Modelling Search and Stopping in Interactive Information Retrieval

David Martin Maxwell

School of Computing Science
College of Science and Engineering
University of Glasgow
Scotland

A thesis submitted for the degree of
Doctor of Philosophy (PhD)

Thesis Abstract

Searching for information when using a computerised retrieval system is a complex and inherently interactive process. Individuals during a search session may issue multiple queries, and examine a varying number of result summaries and documents per query. Searchers must also decide when to stop assessing content for relevance – or decide when to stop their search session altogether. Despite being such a fundamental activity, only a limited number of studies have explored stopping behaviours in detail, with a majority reporting that searchers stop because they decide that what they have found feels *"good enough"*. Notwithstanding the limited exploration of stopping during search, the phenomenon is central to the study of *Information Retrieval,* playing a role in the models and measures that we employ. However, the current *de facto* assumption considers that searchers will examine k documents – examining up to a *fixed depth.*

In this thesis, we examine searcher stopping behaviours under a number of different search contexts. We conduct and report on two user studies, examining how result summary lengths and a variation of search tasks and goals affect such behaviours. Interaction data from these studies are then used to ground extensive *simulations of interaction,* exploring a number of different *stopping heuristics* (operationalised as twelve *stopping strategies*). We consider how well the proposed strategies perform and match up with real-world stopping behaviours. As part of our contribution, we also propose the *Complex Searcher Model*, a high-level conceptual searcher model that encodes stopping behaviours at different points throughout the search process. Within the Complex Searcher Model, we also propose a new results page stopping decision point. From this new stopping decision point, searchers can obtain an impression of the page before deciding to *enter* or *abandon* it.

Results presented and discussed demonstrate that searchers employ a range of different stopping strategies, with no strategy standing out in terms of performance and approximations offered. Stopping behaviours are clearly not fixed, but are rather *adaptive* in nature. This complex picture reinforces the idea that modelling stopping behaviour is difficult. However, simplistic stopping strategies do offer good performance and approximations, such as the *frustration*-based stopping strategy. This strategy considers a searcher's tolerance to non-relevance. We also find that *combination* strategies – such as those combining a searcher's *satisfaction* with finding relevant material, and their frustration towards observing non-relevant material – also consistently offer good approximations and performance. In addition, we also demonstrate that the inclusion of the additional stopping decision point within the Complex Searcher Model provides significant improvements to performance over our baseline implementation. It also offers improvements to the approximations of real-world searcher stopping behaviours.

This work motivates a revision of how we currently model the search process and demonstrates that different stopping heuristics need to be considered within the models and measures that we use in Information Retrieval. Measures should be reformed according to the stopping behaviours of searchers. A number of potential avenues for future exploration can also be considered, such as modelling the stopping behaviours of searchers individually (rather than as a population), and to explore and consider a wider variety of different stopping heuristics under different search contexts. Despite the inherently difficult task that understanding and modelling the stopping behaviours of searchers represents, potential benefits of further exploration in this area will undoubtedly aid the searchers of future retrieval systems – with further work bringing about improved interfaces and experiences.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this doctoral thesis are original and have not been submitted in whole (or in part) for consideration for any other degree or qualification in this (or any other) university.

This doctoral thesis is the result of my own work, under the supervision of Dr Leif Azzopardi *(University of Strathclyde)* and Professor Roderick Murray-Smith *(University of Glasgow)*. Nothing included is the outcome of work done in collaboration, except where specifically indicated within the text.

Permission to copy without fee all or part of this doctoral thesis is granted, provided that copies are not made or distributed for commercial purposes and that the name of the author, the title of the thesis and date of submission are clearly visible on the copy.

David Martin Maxwell
Glasgow, Scotland 🏴󠁧󠁢󠁳󠁣󠁴󠁿 🇪🇺
3rd March 2019

Original Publications

Portions of the research presented in this doctoral thesis are included in the following selected peer-reviewed publications. These are listed chronologically by publication date.

- Maxwell, D. and Azzopardi, L. (2014). Stuck in traffic: How temporal delays affect search behaviour. In *Proceedings of the 5th IIiX*, pages 155–164

- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015a). An initial investigation into fixed and adaptive stopping strategies. In *Proceedings of the 38th ACM SIGIR*, pages 903–906

- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015b). Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM CIKM*, pages 313–322

- Maxwell, D. (2016). Building realistic simulations for interactive information retrieval. In *Proceedings of the 1st ACM CHIIR*, pages 357–359

- Maxwell, D. and Azzopardi, L. (2016b). Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proceedings of the 39th ACM SIGIR*, pages 1141–1144

- Maxwell, D. and Azzopardi, L. (2016a). Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM CIKM*, pages 731–740

- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of the 40th ACM SIGIR*, pages 135–144

- Maxwell, D. and Azzopardi, L. (2018). Information scent, searching and stopping: Modelling SERP level stopping behaviour. In *Proceedings of the 40th ECIR*, pages 210–222

- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2019). The impact of result diversification on search behaviour and performance. *Information Retrieval Journal*. In press.

Presentational Conventions

A number of different presentational conventions have been employed in this thesis for a consistent (and different) look, and to maximise understandability. This section outlines the conventions that have been used.

Spelling

- Spelling is to the *Oxford English Dictionary* (British English). The version that was referred to is searchable online at `https://en.oxforddictionaries.com/`. We prefer a *s* to a *z*!

Fonts and Emphasis

- *Italicised text* is used to define a term and/or concept, but not thereafter. This applies to acronyms, where the full expansion is presented initially; associated abbreviations are used thereafter. Full expansions of an acronym may be reused if required (i.e. in later chapters).

- The main body of this thesis is typeset in 12-point Palatino (body) with 1½ line spacing. Headers, figures and tables (along with their associated captions) use **Foundry Sterling**. The names of tools used and other minor components of this thesis (e.g. table groupings) are also represented using **Foundry Sterling**. For example, the fictional retrieval system Search is used to demonstrate various concepts.[1]

[1] Any resemblance of Search to real-world retrieval systems is unintentional and purely coincidental.

- Emphasis is provided in the form of shaded boxes, such as a section header. These boxes also appear inline to emphasise the introduction of an important concept or term that is key to the thesis.
    - Research questions and hypotheses are also highlighted inline.
    - Important terms and descriptors to this thesis are highlighted when first introduced.
    - We refer to a number of different *stopping strategies* throughout this thesis, each with their own name and at least one variable. These strategies are presented using the notation Name @Value.
    - Cell highlighting is used throughout tables presented within this thesis to represent values of interest – whether they simply are the best reported, or to highlight statistically significant differences. Refer to the caption of a given table for the specific meaning of what cell highlighting denotes.
    - Emphasis is also used to denote labels used in figures presented throughout this thesis. For example, these labels are used to name individual components illustrated within a figure.

URLs

- URLs are used to provide references for claims and to refer readers to external resources. As these resources may become unavailable over time, the date of **L**ast **A**ccess follows each URL – e.g. `http://www.dmax.org.uk` LA *2018-06-07*.[2]

Presenting Concepts and Results

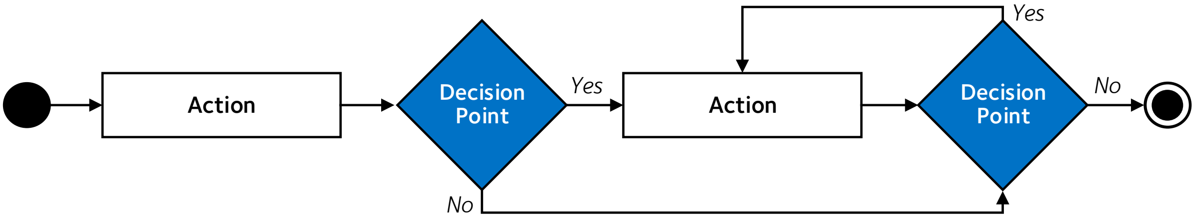
- Pseudo-code that is presented within this thesis uses the *HAGGIS* high-level reference programming language (Cutts et al., 2014), as used by the *Scottish Qualifications Authority (SQA)* for computing science exams.

[2] If a URL becomes inaccessible, the *Wayback Machine* (`https://archive.org/web/`) may provide an archived copy of the page being referenced.

- Plots use a consistent colour scheme across chapters to maximise understandability and comparability. Colours employed are based upon colour schemes as demonstrated to be effective in the online tool outlined by Harrower and Brewer (2003).

- Results are presented to three significant figures. Some representations require a greater degree of accuracy, in which case an appropriate representation will be used.

- In addition to the points described above, *flowcharts* are also used extensively in this thesis to demonstrate the conceptual models that we outline. A standard design for flowcharts is used. It follows the design guidelines provided in `ISO 5807:1985`[3]. Other models presented in the literature also employ such an approach (e.g. Thomas et al. (2014)). The following example demonstrates the symbols used.



The sequence of events begins at the ● and ends at the ◉.[4] Diagram flow can be deduced by examining the direction of the arrows. Actions (or events) are denoted by the text contained within unfilled rectangles □, with decision points represented as ◆. The different outcomes of decisions are denoted by the *italicised* text at each output point of a ◆.

Use of Illustrations Illustrations are used extensively to make the process of reading this thesis a little more enjoyable, as well as (hopefully) providing the reader with a better understanding of points and concepts being conveyed. The author of this thesis drew a majority of the illustrations in *Adobe ® Illustrator ® CS6*.

However, a number of free vector artworks have also been downloaded from `freepik.com` and incorporated within illustrations in this thesis. This statement serves as an acknowl-

[3] `ISO 5807:1985` defines symbols to be used in information processing documentation and gives guidance on conventions for their use in data flowcharts, program flowcharts, system flowcharts, program network charts, system resources charts.

[4] Note that these symbols are not part of the `ISO 5807:1985` standard; they are part of the *Unified Modelling Language (UML)* specification and have been included to ensure diagrams are simple to understand.

edgement that such artworks have been incorporated within this thesis, and are included on the assumption that no part of this work will be used for commercial purposes.

Secondly, the *IKEA* assembly man has been included at the start of Part II to convey the idea of assembling the searcher model proposed in this thesis. Permission has been sought and granted from *Inter IKEA Holding S.A.* to incorporate the IKEA assembly man within this work. *Thank you, IKEA!*

Finally, Picture 1 in the **PhD Journey** shows several of my friends from the School of Computing Science at the University of Glasgow. I sought permission from everyone sitting at the table before including the image in my thesis. *Thanks, team!*

Document Compilation, Rules and Regulations This thesis is typeset using XƎTEX, version `3.14159265-2.6-0.99999`. A custom TEX class (`.cls`) has been developed and used for typesetting. The layout meets University of Glasgow PhD thesis regulations; core requirements of margins, font sizes and line spacing are fully complied with.

Apparatus Used

The user studies reported in this thesis made use of the **TREConomics** framework. The two user studies were crowdsourced in nature, and as such were run over the *Amazon Mechanical Turk (MTurk)* platform.

For the extensive *simulations of interaction*, three computers hosted by the School of Computing Science at the University of Glasgow were used. Basic hardware and software specifications are listed below. FQDNs are obscured to avoid potential security issues.

- `****.***.gla.ac.uk`
  2× *Intel*® *Xeon*® CPU E5-2660, 32 logical cores
  128GB RAM
  Scientific Linux 6.10 *(Carbon)*, `2.6.32-696.1.1.el6.x86_64`

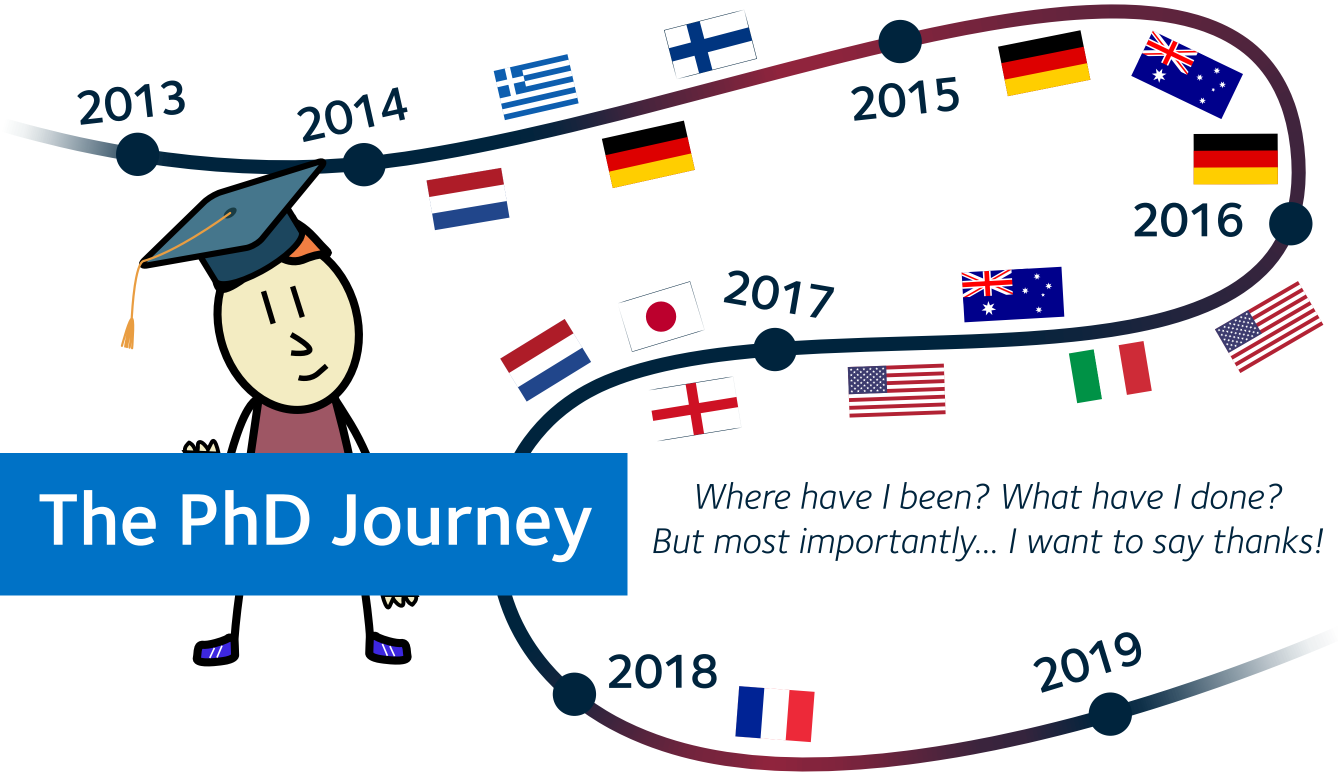
- `******.***.gla.ac.uk`
  2× *Intel*® *Xeon*® CPU E5-2660, 32 logical cores
  128GB RAM
  Scientific Linux 6.10 *(Carbon)*, `2.6.32-573.12.1.el6.x86_64`

- `******.***.gla.ac.uk`
  8× *AMD Opteron*™ processors 6366 HE, 64 logical cores
  512GB RAM
  Fedora 18 *(Spherical Cow)*, `3.11.10-100.fc18.x86_64`

All three computers ran experiments with *Python* 2.7.14, *Whoosh* 2.7.4 and the **SimIIR** framework. All stochastic simulation components were seeded, ensuring reproducible results. Results from seeded random number generation were checked across all three computers – identical results were obtained.

Thanks go to Douglas Macfarlane, Stewart MacNeill and the rest of the support team at the School of Computing Science for ensuring these computers (and the DCS network) were available for all of the experimental work that was undertaken over the years.

The PhD Journey



Where have I been? What have I done?
But most importantly... I want to say thanks!

A good friend of mine (and a fellow PhD student) once said to me that when the time came to write his PhD thesis, he would avoid an acknowledgements section where *everyone and their dog* would be thanked for helping him reach his target of attaining a PhD. I, on the other hand, hold a very different opinion on this matter. There are a lot of people, who have in one way or another, helped me reach where I am today. Whether these people actively guided me in my studies, or were individuals who I was fortunate to become acquainted with over the past five years, they all *"cajoled"*[1] me in one way or another towards the finishing line.

I firmly believe that everybody who I have the pleasure of meeting and working with over the past five years should be acknowledged – whether they feel they contributed in any meaningful way. If you are one of these people and are left wondering, believe me: *you did make a difference.* While acknowledgements may not merit enough gratitude to those who have helped me along the way, I still wish to thank all of you. ***To show my sincere appreciation, I want to dedicate this work to each and every one of them*** – regardless of whether they have a dog or not.

Hindsight tells me that doing a PhD is much like embarking on a *very long,* solo journey. Unless you have experienced it yourself, you won't appreciate how tough (and lonely) it can be at times – especially when things don't go according to plan. Three years in, I found myself sitting in my lab all alone on a Friday night, wondering why my experiments weren't

[1] Professor Ian Ruthven used this term in his PhD thesis (Ruthven, 2001) as a means of describing the individuals who were there for him, behind the scenes, *"cajoling"* him towards the finishing line.

producing the results I had expected and hoped for.[2] It can at times all seem so incredibly pointless, and you find yourself questioning what you're doing with your life. I experienced these lows more times than I care to admit. It can be tortuous. *Impostor syndrome* is something every PhD student feels, and I was no exception.

However, I got to the finishing line. Doing a PhD isn't just about learning your field of study and making an original contribution to it; no, it's much more than that. It also involves learning about yourself. It's *character building.* It involves steely grit and determination to get through the difficult times. Even when everything comes crashing down around you, *you will get through it.* My PhD taught me this more than anything, and for that I am incredibly thankful. I'm definitely a different person for having done it[3] – a much better one (I think so, anyway!), equipped with a good skillset to enter the world and make a positive contribution. Even though every PhD comes complete with negative moments, I took positives from all of them. From this, I could enjoy the good times even more. And believe me, there were *heaps* of good times during the past five years.

One of the many great things about my experience as a PhD student was the office I was given to work in. It's in the *Sir Alwyn Williams Building (SAWB),* room 221. Being a contemporary building, there are lots of windows – and you get a really nice view of the grass next to Lilybank Gardens, looking down to *Brel*, and, yeah, the *Boyd Orr Building,* too. However, in moments of reflection, I always found myself staring vacantly out the window at people walking past outside, going about their lives. Everyone's experiences – from all walks of life – are different, making for a virtually limitless number of stories to listen to, and to learn from. I have always found this truth about life to be absolutely fascinating.

So, on that basis, I want to spend the next few pages writing about *my PhD journey,* acknowledging everyone who made a positive impact along the way. I think that investing the additional time in writing this short passage is a good reflective experience, and also goes a little to say thanks for the amazing things these people have done for me.

I hope you enjoy reading it as much as I enjoyed writing it.

[2] It was a stupid mistake, of course. From memory, I think I forgot to increment a counter in a loop. But it took an entire evening to figure that out. *Of course it did!*

[3] This is something most people will agree with. My friend James gave a nod to this in the acknowledgements of his excellent PhD thesis, too (McMinn, 2018).

Day One — and Looking Back Day one was October 1st, 2013. I remember this day well. In particular, I have vivid memories of sitting down at my new desk in the morning and thinking something along the lines of *"what have I just let myself in for?!"*. The very idea that I was now a PhD student in itself felt really daunting because from reading research papers in my MSci year, I was humbled by how much knowledge there was out there – and I had to get myself to a level to contribute to that knowledge. After being assigned my first task by my supervisor, Leif, I set to work – but it did seem very overwhelming.

However, I chipped away at it. As one task was completed, the next one fell into place – and I found I could do that, too (with some guidance, of course!). I started to produce things, got a paper accepted after six months, and presented it at a conference (as I'll talk about shortly). But as I worked away, I started to find another area of research[4] to be much more interesting. The work presented in the thesis you are reading is actually pretty different from what I thought about doing back on day one. It just goes to show how when you think you have something laid out before you, it's by no means certain that it'll happen.

Life in Glasgow One constant that was present throughout my time as a PhD student at Glasgow was the people. There were always individuals I could rely on for support, advice, or a simple chat. We'd often find ourselves down at Brel when the sunshine was out (which did happen *sometimes*). These are the people that I'd like to acknowledge first – and what better than to start with those who I shared an office with over the past five years?

SAWB221 To Stuart Mackie, فاطمه امين ابراهيم الصافوري (Fatma Elsafoury), my *tocayo* Jorge David González Paule and 王烯 (Xi Wang) – thank you for your companionship throughout the years in *SAWB221*. The camaraderie and support we gave one another did not go unnoticed, and I am grateful for that. Fatma, thank you for the support and interesting philosophical discussions that we had. There are also two other individuals with whom I also shared *SAWB221* with – and also a home (for four years!). To Horațiu Bota, thank you for the friendship that we had over the years throughout our time as PhD students. To จรณะ มโนธรรมรักษา (Jarana Manotumruksa), thank you for your friendship throughout. It's been an absolute pleasure, and it didn't feel like being in an office – you all made it a happy place.

[4] User modelling and simulation – the scope of this thesis.

I'd also like to say thank you to my friends Colin McLellan and Andrew (James) McMinn. All three of us started at the *University of Glasgow* back in September 2008 as undergraduates in Computing Science. Colin was in my very first *CS1Q* undergraduate lab! By early 2019, all three of us had passed our PhD defences at the same institution, although the routes we took to get to that point were slightly different. *We got through it together!* To Colin in particular, thank you for the support and friendship – especially when we were both writing up at the end. Having the same supervisor kept us in close contact with one another – but I don't think either of us has a bad word to say about Leif!

Team IR With all of us working on some aspect of IR, there's also more people within the wider IR group that I would like to acknowledge. My appreciation goes out to everyone who resided next door in *SAWB220* over the years, including رامي سليمان الخوالدة (Rami Alkhawaldeh), شوقي عبدالرقيب الدبعي (Shawki Al-Dubaee), نجود ابراهيم العشبان (Nujud Aloshban), Phil McParlane, Jesús Alberto Rodriguez Pérez (and his brother, Félix Rodríguez Pérez), Stewart Whiting, 辛鑫 (Xin Xin) and 发杰原 (Fajie Yuan). To Stewart in particular, thank you for your support throughout your time as a PhD student – your guidance was greatly appreciated and valued when I started out. You made things seem a little less daunting.

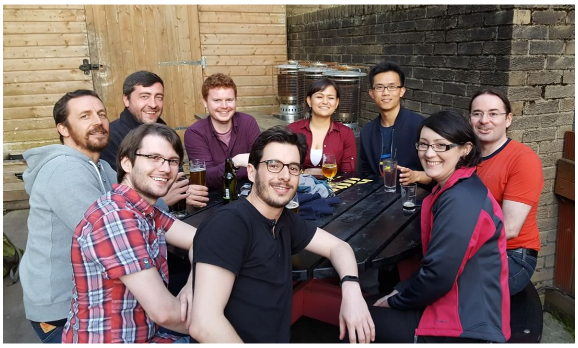
I'd also like to acknowledge those in the rest of the IR group at Glasgow. In particular, I would like to acknowledge Dr Jeff Dalton, Professor Joemon Jose, Dr Craig Macdonald, Dr Richard McCreadie, Graham Mcdonald, 方安杰 (Anjie Fang) and 苏亭 (Ting Su) for their friendship and support throughout the years. Professor Iadh Ounis in particular was a source of great support. Together with Leif and Craig, Iadh taught me many of the basics of IR in my MSci year, for which I am very grateful. Iadh was also one of the examiners for my final PhD defence – and I'll talk about that experience later.

I'd also like to pay particular thanks to ياشار مشفقى (Yashar Moshfeghi) for his friendship and support throughout my time at Glasgow. When you were a PhD student at Glasgow, you were my tutor for the undergraduate *Java Programming 2* course. From the lectures, you helped me to understand and reinforce many of the programming constructs that I use today! Yashar, your expert knowledge and advice on how to run crowdsourced studies was also appreciated. You played an important role in helping me to get the user studies that I define and report on in this thesis up and running. Thank you.

Glasgow Computing Science Of course, I didn't just exclusively interact and socialise with people who studied IR. One of the great things about the *School of Computing Science* at the University of Glasgow is its size and the huge range of different disciplines that are studied. I have made friends with many people along the way, and also learnt things from different research areas, too. It's always interesting to see what other people are working on.

I made some close friends. To Gözel Shakeri, you are the best. I cannot thank you enough for your friendship, support and encouragement that you've given me throughout my time as a PhD student. The support and words of advice through the difficult times – especially when my world came crashing down in mid-2018 – will not be forgotten. Even if I was able to even begin offering you the advice and comfort that you did for me, I will have been a good friend to you, too. And to Frances Cooper, thanks for your friendship and company, especially when you had a brief stay in *SAWB221* during my final writeup phase!

In addition to Gözel and Frances, there are heaps of other people at Glasgow that I want to acknowledge. To name a few... Blair Archibald, Dr Ornela Dardha, Наталья Чечина (Dr Natalia Chechina), Marco Cook, Richard Cziv́a, Euan Freeman, Simon Jouet, Φωτεινή Κατσαρού (Foteini Katsarou), William Kavanagh, Antoine Loriette, Ciaran McCreesh, Stephen McQuistin, Magnus Morton, Алекс Панчева (Alex Pancheva), Craig Reilly, Stefan Raue, Dr Giorgio Roffo, Charlie Rutherford, Kyle Simpson, Robbie Simpson, Dr Michel Steuwer, Lovisa Sundin, Patrizia Di Campli San Vito, Tom Wallis, Dr David White and Михаил Янев (Mihail Yanev) – you all over the years provided a friendly face and support. My appreciation goes out to every single one of you. Even if we simply had a chat and/or a drink, your company meant (and still means!) a lot.



Picture 1 The good old days, back in August 2016. Sláinte, everyone!

I would also like to thank Professor Roderick Murray-Smith. Thank you for your support when Leif left Glasgow in mid-2016 to the *University of Strathclyde*. Your insightful advice and feedback gave me an alternative perspective from which I viewed my work. I was able to incorporate some of your points into the final product, making it a stronger thesis.

And to those friends from outside the School, I want to acknowledge your support and continued friendship. In particular, I'd like to acknowledge Julie Briand, Gary Christie, Adéla Holubová, Nick Swan and Shaun Rew. Julie, thanks for your company throughout the process – we both achieved our goals and got our PhDs! *"Choose your future. Choose life."* I'd also like to mention Sean McKeown – thank you Sean for your friendship throughout the whole experience. I value your advice and feedback, and I hope I have been able to repay that over the years. You have done *Edinburgh Napier* proud.

Tutoring, Exam Collection and More I always said to my friends that when my PhD work was getting tough, I could find some solace in teaching. Throughout my time as a student in the School of Computing Science, I've been incredibly fortunate to take on such important roles – and from those roles, meet and work with some fantastic people. Back when I was a fourth year undergraduate, Professor Quintin Cutts introduced me to the world of teaching. From that moment, I never looked back. Tutoring and demonstrating were some of the best things I did at Glasgow. Sitting down and helping someone understand a solution to a problem that they have been facing in their work was such an enjoyable experience.

I tutored labs for a total of *nine years* – and loved every minute of doing so. While I tutored basics such as *CS1P* and *JP2* (Python and Java programming), my main focus was undoubtedly *web development.* As I'll talk about later, I wrote a book with Leif called *Tango with Django* to make learning the *Django* web application framework a more straightforward experience. I'd like to thank Professor David Manlove and Dr Gerardo Aragón-Camarasa for providing me with the opportunity to continue working on web development with them in my capacity as a tutor. I thoroughly enjoyed working with you both. And to my friend and fellow PhD student Laura Voinea, thank you for your company during the *WAD2* and *ITECH* labs over the years. Working with you was an absolute pleasure.

Of course, there's also the administrative team within the School that kept things flowing smoothly. These were the individuals who supported me when I needed it, too – and I want to acknowledge them here. To Lydia Marshall, Helen McNee and Αναστασία Φλιάτουϱα (Anastasia Fliatoura), thank you for making everything as straightforward as it could have been, at least from an administrative point of view! In particular, I want to thank Anastasia for her help in sorting out the thesis submission dates for me at the end of the PhD.

I also want to acknowledge Teresa Bonner, Helen Border and Gail Reat in the teaching office. You all trusted me to do the job that I did when it came to tutoring, and for that, I am very grateful. One of the other jobs you gave me during my time as a PhD student was to run around the campus during exam season and collect the student's scripts. Although to many this sounds like a nightmare, I actually really enjoyed it. Once again, it provided a nice break from my studies, and I learnt how to sort ~200 exam scripts by matriculation number in the quickest possible time. *Where else would I have got that experience?* Thanks also to Magnus and Laura – as well as Paul Harvey – for your companionship when we spent those days in April-May 2015, 2016 and 2017 running around collecting all the student's scripts!

Of course, all of these extra commitments I took onboard were for the benefit of the students who have studied at the School over the years. As one of their tutors/demonstrators, I've had the good fortune to get to know some wonderful people over the years. Even if I guided them through their studies for a few weeks of their lives, I hope that I left an impact.

In particular, I want to acknowledge Екатерина Александрова (Ekaterina Aleksandrova), Lisa Brooks, 陳文勝 (Winston Chen), Ἄγγελος Κωνσταντινίδης (Angelos Constantinides), David Creigh, Tom Decke, Ивелина Дойнова (Ivelina Doynova), Leisha Hussein, Lisa Laux, Elena Lucchetti, Rebekka Orth, Gabriele Rossi, Vincent Schlatt and Tevhide Turkmen for keeping in touch and your friendship throughout our time here at Glasgow. It has been a pleasure. Sorry that you were inflicted with the pain of having to *Tango with Django* – but I know that in the end, you all tangoed really well!

Broadening my Horizons I had many fantastic times in the School of Computing Science. But one of the perks of being a PhD student is the ability to travel. I had the good fortune to travel for conferences, internships and summer schools. I was originally uneasy at the idea of travelling solo, but after travelling to my first conference, this fear totally disappeared. Travelling became one of my favourite things to do, and I relished the opportunity to visit a new city, country, or continent – and of course to meet new people.

The different flags at the start of this section show the countries I visited throughout my studies. Of particular fondness to me are the memories of my first conference in Amsterdam, The Netherlands *(ECIR 2014)* – and the conference where I did my first presentation, at *IIiX 2014* in Regensburg, Germany. Dr David Elsweiler and Dr Markus Kattenbeck made me feel welcome in Regensburg, and helped to calm my nerves. It went went alright, and I have now given numerous talks about my work all over the world. Although I always feel nervous about doing them beforehand, they always seem to turn out alright in the end. Maybe I'll realise this and learn to calm my nerves one day.

However, individuals that I met helped me put my mind at ease, and made me realise that what I was doing was worthy. At the first *CHIIR* conference in North Carolina, USA, I participated in the Doctoral Consortium, where Dr Jaap Kamps was my mentor. Thank you Jaap for offering me insightful advice for shaping up my work. I'd also like to acknowledge Dr Mark Smucker for the discussion and words of encouragement he gave at the conference.

Support also came from Finland in the shape of a *Short-Term Scientific Mission (STSM).*[5] I was fortunate enough to spend several weeks at the *University of Tampere* back in September 2014 with Professor Kalero Järvelin, Dr Feza Baskaya, Dr Jaana Kekäläinen, Dr Heikki Keskustalo and Teemu Pääkkönen. I am very grateful for the guidance and friendship I received during my stay in Tampere. It was a successful trip, too – the simulation framework **SimIIR** that was used in this thesis makes use of some of the querying strategies that were devised at Tampere, and our work stemmed several interesting collaborations that led to publications at conferences such as *CIKM 2015.*

Picture 2 My name on the door at *Tampereen yliopisto!*

A further collaboration saw me pay a short visit in November 2015 to the *University of Duisburg-Essen* in Germany. Here, I was fortunate enough to work with Professor Norbert Fuhr and my friends Trần Tuấn Vũ and Ιωάννης Καρατάσης (Ioannis Karatassis) on modelling the search process using Markov models. It was an interesting and rewarding time, with a further publication presented at *ICTIR 2017.* I am grateful to all involved for the kindness shown during my stay in Germany. Thank you.

[5] This was funded by the *MUMIA COST Action*, grant no. `ECOST-STSM-IC1002-080914-049840`.

I also interned during the summer of 2017 at the *Alan Turing Institute (ATI)* in London. During this time, I learnt more about mathematics, and contributed to other areas of science. I'm grateful that I was able to apply my computing science knowledge on different problems and would like to thank Professor Terry Lyons and Dr Hao Ni for their guidance throughout the project I undertook with my teammates, Alex Cioba and Radosław Kowalski. Acknowledgements also go to the other wonderful people at the ATI who were there with me, including James Bell, William Kayat, Tim King, Emily Neilson, Bernardo Pérez Orozco, Dr Jeremy Reizenstein and 石海忱 (Haichen Shi) – as well as my friend Dr David White and Faustyna Krawiec for making life outside the ATI during that summer so enjoyable.

However, one of the places that I travelled to that stands out the most in my mind was Australia. In the last quarter of 2016, I had a wonderful time in Canberra working at *Microsoft* with the company of Dr Peter Bailey, Professor Dave Hawking and Dr Paul Thomas (along with Dr Nick Craswell, who was based over in Bellevue, Washington). During my time in Canberra, I learnt so much – not only about how things at Microsoft work, and how to collaborate, but also about myself. The experience was superb, and I wouldn't have had it any other way. To Peter, Dave and Paul – thank you for the kindness and support you showed me during my time there. It was an absolute pleasure working with you all, and I hope that I was suitably able to demonstrate my skills.

Picture 3 Melbourne → Canberra.

Before heading to Canberra however, I visited Melbourne. I'd like to thank my friends Johanne Trippas and Πηνελόπη Αναλυτή (Penny Analytis) for their kindness and hospitality for the first few days of my Australian adventure. I hope I can repay that one day. Driving from Melbourne up to Canberra, although *very* long, was one of my highlights. And once I was at my destination, I didn't look back. I remember the nervousness about approaching a random table of young people in a Canberran bar in an attempt to make friends. I am happy to report that I did it – and in the process gained some fantastic friends! I would like to thank James Dart, Gabrielle McGill and Alan Wu for their friendship and support. In particular, I'd like to thank Gabrielle for her continued support, and for showing me around Sydney. Gabrielle, your country is amazing – and I'd love to pay a return visit.

Indeed, making friends and connections with others was a common theme throughout my travels. It was a brilliant time, and I look back on every place I went to with fond memories. There are so many people that I met during my time as a PhD student, but several people stand out in my mind. These people include ညီညီထွန်း (Nyi-Nyi Htun), 김은정 (Eunjeong Kim), 三井マット (Matthew Mitsui), Հասմիկ Օսիպյան (Hasmik Osipyan), Maya Sappelli, Maria Han Veiga and 유주완 (Juwan Yoo). It was a pleasure to meet you in Pisa, Maria! In particular, I'd like to thank مصطفا دهقانی (Mostafa Deghani) and سمیرا آب‌نار (Samira Abnar) for their amazing company and hospitality in Amsterdam at the end of *ICTIR 2017* when my travel plans fell apart. I hope I can repay your generosity.

The Highs and Lows It goes without saying that the places I mentioned are associated with good times. However, doing a PhD is not a linear process, and I think it's also important to acknowledge that *things didn't always go to plan, either.* As I said back at the start, it's a character building process. I took a positive from every negative experience, and I hope that I'm a stronger individual for having been through the tough times.

Experiments went wrong. Conference deadlines were always a stressful experience. Despite my best planning, I would always find myself staying up late the night before the deadline. One of the lowest points came in early 2017 when after weeks of hard graft, I realised that all of the experimental results for a *CIKM 2017* paper on simulation were totally bogus. It was heartbreaking – but Leif was understanding, and I was able to pick myself back up, fix the problems I identified, and submit the revised work to *ECIR 2018* instead. This kind of thing is something I am surprised I don't hear more of amongst PhD students, because everyone makes mistakes. And from each mistake, you learn something new.

I'd be lying if I said things were straightforward elsewhere in my life, too. The writeup phase is already a difficult and stressful (but somewhat enjoyable) process – and when things went wrong elsewhere in mid-2018, it was a tough pill to swallow. It happens. But I am grateful for my friends (who I have acknowledged) for helping me through that bad period. I got there! And I am grateful to them for helping me get to this point.

The Final Defence *Getting there* led to the *final PhD defence.* I submitted my thesis a few months prior, and it felt like 1,000 tonnes had been lifted from my shoulders. The pressure came back, however, before the final defence. I was super apprehensive about it – *what if I was asked a question I had no idea about?*

However, it was alright! I heard several of my friends who had been through the process in the months leading up to my defence, saying that the process was actually *enjoyable.* I would agree completely. I knew my stuff – having just written a thesis on it – and the questions I was asked were fair and reasonable. It was confidence-inspiring, and when questioned by my examiners, I felt like an *equal to them.* It was an enlightening experience, and I am all the better for having completed it successfully.

To that end, I'd like to say thank you to my two examiners, Professor Iadh Ounis and Dr Suzan Verberne. I am humbled by the warm and encouraging feedback you both gave me on the work I have done, and I hope you enjoyed reading the thesis – and asking me questions on it, too! My thanks also to the convenor of my defence, Dr Michele Sevegnani.

Closer to Home Of course, three of the individuals who I owe the most two are my parents and brother. To my mother, Denise, and my father, William: thank you from the bottom of my heart for the love and support that helped me get to 2019, where I can call myself *Doctor David.* You've both told me that I have done you proud, and I hope that I can continue to do so as I try to figure out my own way through life. The sacrifices that you both made for myself and Alastair have not gone unnoticed, and I am (and I know Alastair is!) eternally grateful for the opportunities that you have both provided us.

To my brother Alastair, thanks for everything you've done for me – especially over the past five years. You've been a strong voice of reason to counter my sometimes clouded judgement, and you've helped steer me through murky waters. Your excellent proofreading abilities are also acknowledged, and I have to thank you for taking the time to read this thesis as I finished chapter after chapter. Keep doing what you're doing. I love you all.

I'd also like to pay tribute to someone who also strongly encouraged me to get into computing – and along with my parents, helped kickstart the idea of attending university in the first place. To Ian Phillips, my computing teacher at *Mearns Castle High School* – thank you.

You saw me through *Standard Grade* and *Advanced Higher Computing*, and you were another individual who was conducive in helping me get to where I am today.

My Supervisor However, there's one more individual I deliberately saved until the end. Without his guidance and support, I am certain I wouldn't be where I am today. Dr Leif Azzopardi was my PhD supervisor, and he also saw me through my fourth and fifth year undergraduate projects. He also taught me many of the IR and programming concepts that I put to good use in this work, and taught me new ways to look at the world.

From the first time I met him as one of my lecturers in 2011, Leif has offered me nothing but absolute support and encouragement. He took the time out of his day to discuss the problems I was facing and warmly welcomed me to his family. He offered me support when I needed it, and never lost his trust in me – even when I lost some trust in myself. And with that trust in me came so many opportunities, perhaps chief of which was co-authoring *Tango with Django* with him, a book that has been used by thousands and thousands of people around the world. He provided me with the ability to travel and meet so many of the fantastic individuals that I have acknowledged in this passage.

Leif, you treated me as an equal throughout, and I look up to you. I would echo Colin in saying that you were both a supervisor and a trusted friend throughout the entire process, and hope that you will continue to be. ***From the bottom of my heart, thank you for everything you have done for me.*** I hope that we can continue to work together in the future.

I will admit that writing this passage took far longer (and *is* longer) than I envisaged. However, as I said at the beginning, I think it's important to acknowledge all of those who played a part in the journey of my PhD – from my supervisor, to all of the friends that I made.

A total of 1,962 days passed from commencing my PhD to successfully defending it in front of Iadh and Suzan. In that time, I've learnt so much – not only about my field of study, but about how the world works. When I left Canberra for the last time in 2016, my friend Gabrielle said to me that even though I really wanted to go back, I'd find that while the buildings would look similar, the place wouldn't feel the same to me. *People move on.* They move on to different places and to do different things. And although this is true, the memories and experiences that I have gained from this five year journey will stay with me for the rest of my days. There's been plenty of ups and downs, yes. It's been a hell of a ride. But I wouldn't change a single thing.

And you know what?

This is only the beginning! 🙂

I would also like to acknowledge my funder, the *EPSRC* (under grant number `1367507`). The scholarship I was provided with allowed me to complete my work without any financial headaches. Thanks also to the *RSC Committee* at the School of Computing Science for all of the funding over the years that allowed me to travel to conferences and present my work.

Table of Contents

List of Figures

List of Tables

Abbreviations Used

For an explanation of a particular abbreviation, *see* the corresponding glossary entry.

AR Aspectual Recall, *see* **AR**

CG Cumulative Gain, *see* **CG**

CSM Complex Searcher Model, *see* **CSM**

DCG Discounted Cumulative Gain, *see* **DCG**

ESL Expected Search Length, *see* **ESL**

FQDN Fully Qualified Domain Name, *see* **FQDN**

HCI Human-Computer Interaction, *see* **HCI**

HIT Human Intelligence Task, *see* **HIT**

HTML HyperText Markup Language, *see* **HTML**

HTTP HyperText Transfer Protocol, *see* **HTTP**

IFT Information Foraging Theory, *see* **IFT**

IIR Interactive Information Retrieval, *see* **IIR**

iP Interactive Precision, *see* **iP**

iPRP Interactive Probability Ranking Principle, *see* iPRP

IR Information Retrieval, *see* IR

MSE Mean Squared Error, *see* MSE

MTurk Amazon Mechanical Turk, *see* MTurk

MVT Marginal Value Theorem, *see* MVT

NIST National Institute of Standards and Technology, *see* NIST

OFT Optimal Foraging Theory, *see* OFT

P@k Precision at *k*, *see* precision

PRP Probability Ranking Principle, *see* PRP

QREL Query RElevance Judgement, *see* QREL

RBP Rank Biased Precision, *see* RBP

RDBMS Relational Database Management System, *see* RDBMS

SERP Search Engine Results Page, *see* SERP

SET Search Economic Theory, *see* SET

TREC Text REtrieval Conference, *see* TREC

URL Uniform Resource Locator, *see* URL

WWW World Wide Web, *see* WWW

Glossary of Terms

AR *Aspectual Recall* considers the number of documents returned by a retrieval system that reference at least one unseen *aspect* of a particular topic. An interactive-based approach can also be considered, where documents identified by searchers are exclusively considered.

CG *Cumulative Gain* is used to measure the effectiveness of a retrieval system (or the searchers that use it). The usefulness or *gain* possessed by each ranked document is considered and accumulated together to produce a final measure. This can be at the *query level* (i.e. considering individual queries), or at the *session level* (considering the total gain acquired over a number of queries).

CSM The *Complex Searcher Model* is the high-level, conceptual searcher model proposed in this thesis. It is a development of existing searcher models provided in the associated literature. The model considers the search session as a whole and incorporates novel improvements to the search process, such as a new *stopping decision point.*

DCG Considered as a natural evolution of CG, *Discounted Cumulative Gain* once again considers the *gain* that can be attained from a document. However, the underlying assumption here is that relevant documents at higher ranks are more desirable. Gain therefore for documents at lower ranks are *penalised,* or discounted, producing a rank-aware measure.

Document In an IR system, a *document* contains information that can be examined. Typically, this would consist of unstructured text (i.e. natural language). However, depending upon the context, a document may contain other forms of information, such as images, audio, or video.

ESL The *Expected Search Length* is an evaluation measure used within IR. It considers the number of non-relevant documents that will have to be examined by a searcher before reaching the desired number of relevant documents. The ESL provides motivation for a number of stopping heuristics used within this thesis.

FQDN A *Fully Qualified Domain Name* is a domain that specifies a host's exact location within a domain name hierarchy. For example, `www` may be a valid hostname, but `www.gla.ac.uk` provides an exact match to the host's location within a wider network.

HCI Human-Computer Interaction is the study of how computer technology is used and designed. It focuses on the interfaces between users and computers.

HIT A *Human Intelligence Task* is the name given to jobs posted on the *Amazon Mechanical Turk (MTurk)* platform.

HTML *HyperText Markup Language* is the standard *markup language* used in the development of web pages and web applications. HTML documents are annotated in a way that is syntactically different from the text, such as through the use of `<tags>`).

HTTP The *HyperText Transfer Protocol* is the underlying protocol used for the transmission of content over the WWW, amongst many other protocols. It defines the rules by which web servers and web browsers can communicate with one another.

Hyperlink A *hyperlink* is a reference to some data source that can be clicked on to jump to the said data source. This concept is most well known as part of the WWW, with the links that hyperlinks create between documents defining the web-like structure.

IFT *Information Foraging Theory* applies the theory and constructs provided as part of OFT. First considered in the 1990s with seminal work by Pirolli and Card (1999), IFT considers searchers as individuals when searching for information. This analogy allows one to consider how instinctive foraging mechanisms employed by animals looking for food in the wild can be applied to humans when *foraging for information*.

IIR A simplistic description of *Interactive Information Retrieval* would be the study of how humans interact with retrieval systems, considering aspects such as their behaviours and experiences. This is in contrast to the study of IR, considering purely *system-sided* aspects.

Information Need A searcher can develop an *information need* when observing some phenomenon in the real world. It is a desire to locate and obtain information to satisfy a conscious or unconscious need. This is typically considered to be one of the first steps of the IIR process.

iP Similar to precision, *interactive precision* considers the fraction of relevant documents relevant to the issued query, as identified *by the searcher.*

iPRP The *Interactive Probability Ranking Principle* (Fuhr, 2008) is an update to the PRP. Within its framework, interaction is considered. This allows for costs over different activities (i.e. issuing queries or examining result summaries and documents) and changes in a searcher's information need.

IR As a field of academic study, *Information Retrieval* could be defined as the study of *"finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)"* (Manning et al., 2008).

KL-Divergence *Kullback-Leibler* divergence, or *relative entropy*, is a measure of how one probability distribution is different from a second.

MSE The *Mean Squared Error* measures the average squared difference between estimated values, and what is being estimated. It considers the notion of bias and standard error, with the lower the MSE, the better the estimation to the real-world observation.

MTurk *Amazon Mechanical Turk* is a *crowdsourcing* platform, allowing for one to coordinate the use of human intelligence to perform tasks that computers cannot presently undertake by themselves.

MVT The *Marginal Value Theorem* (Charnov, 1976) is an optimisation model used to describe the behaviour of individuals foraging in a system where resources are located in discrete patches.

NIST The *National Institute for Standards and Technology* is a laboratory and non-regulatory agency of the *U.S. Department of Commerce.* NIST has been central in providing support to the TREC evaluation effort.

OFT *Optimal Foraging Theory* (Stephens and Krebs, 1986) is a behavioural ecology model that helps to predict how animals behave when searching for food. From the theory, an optimal foraging strategy can be derived and employed that provides the most gain (energy) at the lowest cost.

Patch Considering *Optimal Foraging Theory (OFT)*, a *patch* is considered an area a forager's surrounding environment. In each patch, the forager can extract gain. Using the example by Pirolli and Card (1999), a bird foraging for berries would find berries on different berry bushes. Each bush can be considered an individual patch with different levels of gain. The bird would expend time on a particular bush (within-patch) and then fly to the next patch (between-patch). Under IFT, a patch is typically considered as a SERP.

Precision One of the simplest performance measures, *precision* is defined as the fraction of documents retrieved that are relevant to the searcher's query. This is typically presented as *P@k* or the fraction of relevant documents up to some rank *k*.

PRP The *Probability Ranking Principle* (Cooper, 1971; Robertson, 1977) is a fundamental theory of IR, outlining that for a retrieval system to be effective, it must present results to a searcher in decreasing order of likelihood of the results being relevant.

QREL *Query REelevance Judgements* are a series of judgements that are assigned to documents within a corpus. Typically, these are considered over a per-topic basis, with *binary* or *graded* judgements assigned. As an example, binary judgements would denote that some item is either *relevant* and *not relevant*. These are considered as the ground truth or gold standard judgements of relevance.

Query A *query* is a precise request issued to a retrieval system. Here, a searcher's *information need* is formulated as one or more *query terms*.

RBP *Rank-Biased Precision* (Moffat and Zobel, 2008) is an evaluation measure used within IR. It encodes within it a simple model of searcher behaviour, with a *patience* factor denoting how far down a list of ranked results a searcher is prepared to go.

RDBMS A *Relational DataBase Management System* is a type of database management system based upon the relational model. At a minimum, a RDBMS provides data as a series of *tables*, comprised of rows and columns, and the ability to create *relationships* between the said tables and data.

Recall *Recall* denotes the number of relevant documents that were matched against a searcher's query by a given retrieval system.

Result Summary On a SERP, a *result summary* provides a title, summary and source for a document that was matched to the searcher's query. Result summaries are the *ten blue links* one is accustomed to when interacting with a retrieval system.

Result Summary Level Stopping In this thesis, *result summary level stopping* denotes the stopping decision point when a searcher is interacting with a SERP. It considers the *depth* to which a searcher will examine result summaries. This is typically referred to as *snippet level* or *query level stopping* in the literature.

Searcher A *searcher* is an individual who uses a retrieval system to help him or her satisfy some given information need.

SERP A Search Engine Results Page is the primary output of a contemporary retrieval system (typically WWW-based). It is a page consisting of a series of results that were matched by the retrieval system to the searcher's *query.*

SERP Level Stopping *SERP level stopping* denotes the stopping decision point where a searcher can choose to either *enter* a SERP and begin examining content in detail, or *abandon* the SERP and move on to the next action.

Session Level Stopping *Session level stopping* considers stopping in terms of the overall search session. Typically, this would be evaluated in consideration of time limits or search session goals (i.e. find x relevant documents).

SET *Search Economic Theory* (Azzopardi, 2011) is a theory explaining the search process in terms of economics – in particular *microeconomic theory*. Under this approach, the search process is viewed as a series of *inputs (queries, assessments)* that are used to produce an *output (relevance).*

Stopping Decision Point Core to this thesis, we refer to *stopping decision points* as the decision points within the CSM that permit a searcher to stop their current activity (i.e. examining result summaries or the search session).

Stopping Heuristic A *stopping heuristic* is defined in this thesis as a heuristic that describes the stopping behaviour of a searcher. A heuristic may consider one or more *stopping criteria* when determining a stopping point.

Stopping Strategy A *stopping strategy* is an operationalised stopping heuristic. The heuristic is converted to a series of rules that can be subsequently operationalised – and later implemented – as part of a wider searcher model.

TREC The *Text REtrieval Conference* is an IR evaluation forum, considering a number of different research areas, or *tracks*. Central to the TREC effort is the development of topics, tasks and document collections *(corpora)* that are commonly used in IR experimentation – with this thesis included.

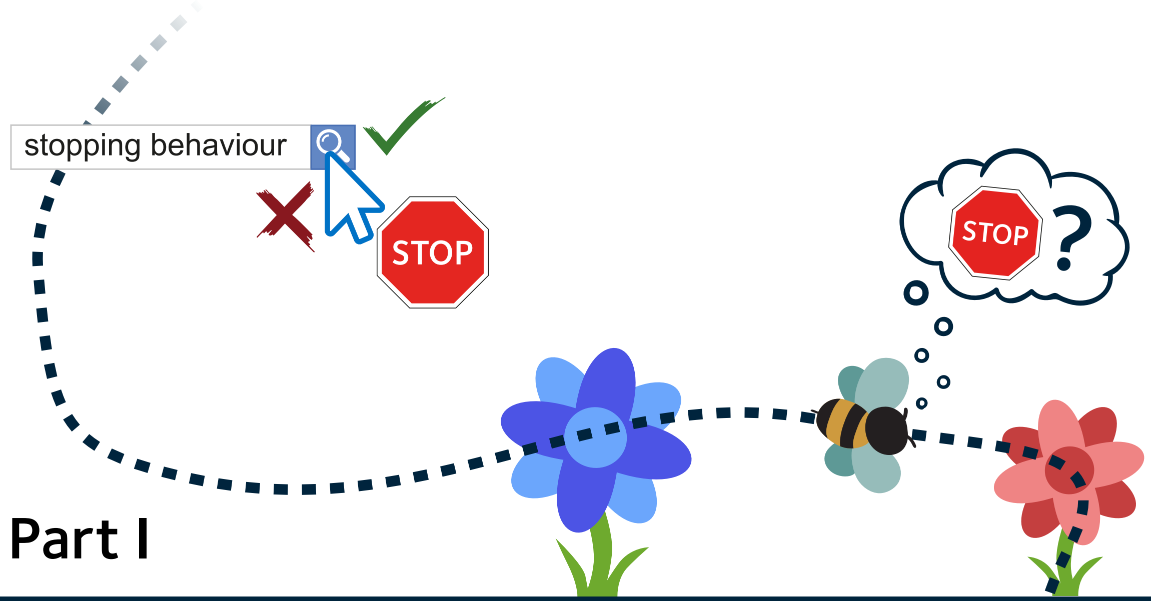
URL A *Uniform Resource Locator* is a reference to some resource hosted on a computer network. It contains the address to the resource and the means by which it can be retrieved. For example, the URL `http://www.dmax.org.uk` specifies that HTTP is used to retrieve content at the address `www.dmax.org.uk`.

User Analagous to a *searcher*.

WWW The *World Wide Web* is an information space in which documents and other resources, linked together via hyperlinks, can be accessed via the *Internet*.

> *“Essentially, all models are wrong,* ***but some are useful****.”*
>
> George E.P. Box, 1919–2013

Oh well. I hope that my model is at least *useful*…



Part I

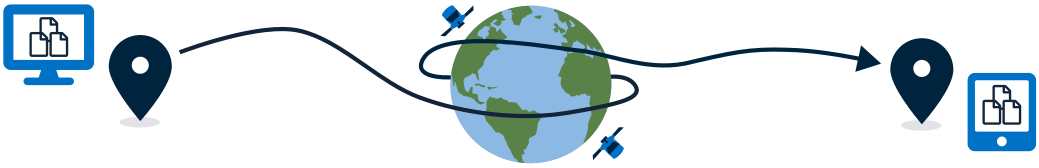
Introduction and Background

In this opening part, we provide an overview of the thesis, present the overarching research questions, and detail the thesis structure. We also provide background to the problem, with a detailed literature review of the various techniques and components commonly used in both IR and IIR — with an emphasis on stopping.

Chapter 1

Introduction

Today, we live in the *Information Age,* an era of human history characterised by the rapid development of technology. This allows for the creation, transmission and retrieval of large volumes of information. Two key developments that have permitted an increase in information generation are the electronic computer and the associated technologies that allow for near-instantaneous communication with devices all around the planet, including the *Internet* and *World Wide Web (WWW)* (Berners-Lee et al., 1994).



Since the early 1990's, the WWW has emerged as the dominant means of publishing information over the Internet, replacing obsolete technologies such as the *Gopher* protocol.[1] As the amount of information available on the WWW grew, so too did the paradigms that were employed by those wishing to seek information on it.

[1] Gopher was designed primarily with a menu-driven interface in mind (i.e. selecting options from a series of choices). The Gopher ecosystem provided the foundations for the *HyperText Transfer Protocol (HTTP)* protocol, which the WWW today utilises.

Information seekers would traditionally *surf* the WWW, starting from a particular domain. From there, they would navigate through the WWW via a series of *hyperlinks* within web pages (or *documents*). This proved practical, as portal websites typically presented categorised lists of websites, much like a telephone directory. However, as the volume of content available on the WWW grew ever larger, this approach became impractical. The development of *search engines* – referred to as *retrieval systems* in this thesis – provided information seekers with the ability to search the ever-increasing universe of documents available at their disposal (refer to Figure 1.1).[2]

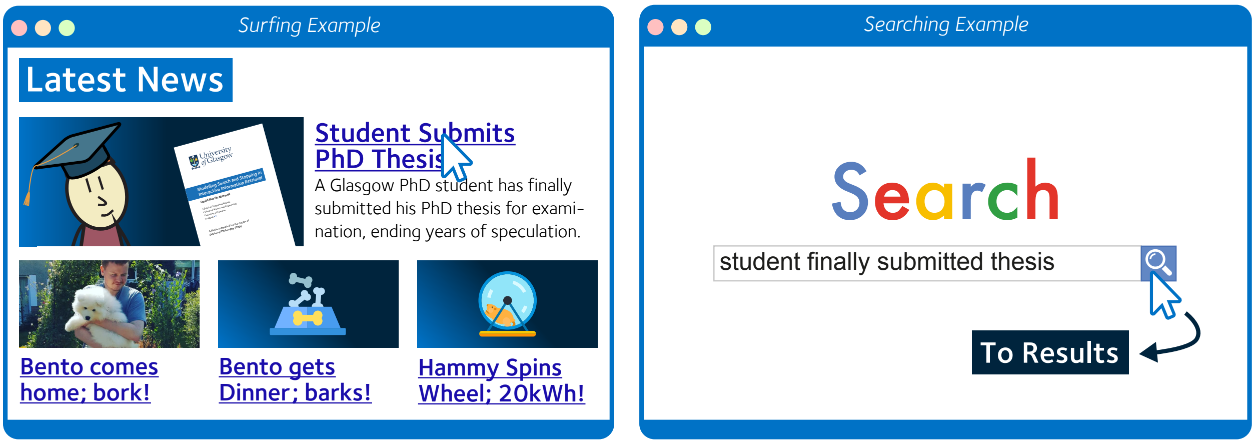
This is not to say that surfing no longer occurs. Information seekers today will often use a retrieval system to find a particular domain. From there, they may then begin surfing within the said domain to find the information that they seek – if such information was not found immediately by the retrieval system. Retrieval systems are however today the most effective way to locate information. Helping searchers realise this by developing efficient retrieval systems is seen as the *raison d'être* of the study of IR.

> *"...but perhaps the key technology that took the web from a useful supplement of current information practice to become the default communication medium is search."*
>
> Wilson et al. (2010)

Contemporary commercial retrieval systems such as *Google* and *Bing* are considered to offer an effective means of finding the proverbial needle in the haystack (Wilson et al., 2010), where near perfect accuracy is regularly attained for popular *queries* (Vaughan, 2004). These retrieval systems, along with the many others in existence today (for use in a variety of contexts[3]), are the product of the collective work undertaken in the field of IR, as we discuss in more detail in Chapter 2.

[2] McBryan (1994) considered a retrieval system as a means of *taming* the considerable number of documents online.

[3] Google and Bing may be the most popular retrieval systems for *general web queries*, but other contexts, for example, can include academic search, enterprise search, multimedia search and patent search.


Surfing Example
Latest News
Student Submits PhD Thesis
Bento comes home; bork!
Bento gets Dinner; barks!
Hammy Spins Wheel; 20kWh!
Searching Example
Search
student finally submitted thesis
To Results

A Glasgow PhD student has finally submitted his PhD thesis for examination, ending years of speculation.

Figure 1.1 The paradigms of surfing and searching. On the left, a seeker will navigate through a series of documents via *hyperlinks* (perhaps without a specific *information need* in mind), while a searcher (right) will issue a query articulating their information need, relying on a *retrieval system* to retrieve a series of documents that are judged to be useful to the seeker.

Retrieval systems aim to make it easier for searchers to satisfy their underlying *information need*. A searcher will develop an information need from a perceived problem – either from a knowledge gap, an internal inconsistency, or a conflict of evidence. This state has been referred to as the *Anomalous State of Knowledge (ASK)* (Belkin, 1980). A searcher, once they have realised this information need, will formulate a query – an expression of what they are looking to seek (Borlund, 2003), typically consisting of a number of different terms. This query is then submitted to the retrieval system, before a potentially relevant set of documents – as judged by the retrieval system – are returned to the searcher. From this set of documents, the searcher can then begin the process of examining them for relevance.

A number of complex interactions take place between an individual seeking information and the retrieval system being utilised (Ingwersen and Järvelin, 2005). This interactive process, where the searcher engages in dialogue with the retrieval system, is considered the study of *Interactive Information Retrieval (IIR)* (Borlund, 2003). One of the fundamental aspects of IIR is that of stopping – where, for example, a searcher must decide when to stop examining the list of results returned to him or her.

Examining stopping behaviour is one of the many different aspects of interaction that have been examined to help us better understand a searcher's behaviours. This knowledge can be used to make the search process a more seamless experience for the individuals using a retrieval system. As we discuss in the next section, much of the research in both IR and IIR has been limited in terms of examining stopping. Subsequently, these limitations provide motivation for the work that we present in this thesis.

1.1 Motivation and Context

Central to much of the work undertaken in the field of IR is the **Cranfield paradigm**, a term denoting a standardised approach of IR evaluation (Cleverdon et al., 1966). Primarily credited to Cyril Cleverdon at Cranfield University[4], the paradigm revolves around the notion of standardised test collections – standardised corpora of documents that can be used by different researchers, providing a uniform foundation for IR experimentation.

While the basic principles of the Cranfield paradigm have remained in place since it was established in the 1960's, aspects of the approach have evolved over the years to cater for the ever increasing complexity of the tasks trialled (Harman, 2010). The approach is widely used in evaluation forums, such as *NTCIR (NII Testbeds and Community for Information access Research)* and *CLEF (Conference and Labs of the Evaluation Forum)*. However, one of the best-known evaluation forums following the paradigm is the *National Institute of Standards and Technology (NIST)* sponsored **Text REtrieval Conference (TREC)** (Harman, 1993). Indeed, the work reported in this thesis extensively utilised material generated as part of TREC efforts, provided as part of different *TREC Tracks* over a number of years.

With the Cranfield paradigm, significant advances have been made possible regarding the evaluation of IR systems. However, the approach can be argued to be somewhat limited from the context of IIR as it highly abstracts the interactions that take place between a

[4] Cranfield University is located at Cranfield, Bedfordshire, England. It is a unique university in that it has a semi-operational airport, given its heritage with aeronautics research.

searcher and a retrieval system (Borlund, 2000; Ingwersen and Järvelin, 2005). In other words, the paradigm broadly fails to consider the complexities of the IIR process. As an example of such a complexity, searchers could issue multiple queries during the course of a search session. Subsequently, they would adapt their interactions based upon the perceived quality of presented ranked result lists (Moffat et al., 2013).

A key example of such behaviour adaption is the searcher's stopping behaviour. For example, a poor set of results may mean that searchers would stop examining results comparatively early than a set of results perceived to be of good quality. Searchers also often stop once they feel that they have found sufficient information to satisfy their information need (Zach, 2005). Indeed, selecting good terms to use within a query is difficult yet important (Efthimiadis, 2000). The initial query posed in a search session often acts as an entry to the search system, followed by phases of browsing and query reformulations (Marchionini et al., 1993). Searchers also will typically abide by the *principle of least effort,* whereby they strive to minimise the expected rate of work expenditure over time (Zipf, 1949).

The experimentation paradigms that have evolved from Cranfield make a series of different assumptions that are largely at odds with how searchers interact with retrieval systems. Namely, these assumptions state that a searcher will:

- issue a *single query* over the course of a search session;
- *examine documents to a fixed depth* (typically $1,000$ in TREC experimentation); and
- *assess all documents to the fixed depth.*

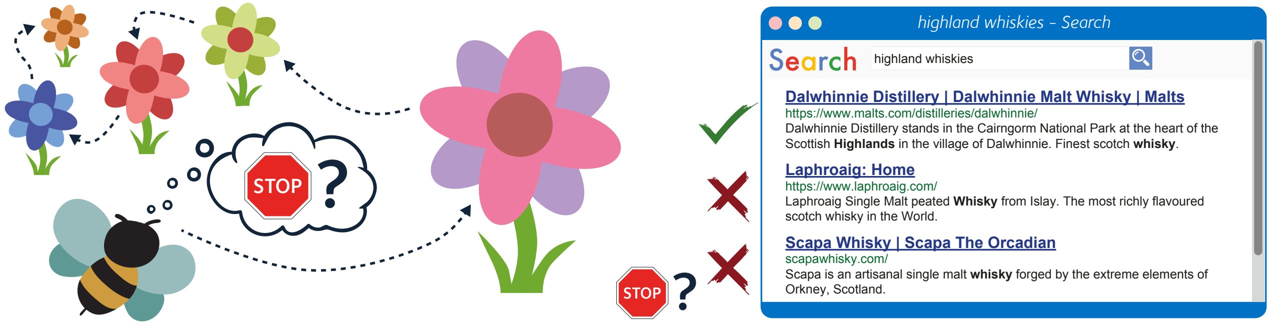
While providing a simple platform for performing retrieval system evaluation, such assumptions are unrealistic. Herein lies a fundamental disconnect between the studies of IR and IIR – the naïve assumptions made of searchers within IR experimentation listed above do not hold when considering the complex interactions that actually take place during the IIR process (Ingwersen and Järvelin, 2005). In order to address the fundamental disconnect between the two fields, we need to create **more realistic searcher models** that better articulate what real-world searchers actually do. A better searcher model would ultimately mean a better understanding of the complex interactions that take place, which would lead to an improved understanding of how to assist searchers. Work to improve our understanding has been undertaken in the field of IIR to address this, examining searcher behaviours under a number of different phenomena – including (but not limited to) the following:

- *query formulation and suggestions* (Azzopardi, 2009; Azzopardi et al., 2007; Baskaya et al., 2013; Carterette et al., 2015; Jordan et al., 2006; Keskustalo et al., 2009; Verberne et al., 2015);
- *browsing behaviours* (Carterette et al., 2015; Chuklin et al., 2015; Guo et al., 2009; Pääkkönen et al., 2015; Smucker, 2011);
- the influence of *costs and time* (Azzopardi, 2011; Baskaya et al., 2013); and
- *performance over search sessions* (Luo et al., 2014, 2015).

When considering how we model searcher interactions, a further (and particularly important) phenomenon largely overlooked in the above is a **searcher's stopping behaviour**. Indeed, given its title, this is what we consider in this thesis – *how can we make improvements to searcher models when considering stopping behaviours?* This phenomenon is now seeing an increasing amount of time devoted to its examination. In the following subsection, we provide an argument as to why examining this phenomenon is important.

1.1.1 Considering Stopping Behaviours

Knowing when to stop is a fundamental aspect of animal – and by definition, human – thinking and behaviour. There must come a time when an animal must stop what it is doing. In the natural world, for example, a honeybee, when *foraging* for pollen, will eventually make a decision to stop collecting pollen on the flowerhead it finds itself on and flies away



Dalwhinnie Distillery | Dalwhinnie Malt Whisky | Malts
https://www.malts.com/distilleries/dalwhinnie/
Dalwhinnie Distillery stands in the Cairngorm National Park at the heart of the Scottish **Highlands** in the village of Dalwhinnie. Finest scotch **whisky**.

Laphroaig: Home
https://www.laphroaig.com/
Laphroaig Single Malt peated **Whisky** from Islay. The most richly flavoured scotch whisky in the World.

Scapa Whisky | Scapa The Orcadian
scapawhisky.com/
Scapa is an artisanal single malt **whisky** forged by the extreme elements of Orkney, Scotland.

Figure 1.2 Examples of stopping. On the left, when will the bee move from one flowerhead to the next? On the right, under the context of information seeking, how far down a list of ranked results will a searcher go before he or she decides to stop examining content? In the example above, Search has failed to return a comprehensive list of highland single malt whiskies 🏴󠁧󠁢󠁳󠁣󠁴󠁿. Will the searcher become frustrated with this, and stop examining results early?

to another flowerhead. The honeybee is in essence attempting to maximise the amount of *gain* (pollen) she accumulates over time on each *patch* (flowerhead) that is visited.

If we consider stopping from an information seeking context, there are many different examples we can use to demonstrate why this behaviour is of great importance. For example, a searcher may decide to stop searching for information when the documents presented show a large volume of non-relevant material, frustrating the searcher (Cooper, 1973b) – perhaps because the retrieval system failed to gauge the searcher's *query intent* (Ashkan et al., 2009), as demonstrated in Figure 1.2. Searchers could also stop examining content after they have become satisfied with the information found previously in a search session (Cooper, 1973a; Gibb, 1958; Simon, 1955), or if they feel that the information being presented is too similar to what has been found earlier (Nickles, 1995).

A number of different *external factors* can influence the decision of when one should stop. Examples of these include the bee finding a flowerhead with no pollen, or time pressures when searching for information. However, Nickles (1995) argues that knowing when to stop is largely determined by a series of *internally defined stopping criteria* that the decision maker employs, just like the examples defined above. Therefore, this internal construct makes stopping a phenomenon that is difficult to model in an effective way. Given that internal factors are a major drive in determining when to stop, studies have largely been unable to quantify *why* searchers stop, other than what they find during the search process gives them the feeling that the located information is *"good enough"* (Zach, 2005).

In contrast to this vague definition of stopping behaviour, several researchers have attempted to create a series of reasoning- and judgement-based **stopping heuristics** that attempt to formally define when a searcher should stop. It is these stopping heuristics that we will primarily consider in this thesis. These heuristics can then be integrated within a wider searcher model, allowing us to determine whether they improve or worsen approximations of actual searcher stopping behaviours. From here, we can then begin to ascertain potential answers to what the feeling of *"good enough"* (or even *not* good enough!) may entail. The searcher model can incorporate stopping behaviours at a variety of different **stopping decision points** – such as at an individual result summary level *(how far down this list of ranked results should I go?)*.

Examining stopping behaviours during search is important because it considers the judgements of a searcher as part of their interactions. For example, it would be prudent of a searcher examining a ranked list of results that are mostly non-relevant to stop early, thus saving time and effort (thus making the searcher more *efficient*). Stopping behaviour is also implicitly or explicitly encoded within a variety of different IR and IIR measures. Obtaining a better understanding of when searchers stop means that we can encode this information within measures of search (improving their credibility), and provides an evidence-based approach to mapping these measures with what actually takes place in reality.

1.2 High-Level Research Questions

Having set out the problem space above, we can now begin to formulate the four high-level research questions that the work in this thesis addresses, denoted as **HL-RQx**. Our first research question considers the concept of modelling searchers, and how, with an emphasis on examining stopping decision points, we can improve current models to better reflect actual searcher behaviours – in particular, their stopping behaviour.

- **HL-RQ1** How can we improve searcher models to incorporate different stopping decision points?

As previously stated, being able to improve upon the current searcher models from the perspective of stopping should allow those subscribing to such a model to become more efficient as to how they search. Closely related to this advancement in modelling this process is the consideration of the various stopping heuristics.

- **HL-RQ2** Given the stopping heuristics defined in the literature, how can we encode these heuristics into a series of *operationalised,* programmable **stopping strategies** that can be subsequently incorporated into the searcher model and evaluated?

Stopping heuristics that we detail later in Section 3.2 are high-level in nature and do not provide an explanation as to how they can be operationalised within a wider system. The challenge that must be addressed in order to answer this second high-level research question will be how we can operationalise such stopping heuristics.

With a more realistic searcher model from **HL-RQ1** and a series of stopping strategies defined by addressing **HL-RQ2**, how well does this combination perform?

- **HL-RQ3a** Given the aforementioned operationalised stopping strategies, how well does each one perform?
- **HL-RQ3b** How closely do the operationalised stopping strategies compare to the actual stopping behaviours of real-world searchers?

These questions are of course of a very broad nature, and it is simply not possible to evaluate them in every conceivable search context. As such, we will examine different contexts that are likely to impact upon searcher stopping behaviours. Specifically, we will examine topical *interactive search* in the domain of news, where we will consider various conditions: search goals and task types; retrieval systems; and result summary length. In the following section, we expand upon these conditions to provide a concrete set of thesis contributions.

1.3 Thesis Contributions

This thesis presents a number of key contributions. Listed below, we consider primary contributions from conceptual, theoretical, methodological and empirical standpoints.

Conceptual **Complex Searcher Model** Our first contribution is a new searcher model. Taking current searcher models, we propose an updated, high-level model of the search process called the *Complex Searcher Model (CSM)*. This provides us with a solution for addressing **HL-RQ1** . Outlined in Chapter 4 (page 107), the conceptual CSM outlines a series of different activities and decision points that searchers undertake throughout the search process, and establishes a flow of interaction based upon established models. Within the CSM are a number of different innovations, key of which is the new stopping decision point. For example, this improvement allows us to ascertain a better understanding of the search process, and the complex interactions that occur between a searcher and retrieval system. Being a conceptual model, we can take the CSM and instantiate it in a number of different ways. The stopping strategies that we consider in this thesis, for example, provide a means for instantiating stopping decision points within the CSM.

Theoretical **Stopping Strategies** As previously discussed, there is a range of different stopping heuristics that have been defined in the literature that provide an explanation for when searchers should stop examining content. The second major contribution of this thesis is the development of twelve operationalised stopping strategies. These may then be subsequently deployed as the logic underpinning a stopping decision point of the CSM (as defined above). These twelve strategies encode a total of seven different stopping heuristics and IR measures. The operationalised stopping strategies provide a solution to HL-RQ2.

Methodological The proposed CSM and the twelve stopping strategies that we operationalise need to be evaluated, such that we can then subsequently address the two remaining high-level research questions, HL-RQ3a and HL-RQ3b. To do this, a general methodology outlines an approach undertaken for user studies. Simulation is then used to determine how the different stopping strategies perform over each of the different search contexts trialled, and how the stopping strategies compare to actual searcher behaviour.

Empirical Varying Result Summary Length We report on a study where the length of individual result summaries presented to searchers are varied to determine what impact that this has on searcher stopping behaviours. As we modify the length of result summaries, we also argue that we influence the overall quality of result summaries. We then perform a simulated analysis examining each of the stopping strategies, determining what strategies perform best and offer the closest approximations to real-world stopping behaviours.

Empirical Varying Goals, Tasks and Systems We report on an additional user study, examining the impact of stopping behaviours when the search task and goals are changed. For this, we consider topical *ad-hoc retrieval*[5], along with a diversified search task, changing the overall goal of what searchers are looking to find. This is then complemented by a further simulated analysis, examining the individual stopping strategies like above.

Empirical New Stopping Decision Point The final empirical contribution complements the conceptual contribution of this thesis, addressing HL-RQ1. We perform a further simulated analysis, examining how well the new stopping decision point performs when incorporated within the CSM – and whether it offers better approximations to actual searcher stopping behaviours.

[5] The ad-hoc search task is explained in detail in Section 2.3.1.1 – it is one of many different types of search task that can be performed by searchers.

1.4 Thesis Statement

Given the above, the major claim of this thesis is that by considering stopping behaviours at different points throughout the search process, we can develop more credible and realistic models of the said search process. These more advanced models can be used as a tool for improving our understanding of stopping behaviours and other complex interactions that occur when searching. Findings from this work can then subsequently aid researchers in the development of more intuitive (and realistic) measures used to facilitate the evaluation of retrieval systems and their users.

1.5 Origins of the Material

Material presented in this thesis has appeared in several conference papers and journals throughout the duration of the author's PhD programme, from October 2013 to March 2019. All are listed in the front matter of this thesis in chronological order. In this section, we provide a narrative, explaining how the developments in the listed publications led to the contributions of this thesis. Work can be considered over three main strands:

- the development of the conceptual and theoretical contributions to this work;
- the development of the **SimIIR** framework; and
- a series of empirical studies.

Conceptual and Theoretical Work on the *Complex Searcher Model (CSM)* has been undertaken over a number of years, and were presented in various publications. Several iterations of the CSM have been developed, with each iteration offering refinements to improve its realism.[6] The first iteration of the CSM – essentially analogous to prior models of search

[6] To simplify reporting (and use) of the CSM in this thesis, we consider only the latest revision of the model.

outlined in Sections 2.3.1.2 and 2.3.5 – was used in simulated analyses, as reported in the two publications listed below.

- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015a). An initial investigation into fixed and adaptive stopping strategies. In *Proceedings of the 38th ACM SIGIR*, pages 903–906
- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015b). Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM CIKM*, pages 313–322

These publications are notable for also including a number of operationalised stopping strategies, providing the foundations for the second major contribution of this thesis. The stopping strategies defined in these publications were used in subsequent publications. Further developments to the CSM were found in a subsequent publication which experimented with the notion of developing *intelligent search agents*.

- Maxwell, D. and Azzopardi, L. (2016a). Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM CIKM*, pages 731–740

The final development of the CSM led to the inclusion of an additional stopping decision point. This new stopping decision point was tested with a thorough empirical analysis, as reported in the publication enumerated below.

- Maxwell, D. and Azzopardi, L. (2018). Information scent, searching and stopping: Modelling SERP level stopping behaviour. In *Proceedings of the 40th ECIR*, pages 210–222

SimIIR Framework One of the major pieces of scientific apparatus utilised throughout all of the aforementioned studies is the **SimIIR** framework, which we discuss in Section 6.4.1 on page 159. Conducting the extensive simulations of interaction we report in this thesis would not have been possible without it. A demonstration paper presenting the framework and the various components that could be instantiated within it has been published.

- Maxwell, D. and Azzopardi, L. (2016b). Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proceedings of the 39th ACM SIGIR*, pages 1141–1144

Empirical Studies The general methodology that we employ for the third major contribution of this thesis has been introduced and refined in the publications listed previously. In addition to this, a basic description of the methodology is provided in a Doctoral Consortium paper that the author presented at the first *ACM Conference on Human Information Interaction and Retrieval (CHIIR)* in Chapel Hill, NC, USA.

- Maxwell, D. (2016). Building realistic simulations for interactive information retrieval. In *Proceedings of the 1st ACM CHIIR*, pages 357–359

The results of two user studies have also been published, and are of direct relevance to the work detailed later in this thesis.

- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of the 40th ACM SIGIR*, pages 135–144
- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2019). The impact of result diversification on search behaviour and performance. *Information Retrieval Journal*. In press.

These studies provide the grounding for simulated analyses that we also consider later in this thesis. The data extracted from these user studies provides credibility to our simulations through the extraction of aspects such as interaction costs and probabilities.

1.6 Thesis Outline

This section provides a brief summary of the remaining parts and chapters of the thesis.

Part I The remainder of Part I concerns prior work that has been undertaken in the fields of IR and IIR. Two chapters outline the basics of IR and IIR (with particular emphasis to how models and measures that we commonly employ consider stopping), before examining the literature that has explicitly considered searcher stopping behaviours.

Chapter 2 Beginning on page 21, this chapter provides an overview of the key concepts of the fields of IR and IIR. We focus on core IR concepts, such as the indexing and retrieval processes (including retrieval models). We then move towards a more user-centric examination of established methods in the field of IIR, such as the consideration of various evaluation measures that are commonly used. We also outline different searcher models that have been previously defined in the literature. These capture the activities and decisions that individuals perform while searching.

Chapter 3 We then consider work that has considered stopping in relation to search. In this chapter, we begin by describing various stopping heuristics defined in the literature. We summarise previous user studies that have examined searcher stopping behaviours, and then consider key theoretical models of search that provide explanations for when individuals stop.

Part II Beginning on page 106, Part II presents the conceptual and theoretical contributions of this thesis, including a discussion of the CSM. In this part of the thesis, we also provide an outline of the general methodology that is used in Part III.

Chapter 4 This chapter introduces the CSM, discussing the advances that the conceptual model provides over contemporary searcher models. We discuss the key stopping decision points provided by the CSM that are central to this thesis, before discussing the assumptions of the model. This partly addresses **HL-RQ1** – evaluation of the model is also required, and is discussed in Chapter 9.

Chapter 5 In this chapter, we introduce and discuss the various stopping strategies that we operationalise as part of the contributions of this thesis, thus addressing **HL-RQ2**. Each of the different stopping strategies, complete with examples, are discussed in depth. The chosen stopping strategies are linked back to their originating stopping heuristics, which are detailed in Chapter 3.

Chapter 6 This chapter outlines our general methodology, detailing the high-level structure of the scientific method used in our empirical work. We also provide a discussion of common approaches that we used across all subsequent chapters.

Part III The third part of this thesis considers our empirical contributions. In this part, we present the user studies that were undertaken, as well as a number of simulated analyses that allow us to address research questions **HL-RQ3a** and **HL-RQ3b**.

Chapter 7 The first empirical chapter considers how stopping behaviours vary when the length (and thus quality) of result summary snippets are varied. We provide a discussion of a user study that examined this phenomenon, before summarising the findings of simulated analyses that were conducted in order to determine what stopping strategies offered the best performance and approximations of real-world searchers under this context.

Chapter 8 In this chapter, we report on a user study examining how a searcher's stopping behaviour varies when subjected to conditions that vary the task, goal, and system used. We then again perform simulated analyses to examine these stopping behaviours in more detail.

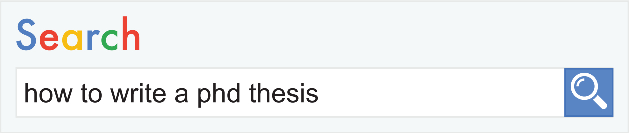
Chapter 9 The final chapter wherein novel findings are presented considers the new stopping decision point that is provided by the CSM. We empirically test the CSM, allowing us to determine whether the inclusion of the new stopping decision point discussed in Chapter 4 provides improvements in overall performance and approximations of actual searcher stopping behaviours. As such, this chapter provides sufficient evidence, in conjunction with Chapter 4, to address **HL-RQ1**. We utilise data from user studies discussed in Chapters 7 and 8 to ground our simulations.

Part IV The final part of this thesis consists of a solitary chapter, **Chapter 10**. The concluding chapter of this thesis provides a summary of the work that was undertaken, and the results obtained. We then discuss potential avenues for future work.

Chapter 2

Information Retrieval: A History and Background

Searching for information on computers is today commonplace, thanks to the proliferation of the WWW and commercial search engines[1]. Despite potential negatives that these technologies may bring – turning us into *shallow thinkers* (Carr, 2008), for example – retrieval systems today by and large make our lives easier, allowing us to find the proverbial needle in the haystack with minimal effort. These results are returned to us while honouring the implicit searcher contract of a timely response (in the order of milliseconds).



Search

how to write a phd thesis

Central to the development of retrieval systems is the study of *Information Retrieval (IR)*. One of the key developments in the advancement of IR as a field was the creation of a *de facto* approach to studying IR and performing IR experimentation. This was developed in tandem with a series of *retrieval models* underpinned by different theories, and the means by which we could *evaluate* their effectiveness. We begin this chapter with a brief overview of

[1] Or, as we refer to them in this thesis, *retrieval systems*.

the history of IR, acknowledging the manual and mechanised systems that predate contemporary computer-based retrieval systems. After this, we move on to discussing the basics of what constitutes an IR system. From there, we discuss work that has switched the core focus of research from the *system* to the *searcher*, introducing the field of *Interactive Information Retrieval (IIR)*. Included in our discussion of the searcher are some of the current *searcher models* that encapsulate the different activities that they may perform. We then conclude the chapter with a discussion of the various measures used for IR and IIR evaluation.

2.1 A (Brief) History of Information Retrieval

While many associate the study of IR with computers, the need to seek information in a quick and effective manner has existed throughout human history. In this section, we provide a very brief overview of some of the key advancements in what can be considered to be the study of IR – from library cataloguing approaches, to contemporary retrieval systems.[2]

2.1.1 Libraries and Mechanisation

Containing a large volume of books discussing a virtually unlimited range of categories, *libraries* require the need for a means of organising (and thus easily locating) information with relative ease. *Catalogues* provide a way in which to achieve this, with ancient Greek poet Callimachus being the first person to create a catalogue in the third century BC (Eliot and Rose, 2009). A more recognisable approach to categorising content was devised by Dewey (1891) with the *Dewey Decimal System*. The use of cards as an *indexing system* was also considered by individuals such as Soper (1918) who invented a system of providing information on what category a card belonged to based upon a punched hole.

However, finding information using these techniques was *slow*. In order to speed up the process of finding useful material, mechanised techniques were also used. Allowing for

[2] An excellent, in-depth discussion on the history of IR is provided by Sanderson and Croft (2012).

searching at the rate of 600 cards per minute, Luhn devised in the early 1950s a mechanised system that utilised punchcards and light. As stated by Sanderson and Croft (2012), this was also around the time that the term *Information Retrieval (IR)* was used (Mooers, 1950). From this point in history, computer technology was developing at a rapid rate. Ultimately, this led to computerised systems superseding mechanised approaches (Jahoda, 1961).

2.1.2 The Rise of Computers

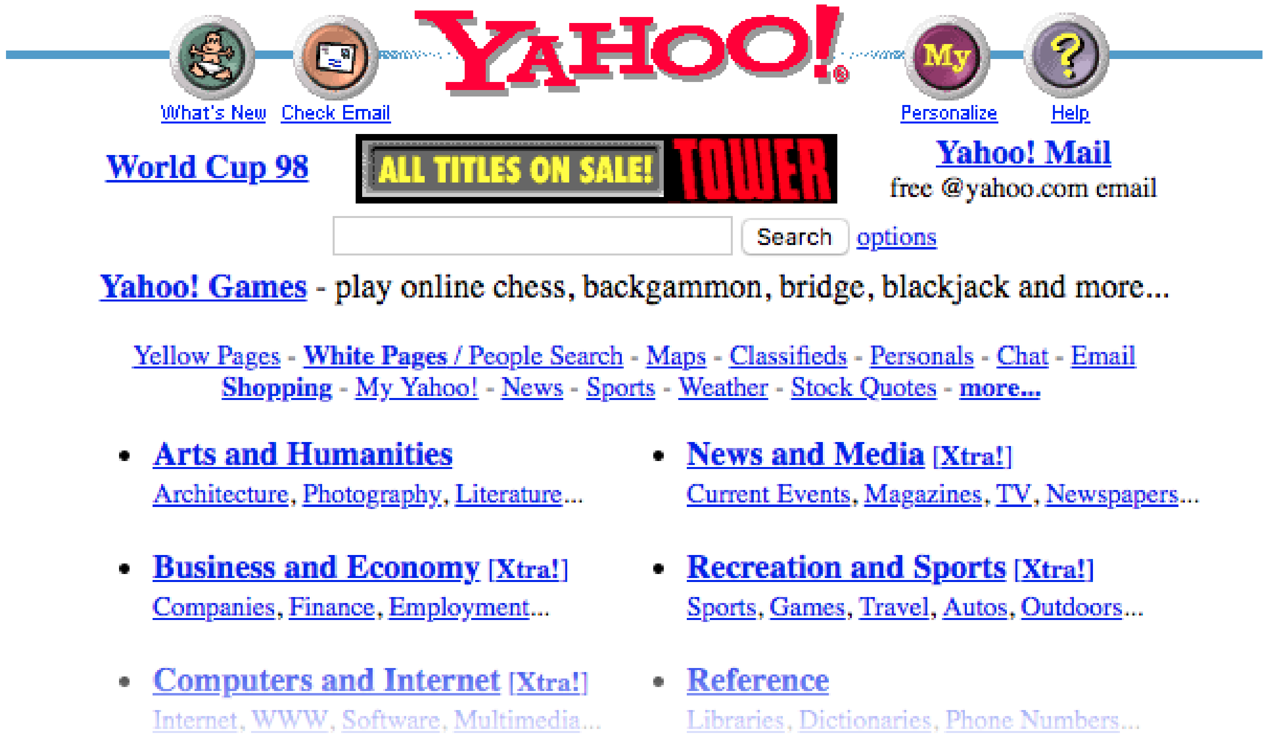
Computers now provide the underlying technologies with which we closely associate with a typical, contemporary IR system. Sanderson and Croft (2012) state that digital storage capacity (e.g. hard disks, and more recently, solid state storage) roughly doubles every two years. This claim is essentially analogous to the famous *Moore's Law* (Moore, 1965), which observes that the number of transistors in a processor (or other integrated circuits) doubles roughly every two years.[3] Indeed, the speed at which modern day computers can search vast indexes and databases of content is vastly superior to traditional cataloguing approaches. These technological advances permit the near instantaneous returning of results from an initial request, with searchers expecting a set of results in the order of milliseconds.

Progressing from computers was the development of *computer networks,* permitting the transmission of information between computers over increasingly large geographical distances. With the development of the Internet, the scene was set for the introduction of one key technology – the *World Wide Web (WWW).*

2.1.3 The World Wide Web

The distribution and ability to search for information over computer networks such as the Internet was traditionally undertaken with legacy protocols such as *Gopher.* Gopher would

[3] As of 2019, it is becoming increasingly difficult to develop integrated circuits that meet this rule of thumb.



World Cup 98

Yahoo! Mail
free @yahoo.com email

Yahoo! Games - play online chess, backgammon, bridge, blackjack and more...

Yellow Pages - White Pages / People Search - Maps - Classifieds - Personals - Chat - Email
Shopping - My Yahoo! - News - Sports - Weather - Stock Quotes - more...

- Arts and Humanities
  Architecture, Photography, Literature...
- News and Media [Xtra!]
  Current Events, Magazines, TV, Newspapers...
- Business and Economy [Xtra!]
  Companies, Finance, Employment...
- Recreation and Sports [Xtra!]
  Sports, Games, Travel, Autos, Outdoors...
- Computers and Internet [Xtra!]
  Internet, WWW, Software, Multimedia...
- Reference
  Libraries, Dictionaries, Phone Numbers...

Figure 2.1 A screenshot of the landing page of *Yahoo!*, as shown on July 5th, 1998. Notice the link for the 1998 *FIFA World Cup* that was taking place at the time. More central to this thesis is the inclusion of a list of page categories in conjunction with the now ubiquitous search box. Screenshot acquired from the *Internet Archive* (under fair use).

provide a series of options for a user to select (i.e. categorisation of content), akin to the traditional library cataloguing approaches described above.

The advent of the WWW in the early 1990s brought about a new type of IR system – *web retrieval systems*. Regarded as the first experimental web retrieval system, *JumpStation* was outlined by McBryan (1994).[4] In this system, *anchor text* within *hyperlinks* of *HyperText Markup Language (HTML)* pages could be exploited to aid the ranking of documents. However, popular web retrieval systems of the 1990s initially followed the categorisation approach hailing back from libraries, as illustrated in Figure 2.1 with a screenshot of *Yahoo!* from 1998. This categorisation approach on the Yahoo! front page ties in with the surfing paradigm described back in Chapter 1. However, as the volume of information on

[4] With *JumpStation* developed and hosted at the *University of Stirling*, could one make the claim that web retrieval systems are a Scottish invention?

the WWW rapidly increased, this way of presenting information became impractical. It was not long before the now contemporary paradigm of search took hold, allowing individuals to dictate their information needs through the issuance of a query.

The processes that take place from query issuance to the returning of results can be considered as the study of contemporary *Information Retrieval (IR)*. As we will discuss throughout the remainder of this chapter, work includes aspects such as the basic components of a retrieval system and approaches used for the evaluation of such systems.

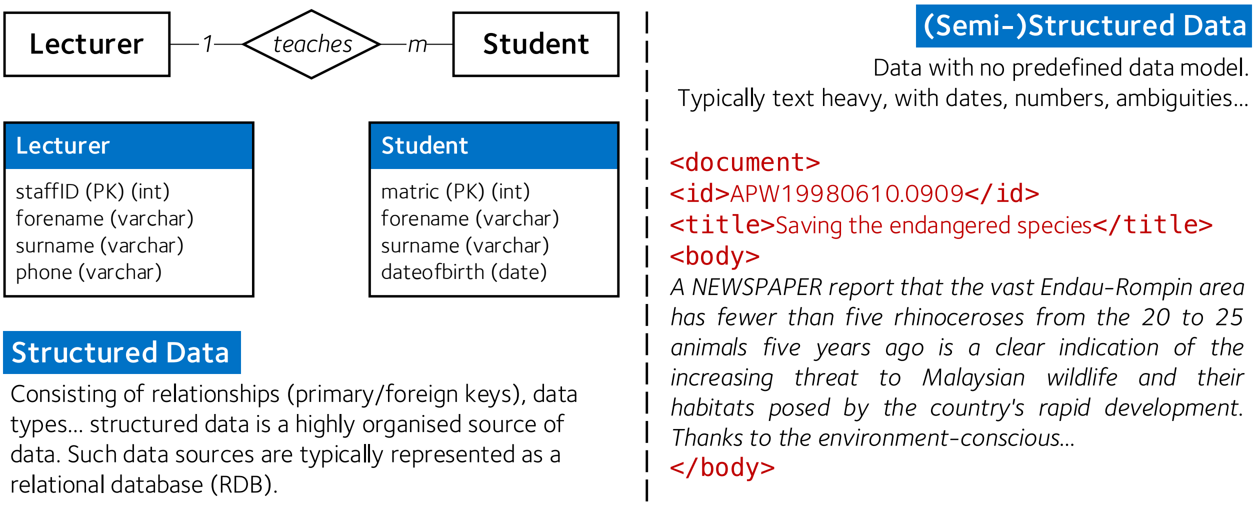
2.2 Information Retrieval Basics

An IR system is expected by the searchers that use it to return results that can be considered **relevant** to their information need. Typically, these results should be **ranked** by decreasing order of relevance. This was originally hypothesised by Luhn (1957), and succinctly expressed by Robertson (1977).

> *"A [reference] retrieval system should rank references in the collection in order of their probability of relevance to the request, or of usefulness to the user, or of satisfying the user."*
>
> Robertson (1977)

Such a system would search through a collection of *unstructured* or *semi-structured* data (such as a collection of web pages or other text documents, or even images or videos, representing *multimedia retrieval*) before returning potential matches to the searcher.

Unstructured and (Semi-)Structured Data A key difference between a traditional database system – or *Relational Database Management System (RDBMS)* – and an IR system is the type of data that they consider. While a RDBMS considers *structured data*, an IR system, in contrast, considers *semi-structured data*, as illustrated in Figure 2.2. With an IR system, such a



Consisting of relationships (primary/foreign keys), data types... structured data is a highly organised source of data. Such data sources are typically represented as a relational database (RDB).

Data with no predefined data model. Typically text heavy, with dates, numbers, ambiguities...

```
<document>
<id>APW19980610.0909</id>
<title>Saving the endangered species</title>
<body>
A NEWSPAPER report that the vast Endau-Rompin area
has fewer than five rhinoceroses from the 20 to 25
animals five years ago is a clear indication of the
increasing threat to Malaysian wildlife and their
habitats posed by the country's rapid development.
Thanks to the environment-conscious...
</body>
```

Figure 2.2 Examples of structured and (semi-)structured data. On the left is a structured RDBMS schema, represented in *compressed Chen notation* (Chen, 1976). Different data types can be specified for each field, representing data in a structured way. On the right is an example of semi-structured data, showing a document from a newswire collection. Note the semi-structured component (containing an identifier and title), and the unstructured body text.

premise for structured data does not exist.[5] Semi-structured data such as an HTML page contains a series of *elements* (e.g. section headers represented within header elements such as <h1>, <h2>, <h3> up to <h6>), but the text within these elements is largely of an unstructured nature. The unstructured data can contain information such as dates or entities (terms describing a real-world object and/or location, such as `canberra` or `dropbear`, and can be (as it is probably written in a natural language) ambiguous. Because of this, examining unstructured data presents a major challenge to researchers.

Being able to effectively sift through large volumes of unstructured data led to the development of retrieval systems. Consisting of a number of key components, the basic process of a retrieval system – along with the inclusion of the users (or searchers) that utilise such systems – can be seen in Figure 2.3. Core to the wider system is the **retrieval engine**, of which many *experimental*[6] retrieval engines can be selected based upon experimentation require-

[5] This may be a slight misnomer; *schemas* can be used for an IR system *index* when considering *fielded retrieval*. For example, a collection of newspaper articles may contain a title and body – but within the fields, the data is unstructured to the retrieval system.

[6] van Rijsbergen (1979) defines a difference between *operational* and *experimental* IR systems. A majority

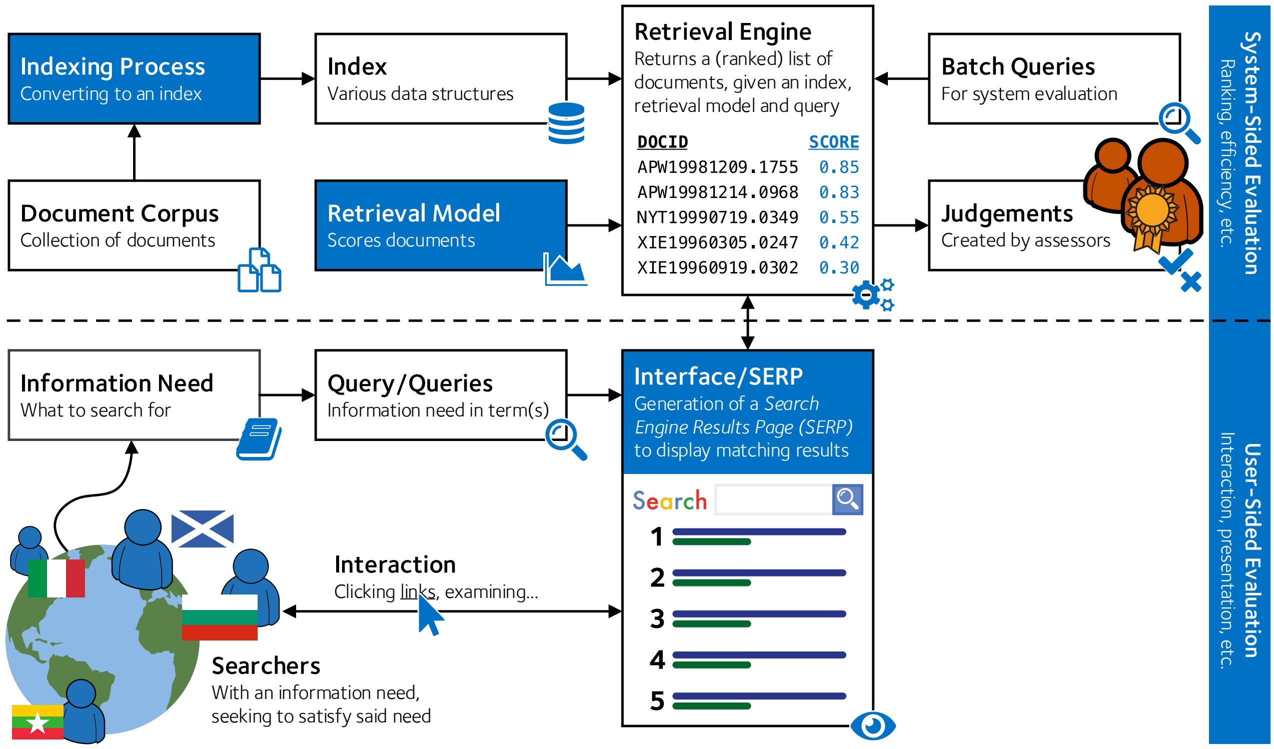


Figure 2.3 The core components of a retrieval system, including the key processes that we discuss in this chapter, highlighted by blue boxes. Central to the discussion in this chapter is the delineation between *system-sided* and *user-sided* evaluation, with both clearly separated in this figure by the dashed line. On the top, system-sided aspects include the *retrieval engine*, *retrieval model* and *index*. Below, user-sided aspects include the *interface*, *interactions* that take place with said interface, and constructs such as the searcher's *information need* and derived *query/queries*.

ments and existing infrastructure available. Examples include *Elasticsearch*, *Lemur/Indri*, *Lucene for IR* (including derived projects such as *Apache Solr*), *Okapi*, the *Terrier IR platform*, *Wumpus* and *Zettair*. Common to all systems are three key inputs, which are:

- an **index** of documents, a specially crafted data structure used for the fast lookup of documents derived from a source collection, or *corpus*;
- a **retrieval model** that scores and identifies documents that may constitute as relevant to what is being searched for; and

of individuals will only ever interact with an operational system (such as Google). The work in this thesis however focuses more on experimental IR systems, and the methodology employed to compare different experimental retrieval systems against each other.

- a query, the construct that represents a given *information need* by a searcher – or one of several queries issued in a *batch environment.*

The retrieval engine combines these inputs to yield an output. This is a ranked list of documents[7] that the retrieval model concludes to be relevant to the given query. This is often called the *matching process.* The retrieval model is responsible for performing the matching of documents from an index. This index typically constitutes a number of different data structures that are generated through the *indexing process*, where a source document corpus is traversed. As highlighted by the blue boxes in Figure 2.3, we discuss the indexing process and various retrieval models in this chapter, explained in Sections 2.2.1 and 2.2.2 respectively. These components are all considered as system-sided aspects of the wider retrieval system, with evaluation of system-sided aspects concerning the quality of returned rankings, how efficient the retrieval engine is, etc.

However, the system-sided aspects only tell part of the story. We build retrieval systems to help searchers satisfy their information need – and hence the study of *Interactive Information Retrieval (IIR)* is devoted to considering the interactions between the searcher and retrieval system. While we discuss more user-sided aspects later on in this chapter (Section 2.3), searchers, given an information need, will issue one or more queries, and *interact* with the presented interface (Ingwersen and Järvelin, 2005), or *Search Engine Results Page (SERP)* (refer to Section 2.3.2.1). Their ultimate goal is to satisfy their said information need. User-sided evaluation is also considered extensively in this thesis. Examples include the examination of the many different interactions that take place, and how the presentation of results affects search behaviours.

Before discussing the user-sided aspects of search, we turn our attention to the *system-sided* components of the wider retrieval system, considering the indexing process and various retrieval models that are commonly employed.

[7] Depending upon the retrieval model used, ranking may or may not occur – refer to Section 2.2.2 for more information.

2.2.1 The Indexing Process

Indexing takes into account the conversion of a collection of documents (or corpus) into a data structure that facilitates fast, full-text search – a key requirement of any retrieval system. Full-text search is typically undertaken in milliseconds, with the goal of finding documents that will be relevant to a given query (and thus information need). The additional storage space and management requirements to maintain an index of documents are considered to be a necessary tradeoff to guarantee timely responses to a searcher's query.

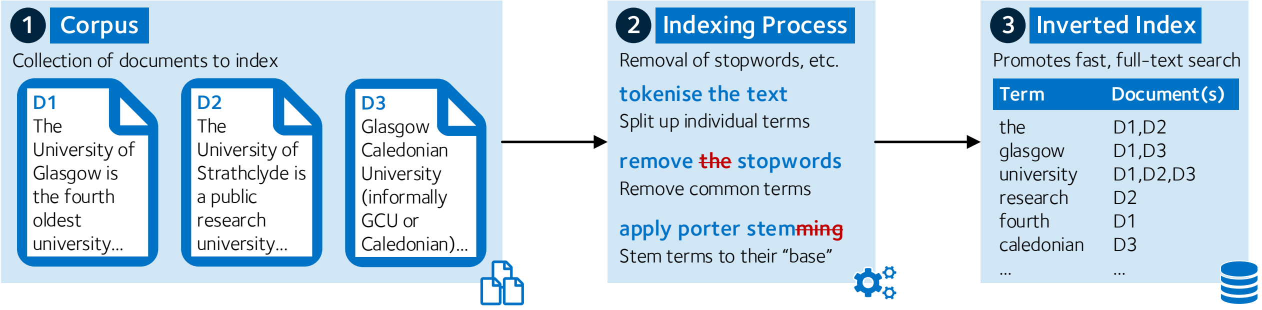
As illustrated in Figure 2.4, the indexing process can be split into three main steps:

1. gathering the corpus of documents to be indexed;
2. performing pre-indexing data preparation; and
3. creating the various data structures that constitute an index.

Experimental corpora are available for use with batch experimentation, typically from various evaluation forums, as discussed in Section 2.3.1.1. For operational retrieval systems, data is collated by other means. For example, web retrieval systems employ a *web crawler* to examine pages on the WWW, and accumulates additional content by following the WWW's hyperlink structure. Google's crawler, *Googlebot,* regularly crawls high impact websites to ensure that the associated index is continually refreshed with up to date information.

An index will contain an entry for each processed document, along with a *vector of terms* that are present within the said document. This is known as the *direct index*[8]. However, a retrieval system needs to support fast full-text search, matching terms from a searcher's query to one or more documents within the index. To support faster query matching, the most simplistic approach is to simply *invert* the index, such that the lookup of the index then corresponds to individual terms, not individual documents. A *vector of documents* can

[8] The *direct index* is sometimes referred to as a *forward index.*


1 Corpus
Collection of documents to index
D1 The University of Glasgow is the fourth oldest university...
D2 The University of Strathclyde is a public research university...
D3 Glasgow Caledonian University (informally GCU or Caledonian)...
2 Indexing Process
Removal of stopwords, etc.
tokenise the text
Split up individual terms
remove ~~the~~ stopwords
Remove common terms
apply porter stem~~ming~~
Stem terms to their "base"
3 Inverted Index
Promotes fast, full-text search

Term	Document(s)
the	D1,D2
glasgow	D1,D3
university	D1,D2,D3
research	D2
fourth	D1
caledonian	D3
...	...

Figure 2.4 An illustration of the main steps to produce an *inverted index*, using a source document collection of three documents as an example. Depending upon the requirements of the IR system, the indexing process may vary; all classical IR systems however rely on an inverted index.

then be provided for each term, yielding much faster access to a potential list of documents. An example of an inverted index is provided at step ❸ in Figure 2.4. The source corpus in this example illustration consists of three documents, with the resultant index shown. The set of documents retrieved can then be sent to a retrieval model for ranking.

Before a document is indexed, a number of pre-indexing steps take place. Three of the most common processes involved include *tokenisation*, *stopword removal* and *stemming*.

2.2.1.1 Tokenisation

Tokenisation is the process of *parsing* a source document and splitting the data within the document into a number of individual *tokens* that may be subsequently indexed. A token is a sequence of grouped characters that provide some semantic meaning.

While we do not go into greater detail about the process of tokenisation, there are many challenges to this process – such as *word boundary ambiguity*. While parsing an English or Latin-based document may be relatively straightforward (with spaces representing *word boundaries*), other language structures (such as 汉语 (Chinese), 日本語 (Japanese), 한국어 (Korean), or ภาษาไทย (Thai)) could present an issue. Considering what words a potential searcher of a retrieval system may use to search with may be a potential pathway for finding a solution to this problem.

2.2.1.2 Stopword Removal

Stopword removal is another popular choice for indexing document collections for use in an experimental IR system. Illustrated in Figure 2.4, extremely common words which would appear to have little value in selecting documents matching a searcher's query (that is, *non-discriminative* words) can simply be removed from a document's vocabulary entirely. Examples of such *stopwords* could be *the*, *a*, or *did*, or even a complete phrase from the famous soliloquy of William Shakespeare's *Hamlet: "to be or not to be"*. Some experiments consider a small list of stopwords, while others consider a larger list. Larger lists often significantly reduce the size of an indexed corpus (Manning et al., 2008). Indeed, it was argued by Fox (1992) that larger lists *"are advisable"*.

While stopword lists may be manually crafted under particular scenarios, automatic extraction from a document corpus is perhaps a more common practice. A simple approach would be to count the *term frequency* for each term within a corpus and sort the resultant list in descending order, selecting some top *k* of the most frequently occurring terms. Readily available lists are also available. van Rijsbergen (1979) for example produced a list of 250 terms, with Francis and Kučera (1985) demonstrating a list of 425 stopwords from the *Brown corpus*[9]. For the experiments detailed in this thesis, *Fox's classical stopword list* (Fox, 1992) is used, consisting of 421 terms. Such an approach may be considered acceptable, but stopwords lists do vary from collection to collection, as stated by Lo et al. (2005).

Issues of course also exist with the removal of stopwords. Removing stopwords from a query may decrease processing time, but what if all terms within a query are stopwords, like the aforementioned soliloquy? The resultant query passed to the retrieval engine could contain zero terms! As such, commercial retrieval systems are *less likely* to employ stopword removal during the indexing process to counter such an occurrence (Manning et al., 2008; Dolamic and Savoy, 2010). Rather, stopword removal may be undertaken on *issued queries*

[9] The *Brown corpus* was a collection of documents representing (then) contemporary American English, compiled by William Francis and Henry Kučera – refer to Francis and Kučera (1979) for more information.

instead (Croft et al., 2009). Techniques such as compression may be used to reduce the size of the index. Queries such as `` `to be or not to be' `` may contain some semantic meaning. Like tokenisation, there is often more to this problem than initially meets the eye.

2.2.1.3 Stemming

Another common pre-indexing process is *stemming*. This is the process of reducing inflected – or sometimes derived – words from their *word stem, base* or *root*. For example, given the terms `fisher`, `fished` and `fishing`, reducing each of these terms to their respective word stem would result in `fish`. Essentially, stemming allows one to group words together with a similar semantical meaning. This provides the advantage of reducing the size of an index, with fewer terms to index. A further benefit may be the potential increase in the number of possible matches that can be found with a stemmed set of query terms, increasing the retrieval system's *recall* (refer to Section 2.4.1.2).

The concept of stemming has been studied since the 1960s, with the *Porter stemmer* (Porter, 1980) emerging over time as empirically the most effective – especially for smaller document collections.[10] Comprised of a series of linguistic rules, the *measure* of a word can be considered as:

> *"loosely checking the number of syllables to see whether a word is long enough that it is reasonable to regard the matching portion of a rule as a suffix rather than as part of the stem of the word."* **(Manning et al., 2008)**

Porter stemming is utilised in the indexing process for the work reported in this thesis; other stemmers do exist, with examples including the original single pass stemmer devised by Lovins (1968), and the Krovetz stemmer (Krovetz, 1993).

[10] The Porter stemming algorithm is not provided in this thesis; refer to Porter (1980) for an in-depth explanation of the algorithm.

Issues such as *overstemming* can impact upon a retrieval system's performance. Here, terms are reduced so far back to the point that meaning is lost, thus negatively affecting the results returned. Terms like `universe`, `university` and `universal` when stemmed will be reduced to `univers`. While the three original terms may be etymologically linked, their modern meanings are however very different. When stemming is applied, documents containing both `universe` and `university` will be returned. While we do not go into depth into the solutions to this problem, two potential workarounds consider: the *lemmatisation* of terms (Manning et al., 2008); and the *n-gram* context of a term, allowing the retrieval system to select the correct meaning (McNamee, 2006). Like stopword removal, stemming is also often applied on issued queries (Croft et al., 2009).

2.2.2 Retrieval Models

Given a generated document index and a searcher's query, the next part of the process is retrieval (or *matching*). For this, a number of mathematically-based *retrieval models* have been developed over the years that attempt to operationalise the notion of relevance. These models provide us with a means for discussion and further refinement. They also provide us with the blueprint from which we operationalise a retrieval system (Hiemstra, 2009). The usefulness of such a model can be subsequently tested via experimentation and evaluation.

Several different types of retrieval model have been defined, ranging from the relatively simplistic to the more complex. More complex approaches not only define a notion of what documents would be considered relevant, but also to what *degree* that is so. This section considers three main retrieval model families, including:

- the *boolean model;*
- the *vector space model;* and
- *probabilistic models.*

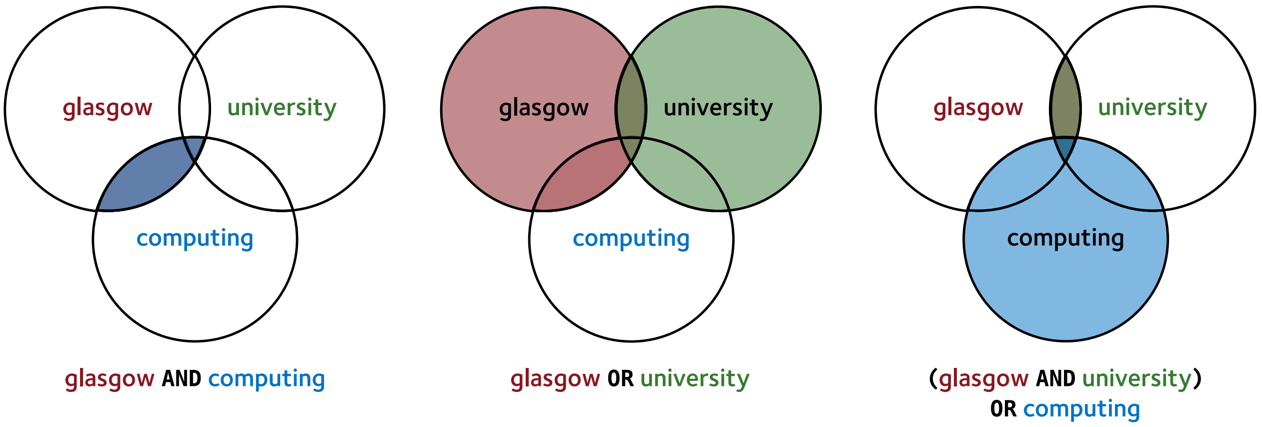
This summary is not exhaustive. A further approach could be a *language model* that considers a probability distribution over a sequence of words (and thus is probabilistic) (Manning et al., 2008). A more contemporary ranking approach would be *neural IR models*. Here, neural networks are used to rank documents in relation to a searcher's query (Mitra and Craswell, 2017). These other approaches are not discussed here. Instead, this section provides a broad overview of models used in this thesis, discussing the benefits and disadvantages of each. We focus the discussion of each retrieval model on how they can potentially influence *stopping behaviours.*

2.2.2.1 Boolean Model

Cited as the first formally defined IR retrieval model, the boolean model is also the most likely one to be criticised (Hiemstra, 2009). The model employs operators of mathematical logic as defined by George Boole (Boole, 1847), or *set theory.* Boole defined three basic operators: `AND`, yielding a logical product between two sets; `OR`, yielding the logical sum between two sets; and `NOT`, yielding the logical difference.

By considering an individual query term and an unambiguous set of documents, logical operations can be applied to retrieve a set of documents. For example, the query term `glasgow` will yield a set of all documents containing the term `glasgow`, yet the query `NOT glasgow` will retrieve the set of documents that *do not* contain any mention of the term `glasgow`. Results of applying logical operators between different sets can be illustrated through a *Venn diagram,* where each set of documents is represented as a disc. Figure 2.5 provides an example of such diagrams, using `glasgow university computing` as an example.

Despite its relative simplicity, there are major limitations to the exact match approach. First, when considering the boolean query, there is no notion of term importance – every term has equal weighting. Issuing a query utilising rules of logic also appears as an unnatural representation of the searcher's information need. Indeed, as an information need becomes



glasgow **AND** computing

glasgow **OR** university

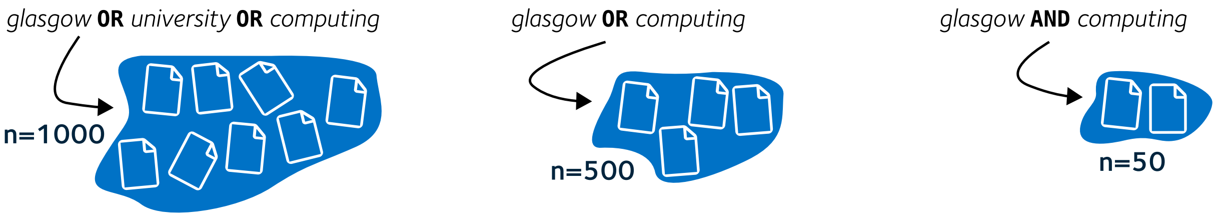
(glasgow **AND** university)
OR computing

Figure 2.5 An example illustration of the boolean retrieval model, using the query terms `glasgow`, `university` and `computing`. Each coloured disc represents the set of documents containing that particular term. In the figure, three Venn diagram examples are provided, demonstrating the key logical operators AND and OR.

more complex, the corresponding boolean query can grow to be disproportionately large and cumbersome. As documents either belong to a set or not, a document is considered to be either relevant (`TRUE`) or not (`FALSE`). As such, one cannot estimate the degree of how relevant a document may be to the searcher's query. Results therefore are provided to the searcher in an unranked manner.

Returning an unranked set of documents would appear as an alien concept to users of contemporary retrieval systems – one would assume that the document presented first would be the document considered to have the greatest relevance, as per the underlying retrieval model. This would make it difficult for a searcher to obtain some notion of how many results he or she should examine before stopping. This is because no ranking means all returned documents are of equal importance.

Instead, a searcher utilising a boolean retrieval system will often find an initial exploratory query will return a large set of documents – too many to examine each in sufficient detail. Rather, what a searcher will likely do is gradually reformulate their query in an iterative manner (Koch et al., 2009) – like in the illustration below – until the document set returned is of a manageable size to process. This is an inherently different kind of stopping behaviour from the examples provided thus far in this thesis which assumes documents are presented in a *ranked* list, with some notion of a *depth* at which a searcher would stop.


glasgow **OR** *university* **OR** *computing*
n=1000
glasgow **OR** *computing*
n=500
glasgow **AND** *computing*
n=50

Despite not being required in contemporary retrieval systems, many systems still do provide support for crafting a boolean query for when returning a good set of results is difficult. Boolean queries may also be of use where ambiguity exists within a searcher's query, and clarification is required to eliminate a set of non-relevant documents (i.e. perhaps using a boolean operator to return a more focused set of results). Indeed, boolean queries still find considerable traction in professional search systems, such as patent search. Here, missing an existing, relevant patent may be incredibly costly – here, *recall* is preferred over *precision*, as discussed in Sections 2.4.1.2 and 2.4.1.1 respectively.

2.2.2.2 Vector Space Model

Further families of retrieval model were later developed to counter the issues and criticisms of boolean retrieval. Luhn (1957) hypothesised that a searcher should prepare a document that is similar to the documents being sought after. By comparing documents against this *representative* document, a retrieval system could then begin to deduce what other documents would be useful, and by *what margin*.

The vector space model proposed by Salton et al. (1975) incorporates the principles as outlined by Luhn (1957). These basic principles are operationalised by representing queries and documents within Euclidean geometry, where both are represented as vectors in multidimensional space. The notion of how close documents appear to each other therefore denotes the relevance of a document.

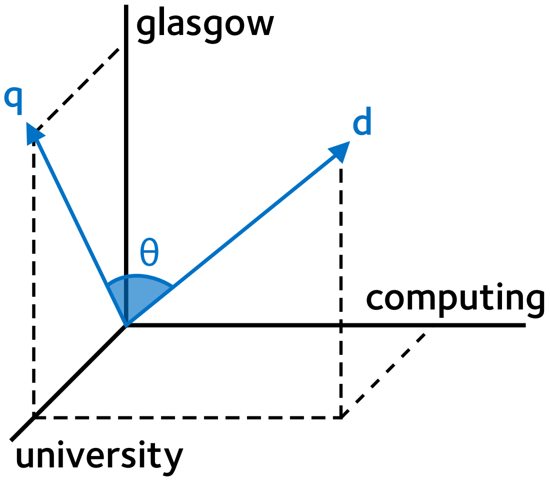


Figure 2.6 An illustration of the vector space model in Euclidean space, with each term representing a dimension. Here, the cosine similarity between query *q* and document *d* is shown.

The vector space model has been very popular as it provides an intuitive means for addressing the overarching problem of a retrieval system. It can also incorporate methods such as *term weighting,* which has been shown to improve retrieval effectiveness (Croft et al., 2009). Furthermore, as queries and terms are represented in Euclidean space, vector similarity methods can be employed to determine relevance. While many approaches have been trialled, empirical evidence has favoured *cosine similarity* (Croft et al., 2009). This is illustrated in Figure 2.6. Using such an approach allows one to then compute the degrees of relevance, meaning that matched documents can be returned in a ranked order. Provision of a ranking then gives cues to searchers interacting with the results list to form an idea of the depth at which examination should stop. However, at what threshold should a searcher stop? As we highlight in the following section, the *Probability Ranking Principle (PRP)* (Robertson, 1977) suggests that such a threshold does exist. If the probability of a document being relevant is greater than the probability of it being non-relevant, a searcher should look at it. Once the probability of a document being non-relevant outweighs the probability of it being relevant, the PRP indicates that a searcher shouldn't invest time examining it. What value this threshold should be is open to interpretation, and will vary from searcher to searcher.

In order to understand the basic workings of the vector space approach, let us consider a query, Q, with each of its constituent terms placed within a term vector in t-dimensional space, leading to $Q = (q_1, q_2, q_3, \ldots, q_{it})$. Consider also a document, D_i, with terms from the document again represented in t-dimensional space, yielding $D_i = (d_{i1}, d_{i2}, d_{i3}, \ldots, d_{it})$. From this notation, d_{ij} represents the *term frequency (TF)* of term j appearing in document i.

With each term represented as a separate dimension within Euclidean space, a weighting scheme can be subsequently applied to emphasise or understate more discriminative or less discriminative terms respectively. By applying weighting schemes, the vector space model ranks documents which promote terms that are more discriminative, thus improving the quality of the returned ranked list.

Term frequency is one of many different term weighting schemes that have been trialled over the years in IR research. A widely used schemes is *inverse document frequency (IDF)*, proposed by Spärck Jones (1972). In the words of its creator, IDF allows for one to define the specificity of a term as *"an inverse function of the number of documents in which it occurs."* (Spärck Jones, 1972). This is useful, as non-discriminative terms that occur frequently within an index (e.g. `the`) would have a small weighting applied, with the inverse happening for more discriminative terms that are better able to return a document.

TF and IDF are typically combined together as a measure of both term appearance and importance, under an approach called *TF-IDF*. For a given term k, one can calculate a TF-IDF score with the following equation:

$$tf_{i,k} \cdot idf_k = \frac{f_{i,k}}{\sum_{j=1}^{t} f_{i,j}} \cdot log\frac{N}{n_k}. \qquad \text{Equation 2.1}$$

Above, $f_{i,k}$ is the frequency of term k, N is the number of documents in the collection used, and n_k is the number of documents in which term k appears at least once.

2.2.2.3 Probabilistic Models

Like the vector space model, probabilistic retrieval models estimate the likelihood of a document being relevant to a given query. One of the most well-known ranking principles, named as the *Probability Ranking Principle (PRP)* in the previous section, was defined by Cooper (1971) and Robertson (1977).

> *"If a reference retrieval system's response to each request is ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."*
>
> Robertson (1977)

Essentially, this states that documents that are considered more likely to be relevant than non-relevant should be retrieved – or where $P(R|D) > P(\overline{R}|D)$. This also implies a cutoff point exists, where probabilities fall below some threshold. While the PRP lays much of the foundation from which probabilistic models have been derived, it does not provide its own concrete implementation of such a model.

A simple approach that implements the PRP is the *Binary Independence Model (BIM).* Simple assumptions that make the implementation of the PRP straightforward are employed. As the name of the model suggests, one such assumption is the notion of binary relevance in term vectors for a given document (i.e. a term either exists in a document or not). A second assumption is that terms are modelled as occurring in documents independently, with no association between terms (represented as a *bag of words*) (Manning et al., 2008).

With the BIM originally designed for documents fairly consistent in length, contemporary corpora have a large variance in term frequencies and document lengths. *Okapi BM25* presented by Robertson et al. (1995) was devised as a way of building a probabilistic retrieval model that combined term frequencies, inverse document frequencies and document lengths (of a document and the average document lengths of documents within a corpus) (Jones et al., 2000). BM25 has had considerable impact upon the IR community, and is still used extensively today. BM25 provides a solid baseline for contemporary research and is the retrieval model employed in the experimentation discussed in this thesis, primarily selected for its effectiveness and popularity.

For a given query Q containing keywords $q1, \cdots, q_n$, the BM25 score of a document D is defined as:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - \beta + \beta \cdot \frac{|D|}{avgdl})},$$

Equation 2.2

where $f(q_i, D)$ represents q_i's term frequency within document D, $|D|$ is the length of document D (represented by the number of terms), and $avgdl$ is the average document length in the corpus of documents used. β and k_1 are free parameters, usually set to $\beta = 0.75$ and $k_1 = 1.2$.[11] Refer to Robertson and Zaragoza (2009) for a more thorough review of probabilistic models – and of BM25 in particular.

Like the vector space model, retrieval models implementing the PRP ranks documents with respect to the issued query. Therefore, these models provide searchers with a gauge as to how relevant a document can be. They intrinsically provide a cue as to the depth at which a searcher should stop (when documents are ordered by decreasing relevance).

2.3 From System to Searcher

So far in this chapter, we have provided a background on several aspects in the field of IR. These developments are focused exclusively on the *system*. Recall however that the purpose of a retrieval system is to satisfy the information needs of the *searcher* (or user) using it. Satisfying this information need is key to any successful retrieval system.

In this section, we discuss a line of research that moves from considering the system to a more extensive examination of the searcher, and his or her interactions with a retrieval system. This is examined in the study of *Interactive Information Retrieval (IIR)*. However, before discussing the IIR process, we must first consider in more detail the paradigms that

[11] The values of β and k_1 reported here are the values used throughout experimentation presented in this thesis.

have been extensively used in traditional, *system-sided* IR research. These paradigms have been one of the cornerstones of IR's scientific methodology for many decades, and are mostly considered to be naïve of a searcher's behaviour. After discussing these paradigms, we then move to our discussion of IIR, emphasising the *spectrum* of research between the system-sided (IR) and user-sided (IIR) extremes. This then leads onto the concept of *searcher models* that attempt to capture the high-level, cognitive processes that searchers undertake.

2.3.1 Experimental Paradigms

The methodology behind the majority of classical IR research has focused on the *Cranfield paradigm*. Developed at Cranfield University in Bedfordshire, England, the experimental paradigm is based upon the *Cranfield II* experiments (Cleverdon et al., 1966). The goal of these experiments was to create:

> *"a laboratory type situation where, freed as far as possible from the contamination of operational variables, the performance of index languages could be considered in isolation."* Cleverdon (1991)

The experimental paradigm required the same set of documents, and the same set of information needs to be used for each language, and the use of common IR measures, *precision* and *recall* (refer to Sections 2.4.1.1 and 2.4.1.2 respectively) to be used to measure a given retrieval system's effectiveness. Core to these experiments was also the notion of a test collection, itself consisting of three key components:

- the corpus (collection of documents) to be used;
- the statements of different information needs hereafter referred to as topics; and
- a set of relevance judgements – a list of documents considered relevant that *should* be retrieved by the retrieval system for each topic trialled.

Given these three components, the Cranfield experiments made a number of major simplifying assumptions, as outlined by Voorhees (2002). The first considers the notion of *topical similarity,* by which their relevance is approximated. In short, all relevant documents are equally desirable, and the relevance of one given document is independent of the relevance of any other document. This also leads to the notion of a *static information need*. Under Cranfield, there is assumed to be no change during the search process for what the searcher is searching. Additionally, the single set of relevance judgements provided as part of the test collection is to be considered to be *representative of an entire population.* This means that for a given topic, every searcher will seek to find the same set of relevant documents. Finally, the list of relevant documents for a given topic is assumed to be *total and complete,* i.e. all documents relevant to a topic have been identified and are listed.

2.3.1.1 The Text REtrieval Conference

A number of different *evaluation forums* have been derived from the Cranfield experimental paradigm, utilising many of the implicit assumptions. These forums promote the development of IR as a field, fostering a drive to develop improvements in the various retrieval models and other retrieval system components. Examples of evaluation forums include *NTCIR* (Kando et al., 1999), *CLEF* (Peters and Braschler, 2001) and *INEX* (Fuhr and Lalmas, 2006). However, one of the most well-known evaluation forums is the U.S. Government funded, NIST sponsored *Text REtrieval Conference (TREC)* (Harman, 1993). Experimentation following the TREC approach is hereafter referred to as TREC-style in this thesis.

TREC provides a platform for annual collaboration between research groups interested in different aspects of IR research. Each year, a series of TREC *tracks* are defined, with each consisting of a test collection, in turn consisting of the three components defined above. Within each track is a set of tasks. Some of the tasks, such as those belonging to the *TREC Interactive Track* (Over, 2001), are known as ad-hoc. This type of task can be considered as one of the most obvious for search, where a searcher, perhaps through curiosity, develops an information need in an ad-hoc fashion. They then begin the search process by issuing an exploratory query to a retrieval system.

These tasks are used in conjunction with the relevance judgements provided by assessors. Assessors are usually employees of NIST (Robertson, 2008), who were in turn previously employed as news analysts by various U.S. security agencies. A series of documents are extracted from the document collection using a simple query (a process called *pooling*). Due to the potentially large size of document collections, pooling is an acceptable solution to reducing the number of documents to be examined. As an example, given the topic *wildlife extinction,* the query `'wildlife extinction'` is issued over a number of different retrieval systems. Documents returned are pooled together and then judged by the assessors.

For many TREC tracks, judgements are binary, with `0` denoting non-relevance, and `1` denoting relevance. Graded relevance judgements can also be used – the initial Cranfield II experiments, for example, used a five-point relevance scale (Voorhees, 2002). Pooling can mean that documents that are potentially relevant can be missed by assessors, and thus will receive no judgement (Keenan et al., 2001).

Institutions wishing to participate in a track receive the associated test collection and tasks. They then index the corpus and run their experimental retrieval system over the provided material. Experiments are typically run over 25-50 different topics (Voorhees, 2002), with a solitary query issued for each (the topic's *title*). These are typically executed in a batch environment, with a large number of results (typically 1, 000) returned from the retrieval engine. Output from the experiments is then produced in a standardised format. Results can then be used in conjunction with the judgements (termed *Query REelevance Judgements,* or QRELs), and fed into a standardised program called `trec_eval`[12] to perform an evaluation of the runs that have been undertaken. The application returns the values for a number of common system-sided evaluation measures, some of which are discussed in Section 2.4.1.

[12] `trec_eval` is downloadable from `http://trec.nist.gov/trec_eval/` **LA** *2018-03-08*. Version 8.1 of the software was used for computing most of the evaluation measures reported in this thesis.

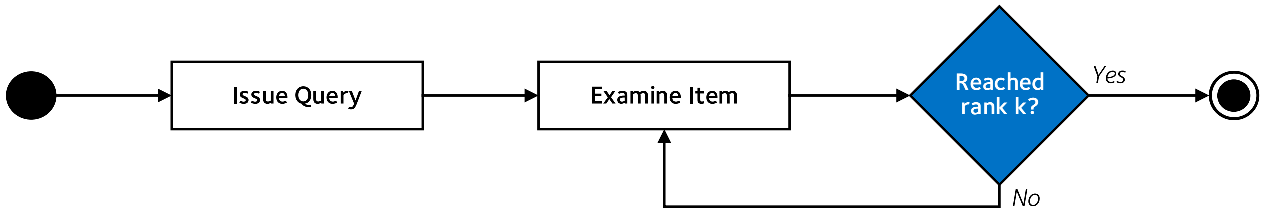


Figure 2.7 The TREC searcher model. Considering a highly abstracted searcher, a query is issued **(Issue Query)**, with each individual result examined in a linear order **(Examine Item)**, up to some depth *k* (typically 1,000), before the searcher stops. All documents up to rank *k* are assessed; no documents are skipped during this process.

2.3.1.2 The TREC Searcher Model

Given the assumptions of the Cranfield paradigm and subsequent evaluation forums such as TREC, one may be forgiven for thinking that the searcher – the target audience of any retrieval system – has been completely ignored from the process. While it is true that the paradigm focuses primarily on system-sided evaluation of retrieval systems, a searcher *is* considered – just in a highly abstracted form. The assumption that a searcher's information does not vary as they search is just one example of an abstraction from reality – a searcher's information need is *dynamic* and typically evolves as they search (Borlund, 2003).

The basic searcher model employed by a TREC-style experiment is illustrated as a flowchart in Figure 2.7. Given the batch-style nature of TREC-style experimentation, this particular searcher model is well suited to such an environment, as the simplifying assumptions and complete lack of interaction from the searcher go hand in hand with the design goals of the initial Cranfield II experiments. When subscribing to this model, a searcher, given an information need (or topic), will issue a single query pertaining to the said information need. This query is simply the topic's title (e.g. `wildlife extinction`). From there, the searcher will then examine each individual document in a linear fashion. This process continues until some rank *k*, at which point the searcher will cease and the search process ends. This rank *k* is typically set to $1,000$ to provide evaluation programs such as `trec_eval` with a large a set of rankings as possible for the various evaluation measures to be computed accurately.

With this highly abstracted searcher model being agonistic of the complex interactions that take place during search, a number of different criticisms can be made. Below, we enumerate on three primary criticisms that have been discussed in the literature.

- **A Single Query** The TREC searcher model assumes that a single query is issued for a given information need. This severely limits the potential for interaction between the searcher and retrieval system – a single query means no *query reformulation* is possible, for example. In reality, searchers *do* reformulate queries, issuing multiple queries during a search session (Keskustalo et al., 2009).
- **Assuming a Fixed Depth** Searchers subscribing to this searcher model will always examine documents up to a depth of k. This assumes a *fixed-depth stopping strategy,* where searchers are agnostic of the results as presented to them. In reality, searchers adapt their stopping depths dependent upon a variety of different factors, such as the number of non-relevant items uncovered thus far (Cooper, 1973b).
- **All Documents are Inspected** The final key, limiting assumption in the TREC searcher model is that *all* documents to depth k are assessed. In reality, searchers may skim through results, or simply decide that a document does not look to be promising, and thus skip it. In the TREC searcher model, there is no concept of a *result summary,* a shorthand overview of the contents of the document. Result summaries are typically expected to be part of the interface of a contemporary retrieval system.

Over time, researchers have begun to examine ways in which to improve upon the basic, rigid assumptions laid out by this searcher model. For example, Smucker and Clarke (2012) introduced *time-biased gain,* where probabilities of interacting with documents were included. Work by Tran et al. (2017) considered the TREC searcher model from the standpoint of a *Markov model.* As such, this representation of the search process was complemented with a series of different *transition probabilities,* dictating the likelihood of a searcher switching from one *state* to another.

These works can be considered as a means of including the complex interactions that take place between the searcher and retrieval system within a searcher model, or consideration of the wider IIR process.

2.3.2 Interactive Information Retrieval

The study of *Interactive Information Retrieval (IIR)* attempts to address our lack of understanding of a searcher's behaviours and interactions, and incorporate new findings into the evaluation of retrieval systems (Callan et al., 2007). IIR studies can include aspects from both user-sided and system-sided research. For example, one might present the results of a user study examining a particular phenomenon of a searcher's behaviour, and also provide details of a system-sided evaluation. As discussed by Kelly (2009), IIR can trace its roots back to a variety of different disciplines, including: traditional IR (i.e. exclusively system-sided research); library and information sciences; psychology; and *Human-Computer Interaction (HCI)*. Typically presented as a branch of IR and/or HCI, arguments also exist to consider IIR as a distinct area of research (Ruthven, 2008).

> *"In IIR, users are typically studied along with their interactions with systems and information. While classic IR studies abstract humans out of the evaluation model, IIR focuses on users' behaviors [sic] and experiences – including physical, cognitive and affective – and the interactions that occur between users and systems, and users and information. In simple terms, classic IR evaluation asks the question, does this system retrieve relevant documents? IIR evaluation asks the question, can people use this system to retrieve relevant documents?"* **Kelly (2009)**

To address the question of whether *people can use a retrieval system*, we begin by examining the wider IIR process. Figure 2.3 on page 27 considered a number of user-focused aspects, as illustrated in Figure 2.8. Given some phenomenon in the natural world (perhaps by observation, reading a book, or through conversation with another human), a searcher will

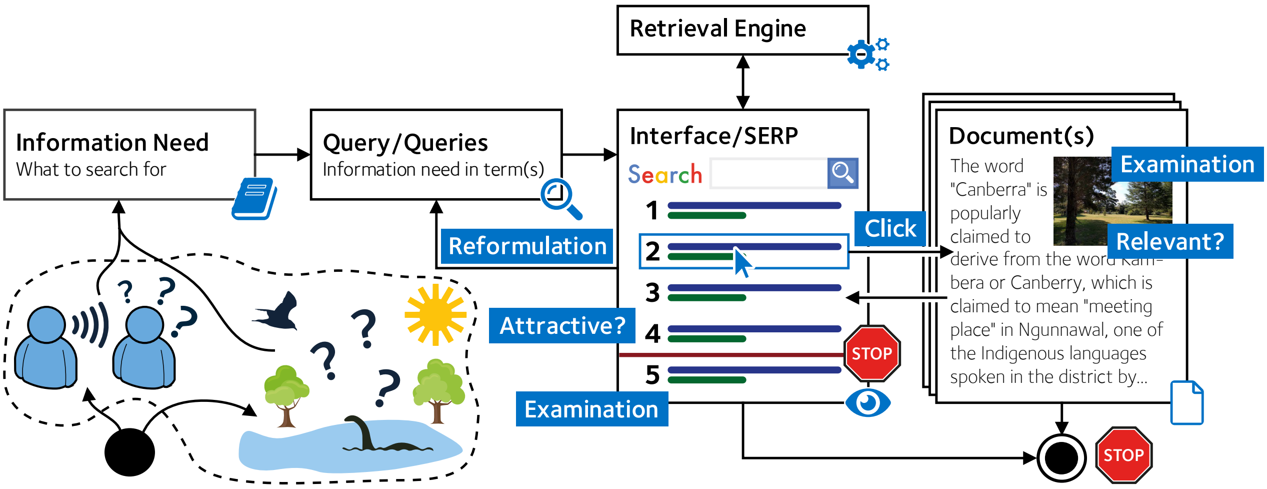

Retrieval Engine
Information Need
What to search for
Query/Queries
Information need in term(s)
Interface/SERP
Search
1
2
3
4
5
Reformulation
Click
Attractive?
STOP
Examination
Document(s)
The word "Canberra" is popularly claimed to derive from the word Kam-bera or Canberry, which is claimed to mean "meeting place" in Ngunnawal, one of the Indigenous languages spoken in the district by...
Examination
Relevant?
STOP

Figure 2.8 The basics of the IIR process, complete with a number of different searcher interactions (as highlighted) that can take place (although this illustration may not be exhaustive). Forming an information need, searchers then begin the interaction process by issuing a query, before examining content presented on the SERP. At given points, searchers may then *stop*.

then begin to formulate an information need. As discussed previously in this thesis, this information need can arise from a knowledge gap in the searcher's mind, an internal inconsistency in what they are experiencing, or a conflict of evidence. In an *Anomalous State of Knowledge (ASK)* (Belkin, 1980), the searcher will then begin the IIR process, with the aim of satisfying their (perhaps vague) information need.

Upon bringing up the interface of a retrieval system, the searcher begins their so-called search session, which begins with the formulation of the information need as a query. Once the query has been submitted, a complex series of interactions begin to take place between the system and the searcher, with these interactions being of great importance to those studying IIR. Results will be retrieved by the underlying retrieval system and presented to the searcher in the form of a *Search Engine Results Page (SERP)*. While we discuss the SERP in more detail in Section 2.3.2.1, it should be noted that a majority of the interactions that take place occur on the SERP.

However, interactions may be dependent upon the searcher's *intent* – that is, what they are aiming to achieve through the satisfaction of their information need. Three intents typically used are *navigational, transactional* and *informational* (Jansen et al., 2008). Navigational intents for example simply refer to the notion of using a retrieval system to navigate the searcher to some *Uniform Resource Locator (URL)*. This means that the searcher will simply need to click the link to satisfy their information need.[13]

Informational intents will undoubtedly require a greater degree of interaction with the retrieval system.[14] For example, searchers will begin to examine the content on the SERP, inspecting individual summaries for potential relevance to their information need. At each stage, the searcher is continually learning, and thus the interaction cycle may prompt the searcher, for example, to provide a *query reformulation* as they begin to develop their underlying mental model of the topic. A revised SERP may then begin to yield more promising results. As the searcher examines these updated results, he or she may find that a particular summary is deemed sufficiently attractive to investigate further, and thus clicks on the provided link. Taking the searcher to the corresponding document, the searcher can then examine the document in more detail, and make a decision as to its relevance. If not satisfied, the searcher may navigate back to the SERP, and continue their examination of further results. At some point, the searcher will make a decision to stop their interactions – either within a given SERP, or the search session as a whole (both are illustrated in Figure 2.8 with stop signs). Stopping may occur for example if a searcher has satisfied their information need, has been frustrated with the retrieval system's inability to return relevant results (Cooper, 1973b), or from a variety of different factors external to the search process, such as time pressure.

While this example above may be highly abstract in nature, it clearly shows that the search process is extremely complex and *inherently interactive* (Ingwersen and Järvelin, 2005). Thus, work in the field of IIR provides a basis for developing more complex, *realistic* models of the search process, improving upon the TREC-style searcher model.

[13] A searcher intent analysis by Jansen et al. (2008) showed that approximately 10% of queries issued to web retrieval systems were either navigational or transactional in nature.

[14] One could argue that an informational intent could now be satisfied through *information cards* presented on contemporary web retrieval systems. Models and measures at present typically do not consider the presence of these components.

2.3.2.1 The Search Engine Results Page

Core to the experience of a searcher when using a given retrieval system are the interactions that take place on its SERPs. Figure 2.9 depicts an example SERP of the fictional retrieval system, Search.[15] The illustration highlights several key SERP components that are extensively referred to in subsequent parts of this thesis. At the top of the SERP is the **query box**, which allows a searcher to enter (and reformulate) queries as and when they require.

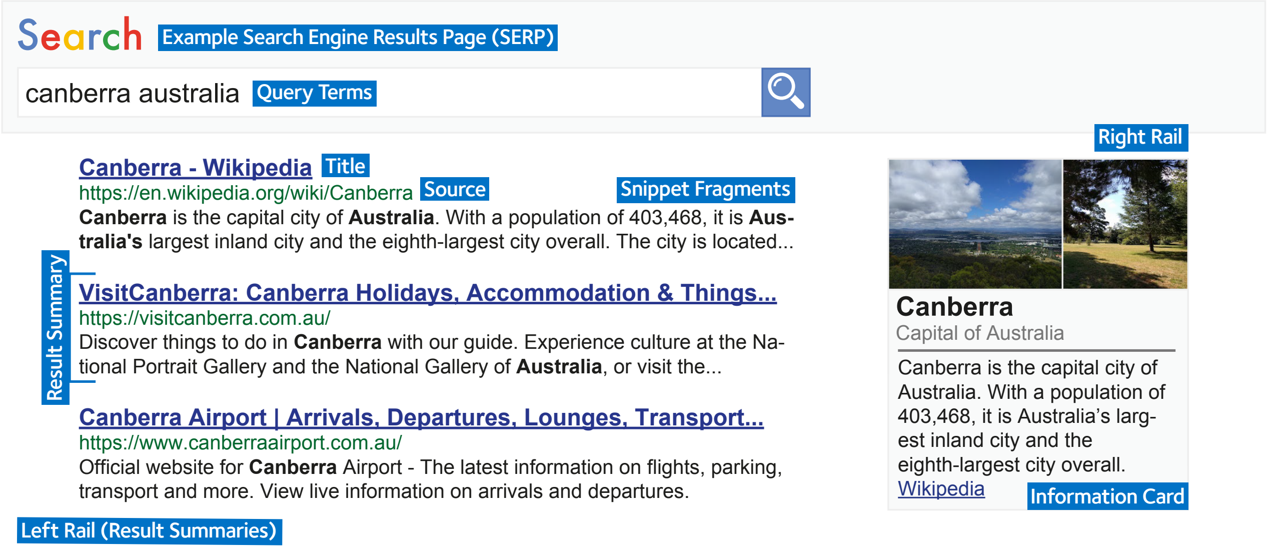
The main body of the SERP is then divided up into the left rail and right rail. Contemporary SERPs utilise the right rail to display additional components such as *information cards* as illustrated in Figure 2.9, with contemporary SERPs thus becoming more and more complex in nature. We however in this thesis exclusively consider simplified SERPs *without* the right rail, or SERPs comprised entirely of *result summaries,* as shown on the left of Figure 2.9.

These result summaries are typically displayed on a SERP as the *ten blue links* (Hearst, 2009), or the first ten ranked results. The document that is judged to be most relevant – as defined by the underlying retrieval model that ranks them – is displayed at the top of the ranked list. These result summaries are short summaries of the corresponding document, and consist of three main components:

- a **title** that represents the title, or headline, of a source document;
- one or more **textual snippets**, providing a summary of the source document such that searchers can determine whether it is worth examining the document in more detail; and
- a **source** for the document, typically an URL if the object is WWW-based.

Snippets are of particular interest to the work in this thesis; we explore the effect of their length on stopping behaviours in Chapter 7. Snippets are typically presented in contempo-

[15] As highlighted in the front matter, we utilise Search throughout this thesis to illustrate various concepts.



Search

canberra australia

Canberra - Wikipedia
https://en.wikipedia.org/wiki/Canberra
Canberra is the capital city of **Australia**. With a population of 403,468, it is **Australia's** largest inland city and the eighth-largest city overall. The city is located...

VisitCanberra: Canberra Holidays, Accommodation & Things...
https://visitcanberra.com.au/
Discover things to do in **Canberra** with our guide. Experience culture at the National Portrait Gallery and the National Gallery of **Australia**, or visit the...

Canberra Airport | Arrivals, Departures, Lounges, Transport...
https://www.canberraairport.com.au/
Official website for **Canberra** Airport - The latest information on flights, parking, transport and more. View live information on arrivals and departures.

Canberra
Capital of Australia
Canberra is the capital city of Australia. With a population of 403,468, it is Australia's largest inland city and the eighth-largest city overall.
Wikipedia

Figure 2.9 An example of a *Search Engine Results Page (SERP)* for the query `canberra australia`. Labels illustrate the names of the key components of a SERP. Of particular relevance to the work in this thesis are the *result summaries*, shown on the left rail.

rary retrieval systems as *query-biased* (Tombros and Sanderson, 1998). This means that the snippet text relates to terms that were present in the searcher's query. Figure 2.9 demonstrates this with the use of **bolded** terms in example snippet text. This approach is alternative to the simple technique of displaying the first sentence or n characters from a document as part of the result summary. Section 7.1.4 on page 195 provides more detail on how snippet text is generated for result summaries.

2.3.3 The IR/IIR Spectrum

In order to aptly describe where IIR fits into the system-sided and user-sided space, Kelly (2009) provided an intuitive spectrum of work that bridges IR and IIR. Figure 2.10 illustrates the spectrum, consisting of eight different categories of study. Moving from left to right in the illustration, categories shift from solely system-sided (TREC-style) studies towards those that are more user-focused, considering a searcher's behaviours when interacting with a retrieval system. Below, we detail four key category types as outlined by Kelly (2009) that have particular relevance to the work detailed in this thesis.

❶ **TREC-Style Studies** With a majority of traditional IR studies falling into this category, TREC-style studies focus upon the development and evaluation of system-sided research, such as retrieval models and indexing techniques. No real searchers are included *in the loop* with this approach, although a simplistic, abstracted searcher model is encoded, as previously discussed in Section 2.3.1.2. While assessors do create relevance judgements used for evaluation, they are not involved in the actual batch-style search process. As such, interaction is assumed to be very simplistic, with a single query issued, for example. This is illustrated in Figure 2.7 on page 44.

❷ **'User' Relevance Assessment Studies** The second category of study does explicitly consider a human *in the loop,* but only in exceptionally limited circumstances. As the name of the category suggests, the humans that are employed for this category of study are used only for generating relevance assessments, perhaps because a specific corpus is used, and no pre-existing TREC relevance assessments are available.

❺ **TREC Interactive Studies** Studies belonging to this category typically are used to evaluate a retrieval system and/or a feature of its user interface, where an experimental retrieval system is used. Typically, aspects such as searcher behaviours, their cognition or the information seeking context are examined. These studies usually aim to assist in aiding our understanding of the search process, and the development of more intuitive and user-friendly search interfaces. Interaction is considered – search sessions in this category of study permit searchers to issue multiple queries through query reformulation and conduct a number of other interactions that studies closer to the left of Figure 2.10 simply do not cater for. These studies are therefore more *realistic,* and are thought to more accurately represent real-world searcher behaviours.

Moving to the right of the spectrum, the study of the searcher becomes ever more prominent, until we reach the following category that considers the *experiences* of searchers.

❽ **Information Seeking Behaviour in Context Studies** In this final category of study, re-

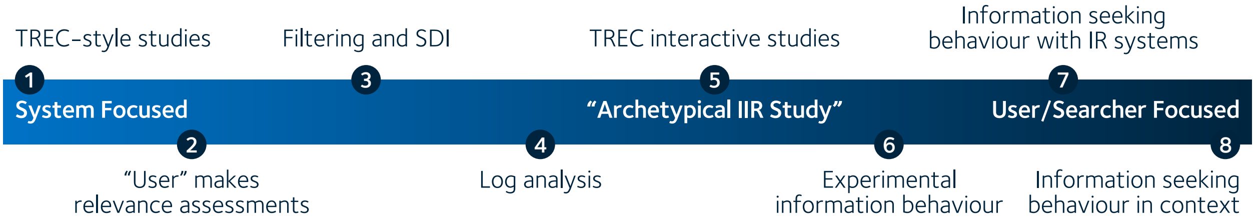

TREC-style studies
Filtering and SDI
TREC interactive studies
Information seeking behaviour with IR systems
1
3
5
7
System Focused
"Archetypical IIR Study"
User/Searcher Focused
2
4
6
8
"User" makes relevance assessments
Log analysis
Experimental information behaviour
Information seeking behaviour in context

Figure 2.10 The spectrum of conceptualising IIR research. Methods on the left consider a more system-focused approach, with those on the right considering a more user-focused approach. The fifth step within the spectrum considering TREC interactive studies is considered to be an *"archetypical IIR study"*. Figure adapted from Kelly (2009), with support of the author.

> searchers consider the information needs of individuals. Researchers would typically observe how individuals conduct their searches, while at the same time collating qualitative data about their differing experiences. These data can then be used to motivate iterations of the design and presentation of search results, thus improving the overall experience for searchers.

The studies represented by Kelly (2009) as category ❺, called *archetypical IIR studies,* are the type of study that we largely consider in this thesis. Indeed, in subsequent chapters, we consider a variety of different aspects that can influence the behaviour and performance of searchers – particularly in relation to their stopping behaviours. This work is done in combination with a series of experiments that can be considered to belong to category ❶. By grounding these experiments with data collected from studies in category ❺, we are able to instantiate more realistic, credible abstractions of the search process which consider some aspects of interaction. We achieve this by simulating the behaviour of real-world searchers to examine what happens to searcher behaviour in different contexts.

2.3.4 The Simulation of Interaction

Simulation is defined as the *imitation of the operation of a real-world process or system over time* (Banks et al., 1996). Such an approach allows one to gain insight into the functioning of some real-world phenomenon, such as the complex interactions that take place during the IIR process. *Computerised simulation* (Heermann, 1990) has become more commonplace today with increasing computational power allowing for the development of ever more complex and realistic simulations. Simulation permits one to solve a large number of problems without resorting to a *"bag of tricks"* (Fishwick, 1995), where special purpose (and often arcane) solutions must be used. Such an example would be a series of linear equations (Fishwick, 1995). With this *closed-form* approach, underlying assumptions can become twisted and altered, drifting the representation further away from the real-world phenomenon that is being modelled.

Simulation avoids such issues by providing one with the freedom and flexibility to reduce the above assumptions. This permits a rapid means of exploring different scenarios, all at a low cost. Additionally, without needing to consider issues such as subject fatigue (within a user study, for example), simulation provides the capability of running experiments with reproducible results (Azzopardi et al., 2011).

Simulation has been employed extensively within classical IR experimentation.[16] TREC-style experimentation may be considered as a form of simulation, where the simple searcher model discussed in Section 2.3.1.2 is used to simulate the searcher's interactions. We consider in this thesis the *simulation of interaction,* where one attempts to mimic behaviours that a searcher exhibits when interacting with a retrieval system (Azzopardi et al., 2011). This means that we can explore different searcher behaviours, methods and techniques to better understand how searchers do, could, or are likely to behave. However, such simulations are questionable and open to criticism if they are not properly motivated, grounded and validated. Therefore, there is a pertinent need to ensure that such simulations are credible abstractions of the search process and that they are seeded with data based on actual human interaction data (study category ❺, as per Section 2.3.3) (Azzopardi et al., 2011).[17]

[16] For an in-depth discussion of various classical IR simulations, refer to Heine (1981).

[17] An exception to this rule would be the exploration of *what-if* scenarios, allowing researchers to examine *what* would happen to a searcher's behaviour *if* a particular scenario were to be applied. One example of such a study is that by Azzopardi (2011). Indeed, we employ such an approach in contributory chapters of this thesis, where we examine many *what-if* scenarios.

Such grounding can, for example, permit *stochastic simulations,* working on the notion of the *probabilities of interaction* (e.g. the probability of clicking on an attractive result summary).

Within the wider IIR process, a number of different individual components have been examined through the use of simulation. These have often been independently analysed from one another (Azzopardi et al., 2011), with examples of different components and associated studies listed below.

- **Query Formulation and Suggestions** This component considers querying, including the generation of pseudo-realistic queries and the development of realistic *querying strategies,* grounded upon the querying behaviours of real-world subjects. Examples include: Azzopardi (2009); Azzopardi et al. (2007); Carterette et al. (2015); Jordan et al. (2006); Keskustalo et al. (2009); Hagen et al. (2015); and Verberne et al. (2015).

- **Browsing Behaviours** This broader component considers the wider behaviours of searchers when examining content, including aspects such as click models, and different browsing strategies employed by searchers (e.g. can a searcher's behaviour be classified as *fast and liberal,* or *slow and neutral?* (Smucker, 2011)). In addition to the work by Smucker (2011), examples include: Carterette et al. (2015); Chuklin et al. (2015); Guo et al. (2009); and Pääkkönen et al. (2015).

- **The Influences of Cost and Time** This component examines how varying *interaction costs* (i.e. the cost of issuing a query or examining a document, considered primarily in terms of the amount of time required) and time constraints can influence the behaviour of searchers. Examples include the economics-based approach outlined by Azzopardi (2011), and work by Baskaya et al. (2012).

- **Performance over Search Sessions** This area of work considers how, when different aspects are changed, the performance of a searcher over a search session varies. Examples include work by Luo et al. (2014) and Luo et al. (2015).

Of particular relevance to this thesis is the work that has been undertaken to examine a searcher's stopping behaviour, with examples including: Carterette et al. (2015); Maxwell et al. (2015a,b); and Thomas et al. (2014). In these works, different stopping strategies and stopping models are proposed. These are considered in depth later in Chapter 3.

Simulation provides a means for examining the aforementioned components from two different standpoints:

1. considering each of the individual components in isolation, as discussed above (e.g. exclusively examining querying behaviours); or
2. considering the interactive search process as a whole, and attempting to capture and model an entire search session, from querying to document examination.

The work in this thesis considers 2, meaning that potential influences of components over others can be considered. This also justifies the need for a more advanced searcher model that captures the interactions of the wider search process. A model is a key component of a simulation or the representation of the real-world phenomenon being simulated (Tocher, 1963). With the TREC-style searcher model outlined, we now turn our attention to considering more advanced searcher models of the IIR process.

2.3.5 Searcher Models

Models of the search process attempt to capture the high level, cognitive processes that a searcher undertakes during a search session – such as issuing a query, or examining a document for relevance. As we have discussed previously, highly abstracted searcher models have been present in system-sided IR research, as well as implicitly encoded within a variety of different evaluation measures, as discussed in Section 2.4.

Carterette et al. (2011) argues that the measures widely used within IR research are themselves typically comprised of three distinct underlying models. These are listed below.

- **Browsing Model** Browsing models describe how a searcher interacts with a retrieval system's interface, including the SERPs that are presented to them.
- **Document Utility Model** These models encode some form of document utility that provides a description of how utility or *gain* can be derived from relevant documents that are examined.
- **Utility Accumulation Model** This final model describes how a searcher accumulates the said utility over the course of an entire search session.

Of particular interest to the work in this thesis are the **browsing models** that attempt to capture the broad array of interactions that take place between the searcher and the retrieval system. The TREC searcher model is a highly abstracted example of such a model, yet there is a disconnect between the assumptions made in this searcher model and reality. As such, we now discuss a number of more advanced searcher models, providing better representations of the complex interactions that take place during IIR.

Seminal work undertaken by Baskaya et al. (2013) presented an explicit, revised searcher model that improved upon the *interactive* capabilities that could be exploited. Improvements over the TREC-style searcher model included the following:

- the ability to **separately judge result summaries** presented on a SERP from the documents associated with each;
- giving a searcher the ability to **stop at a variable depth on the SERP**, avoiding fixed depth stopping behaviours as typically employed in searcher models; and
- permitting a searcher to **issue multiple queries** during a search session, lifting one of the major constraints of the TREC searcher model.

These improvements are demonstrated as a Markov model, illustrated in Figure 2.11. Compared to the TREC-style searcher model illustrated in Figure 2.7 on page 44, the model can be visually seen as more complex, consisting of six main actions.

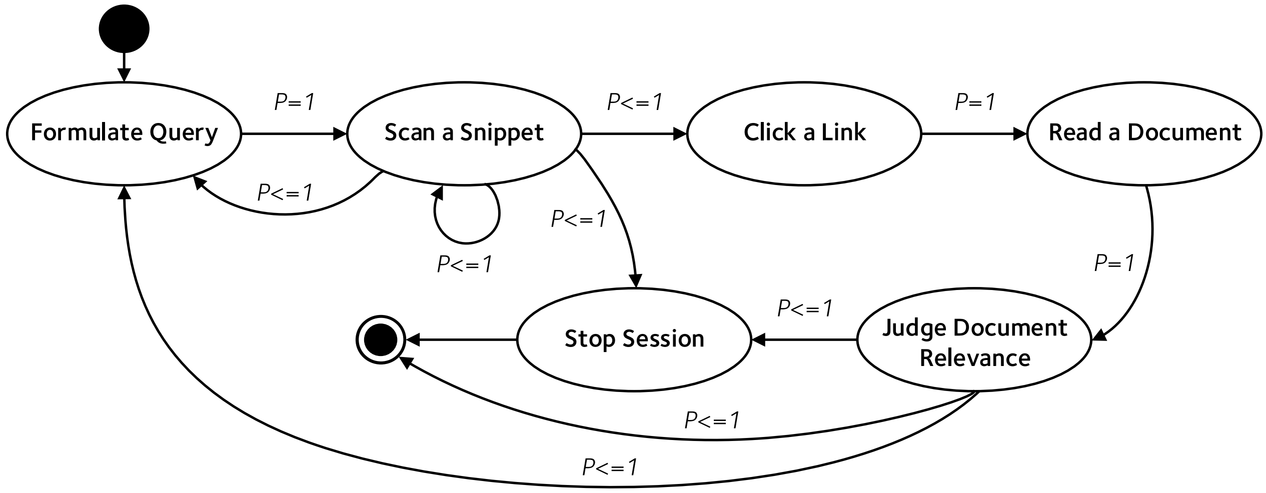


Figure 2.11 The searcher model, as outlined by Baskaya et al. (2013). Represented as a Markov model, the model considers six steps in all. Encoded within two of the steps are decision points that a user following this model must consider in order to continue. Figure adapted from Baskaya et al. (2013), with acknowledgement from the authors.

- **Query Formulation** considers where a searcher formulates the terms that they wish to enter, as well as the issuance of the query to the underlying retrieval system.
- **Snippet Scanning** concerns the action that considers the examination of individual result summaries for attractiveness.
- **Link Clicking** occurs when a result summary is considered to be attractive enough to warrant further examination.
- **Document Examination** considers the process of examining a document for relevance to the underlying information need (considering some form of document utility model (Carterette et al., 2011), or *document triage* (Marshall and Shipman, 1997)).
- **Relevance Judging** determines whether the document is relevant to the searcher's information need.
- **Session Stopping** concerns the action that curtails the search session.

These steps broadly match up the IIR process as outlined in Figure 2.8 on page 47, and described in an abstract form in Section 2.3.2. Azzopardi (2011), Yilmaz et al. (2010), Carterette

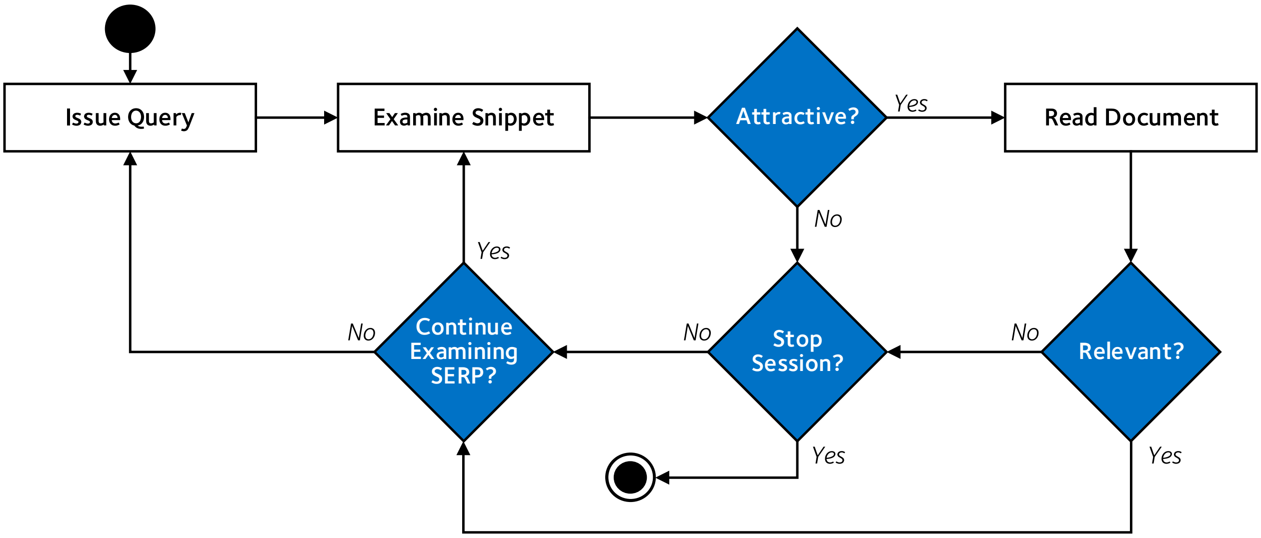


Figure 2.12 The searcher model, as outlined by Baskaya et al. (2013). Adapted from the version of the model illustrated earlier in Figure 2.11, this flowchart illustrates the main processes, decisions and interaction flow that an individual subscribing to this model will follow.

(2011), and more recently Zhang et al. (2017b) have all introduced and utilised searcher models that are similar in terms of the broad set of actions as outlined by Baskaya et al. (2013). Being a Markov model, this searcher model considers each of the different actions as a *state*, with a series of *transition probabilities* linking each of the actions together (i.e. *what is the probability of a searcher moving from scanning a result summary to stopping the search session?)*.[18] Indeed, the typical browsing model components (e.g. inspecting a document for relevance) as per Carterette et al. (2011) are also integrated with query generation and utility accumulation, permitting a searcher to issue multiple queries and gain utility.

Judging result summaries separately from associated documents is to be regarded as a major development in searcher models. The snippet text forming part of each result summary can be generated in a variety of different ways, as discussed earlier. This generation is important – what may look like a result summary that is attractive enough to explore further could potentially lead to a document that is in actuality not pertinent to the given information need. This is a clear development from the TREC-style searcher model which doesn't consider the notion of a separate result summary at all.

[18] This searcher model can also be represented as a flowchart, as illustrated in Figure 2.12.

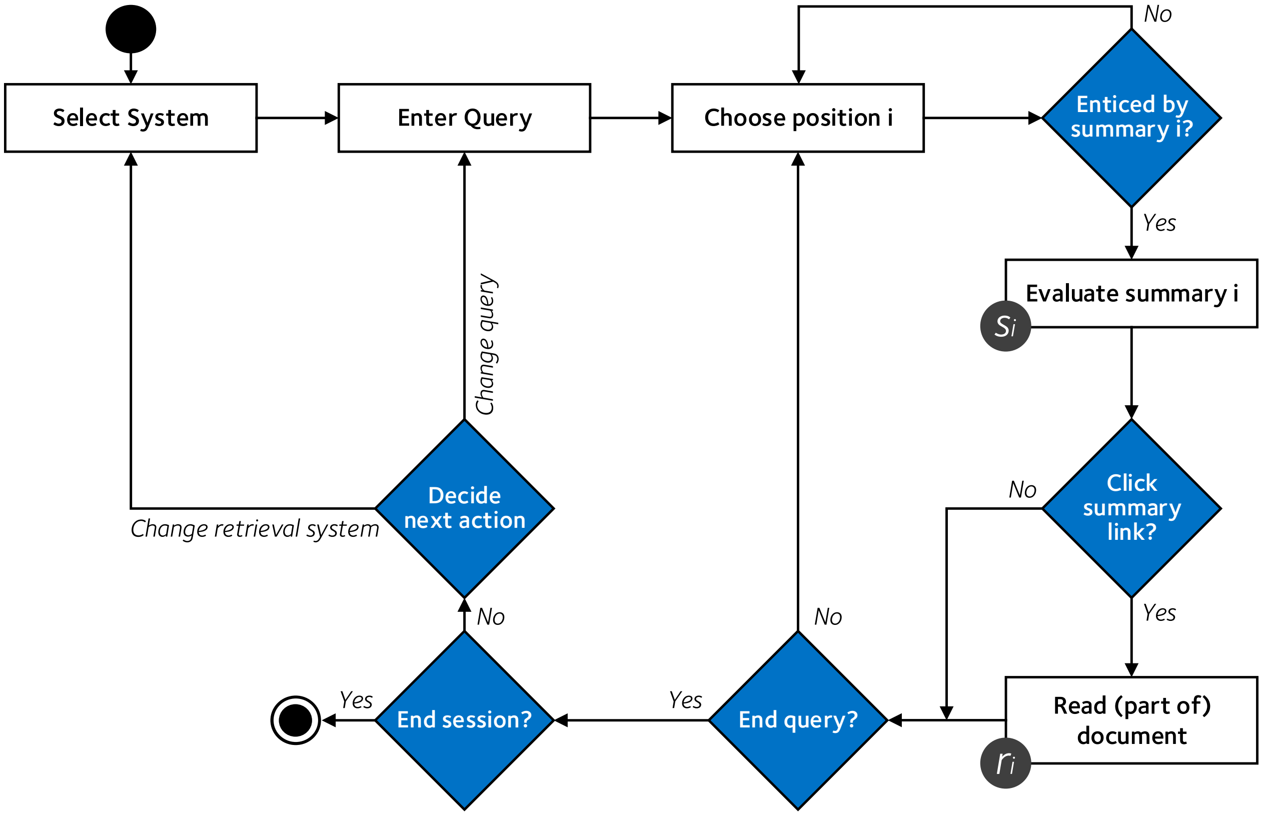


Figure 2.13 A further model of the search process, considering the high-level processes undertaken by a searcher, as outlined by Thomas et al. (2014). Figure adapted, with the support of the authors. Note here the inclusion of the overarching **Select System** process, where a searcher will consider what retrieval system to use before beginning the IIR process.

Effectively modelling stopping behaviour is also important to provide a more realistic model of the search process. The addition of decision points allowing a searcher to judge the attractiveness of a result summary, or the relevancy of a document, provides natural locations for **stopping decision points** (refer to Section 3.1.1).

> *Given that this result summary looks unattractive and therefore not useful for satisfying my information need, should I stop examining the results presented to me on this SERP?*

As an example, a searcher may issue a query that returns very few relevant documents. Once a few snippets and/or documents have been examined, he or she will then conclude that the issued query was unsuccessful and that they would be wasting their time examining more content on the presented SERP. In this case, issuing a reformulated query would be a better course of action. This intuition has been confirmed by empirical analysis, where Azzopardi (2011) for example demonstrated that simulated searchers examined significantly fewer documents when the underlying retrieval system failed to retrieve any relevant material in the top ten results, in contrast to when they did. This shows that searchers are inherently adaptive with their behaviours conditioned upon the quality of the ranked lists. This justifies the inclusion of these additional stopping decision points.

A further explicitly defined searcher model was proposed by Thomas et al. (2014). Illustrated as a flowchart in Figure 2.13, this model encapsulates the same principles as the previously described searcher models: given an information need, a searcher will issue one or more queries and examine a varying number of result summaries and documents for attractiveness and relevance in turn. This searcher model also includes a form of utility accumulation model that considers the utility obtained from a result summary s_i and the relevance judgement of a document, r_i. An addition to this searcher model is the inclusion of an additional step at the beginning of the search session, where a searcher must decide *what retrieval system to use* – or tool selection.

The searcher model proposed by Thomas et al. (2014) also considers the examination of result summaries from an alternative perspective. Instead of assuming that a searcher (perhaps naïvely) examines each individual result summary in detail before making a decision about its attractiveness, searchers subscribing to this searcher model undertake an *initial inspection* of result summaries, where they can examine in either direction on the list of ranked results (by increasing or decreasing some positional counter, *i*). Essentially, this can be regarded as a searcher *skimming* result summaries, determining if they are worthy of further examination, and adds an additional layer of complexity to the model.

Despite the slight variations in how different researchers interpret the IIR process, what can be clearly seen from an examination of these searchers models is the similar workflow between each of them. From query issuance to document examination and judgement, this has emerged as the generally accepted process of IIR. While studies examining explicit searcher models are not common, searcher models have been encoded within many of the different evaluation measures used within IR and IIR. Moving onto the final section of this background chapter, this is now the topic of our attention.

2.4 Evaluation Measures

We now turn our attention towards the *evaluation* of both retrieval systems and the searchers that use them. Careful and thorough evaluation of retrieval systems is required to demonstrate the superior performance of new retrieval models or a new way of presenting of results, for example (Manning et al., 2008). Recalling that the *modus operandi* of a retrieval system is to satisfy the information needs of the searchers who utilise it, Lancaster (1968) provided three criteria by which an IR system can be evaluated:

- the **suitability** of a retrieval system in terms of the searcher's specific tasks;
- the retrieval system's **task performance efficiency**; and
- the extent to which the retrieval system **satisfies** information needs.

These three criteria themselves may be split into two separate categories, considering:

- how well the *system* performs, utilising the first two criterion above; and
- how the system performs in the eyes of the searcher who is using the retrieval system at the time (Voorhees and Harman, 2005).

Measures for both system and user-focused evaluation are considered in this section, in Subsections 2.4.1 and 2.4.2 respectively.

Other Evaluation Measures The measures discussed in this section are only a subset of the range that have been developed and trialled. This section focuses on measures used throughout this thesis. For a more comprehensive summary, refer to Sanderson (2010).

2.4.1 System-Based Evaluation

Considering system-orientated measures of evaluation, one can consider a system's *efficiency* or *effectiveness*. Efficiency concerns some form of operational metrics, such as the speed of the retrieval system. This example is important (especially in commercial retrieval systems), as even a fractional increase of the time taken to return results to a searcher can reduce the number of returning searchers – thus impacting upon the amount of revenue generated through advertising (Brutlag, 2009).

However, when one thinks of the evaluation of a retrieval system, its *effectiveness* is typically considered. Of course, the definition of what defines a retrieval system to be effective depends upon the type of search task being undertaken. A patent searcher would, for example, expect a retrieval system to return *all* relevant patents to avoid missing a related patent (and thus incurring penalties). However, a casual web searcher curious about a topic they know little about (i.e. searching in an *ad-hoc* fashion) may be satisfied with either a singular or a small number of results.

This section provides a brief overview of basic effectiveness measures widely used within IR today from the perspective of system-sided evaluation – or the retrieval system's output – the generated ranked list of results.

2.4.1.1 Precision

The *precision* (P) of a ranked list of results is the fraction of documents that have been retrieved that are considered to be relevant or useful for the searcher's information need. A retrieval system that yields high precision for individual queries is regarded as one that performs well and satisfies searchers.

Figure 2.14 provides a visual illustration of what precision entails (as well as its counterpart, *recall,* which is discussed below). Given the set of all documents within an index, a retrieval system will retrieve a number of these documents that satisfy the criteria set out in the employed retrieval model. Of the documents retrieved, some will be considered relevant to the searcher's information need; others will be considered non-relevant. As such, prior knowledge as to what documents are relevant to a given topic are therefore required to successfully compute the precision of a ranked list – TREC QRELs are an example of prior knowledge, as we discussed earlier in Section 2.3.1.1.

Precision is defined as:

$$P = \frac{|\, relevant\ documents\ \cap\ retrieved\ documents\,|}{|\, retrieved\ documents\,|}. \quad \text{Equation 2.3}$$

IR research reports precision up to a particular rank, i.e. P@k. For example, *P*@10 will provide a fractional value for the number of relevant documents that appeared within the top 10 results for a given query. Herein lies one of the most elementary and basic *stopping models* that we find implicitly encoded within various IR measures. Stopping at a depth of 10, or k (refer to[19]) denotes that documents past this rank are not examined.

2.4.1.2 Recall

While precision considers the fraction of documents retrieved that are relevant, *recall* (R) considers the fraction of documents that were retrieved and relevant to a query against *all*

[19] The value of $k = 10$ is often chosen in IR research as it has been shown that this is typically the depth to which searchers would look through web search results, perhaps due to the effects of pagination (Jansen and Spink, 2006).

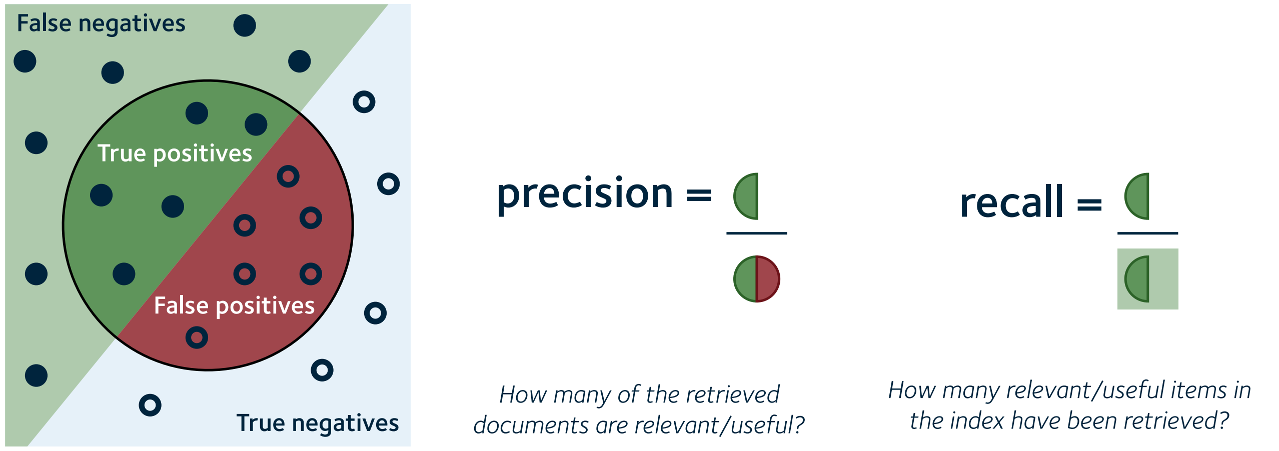

False negatives
True positives
False positives
True negatives
precision =
recall =
How many of the retrieved documents are relevant/useful?
How many relevant/useful items in the index have been retrieved?

Figure 2.14 An illustration of precision and recall. On the left is an illustrated example of an index, containing many documents. The large circle represents the set of documents retrieved for a query. Documents that are relevant to the query are represented as ●, with non-relevant documents represented as ○. Note that not all relevant documents are retrieved; doing so would mean that the retrieval engine used would have produced perfect results. On the right of this illustration, definitions of *precision* and *recall* are also provided.

known relevant documents for a query. Recall can formally be defined as:

$$R = \frac{|\, relevant\ documents\ retrieved\,|}{|\, relevant\ documents\,|}. \tag{Equation 2.4}$$

Considering the patent searching example defined above, a high recall would be more desirable in a patent searching task. This means that more patents matching the searcher's query will be returned, thus reducing the possibility of missing important prior, relevant patent filings (and reducing the risk of any penalties).

Given more modern retrieval models, the notion of ranking would lead a searcher to assume (as per the PRP (Robertson, 1977)) that relevant documents pertaining to their query will be the first results presented. Non-relevant documents will also of course appear, typically leading to some form of a tradeoff between precision and recall, as discussed below. The tradeoff considers the notion that as you increase recall, the number of non-relevant items will undoubtedly also increase, thus reducing the overall precision of the retrieval system.

2.4.1.3 Expected Search Length

The *Expected Search Length (ESL)* (Cooper, 1968) of a searcher considers the number of non-relevant documents that a searcher will have to search through to obtain the *desired number of relevant documents.* Therefore, systems demonstrating a shorter ESL are considered to be more effective than systems with a comparatively longer ESL. In other words, the ESL indicates how much wasted search effort one would expect using a given retrieval system, as opposed to randomly searching until the needed relevant documents are found.

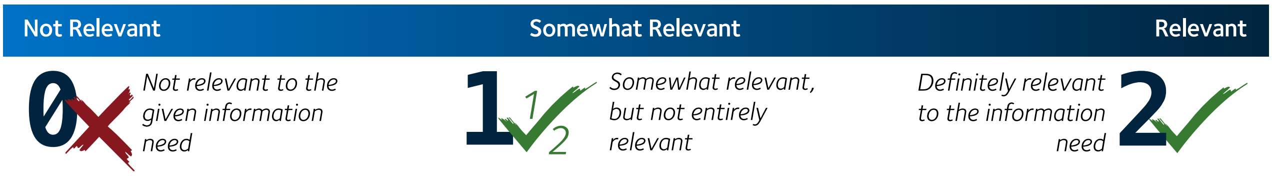
While we do not explicitly use this measure in the contributory work of this thesis, the ESL does nevertheless provide motivation for a number of different stopping heuristics outlined in Section 3.2.

2.4.1.4 Cumulative Gain Measures

An important suite of measures that address the *gain* (or *utility*) that is presented in a ranked list can be derived from *Cumulative Gain (CG)*. Outlined by Järvelin and Kekäläinen (2000, 2002), CG is measured as the cumulation of gain of all relevant documents up to some rank k (or *CG@k*). CG can be measured on a system-sided basis, purely considering the ranking provided. Alternatively, it can also be measured from a user-sided stance, considering only the documents that a searcher has identified (or saved) as relevant to their information need.

Determining a value for the level of gain that can be acquired from a document is not trivial; experimentation in IIR (such as the simulation of interaction study by Pääkkönen et al. (2015)) typically uses the TREC relevance judgement score for a given topic and document combination as the level of gain that is accrued from a document.

This is demonstrated by the illustration below, highlighting the concept of *graded relevance judgements*. When not related to the information need, a score of 0 is assigned, with a relevant document assigned a judgement of 2. Documents partially fulfilling the relevance


Not Relevant
Somewhat Relevant
Relevant
0
Not relevant to the given information need
1
1/2
Somewhat relevant, but not entirely relevant
Definitely relevant to the information need
2

requirements are assigned an intermediary score of 1. These values are then accumulated over the rankings to yield the CG measure for a searcher's interactions.

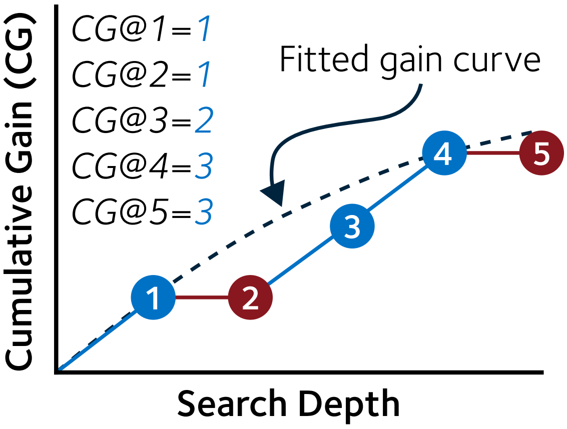


Figure 2.15 An illustration of the *Cumulative Gain (CG)* as outlined by Järvelin and Kekäläinen (2000, 2002). In the example, the first five documents from a list are shown — *1*, *3* and *4* contain information, from which the user gains.

Relevance judgements can also be indicative of a searcher's stopping behaviour. Assuming a perfect ranking, a searcher would be wise to stop his or her interactions with a ranked list once documents that yield no gain begin to appear as they traverse the ranked list. Continuing examination of the ranked list from that point on would be a waste of their time. Given the illustration in Figure 2.15 however, it is unlikely a perfect ranking will occur in reality, and thus searchers need to determine whether they should continue examining a ranked list after encountering several documents yielding no gain in a row.

Given the definition above, CG can be computed as:

$$CG_k = \sum_{i=1}^{k} rel_i, \qquad \text{Equation 2.5}$$

where k denotes the ranking at which CG should be calculated, and rel_i denotes the relevance assessment score for a document at rank i. Given this definition, a computed CG value will be unaffected by changes in rankings. Two lists of rankings with one highly relevant document in rank 1 in one list, and rank 5 in the other, will yield identical CG values.

A further development of CG is *Discounted Cumulative Gain (DCG)*, which, as the name may imply, *discounts* the level of gain accrued by searchers the further down the ranked list of results that they go, thus addressing the problem of CG by increasing the accuracy of the reported values. Specifically, the gain obtained by a new document is discounted according to the rank of that document (i.e. *weighted precision*). A new relevance score is therefore computed for each document by dividing the relevance assessment score by the *log* of its rank. Further developments have included measures such as *Normalised Discounted Cumulative Gain (nDCG),* which addresses the issue of different result lists having different lengths. To address this, nDCG considers the *ideal DCG (iDCG)* of all relevant documents in a corpus for a given query, where the relevance judgements of these documents are ordered by decreasing relevance (i.e. $2, 2, 2, 1, 0$). Dividing the DCG score for a ranked list up to rank k by the iDCG score up to rank k thus normalises the reported value across queries.

2.4.1.5 Rank-Biased Precision

Rank Biased Precision (RBP) (Moffat and Zobel, 2008) is a more contemporary IIR measure that is derived from a simple searcher model. It considers the notion that searchers are not willing to examine every result presented to them in a ranked list. The encoded searcher model, illustrated in Figure 2.16, assumes that a searcher will always examine the first result presented to him or her. The process can be likened to the previously discussed idea of using a Markov model to represent the search process (Tran et al., 2017), modelling the likelihood of a searcher reaching a given depth. Subsequent documents further down the ranking will then be examined with a decreasing likelihood. Given a ranked list of documents, d, RBP is defined as:

$$RBP = (1 - p) \cdot \sum_{k=0}^{d} rel_k \cdot p^{k-1}, \qquad \text{Equation 2.6}$$

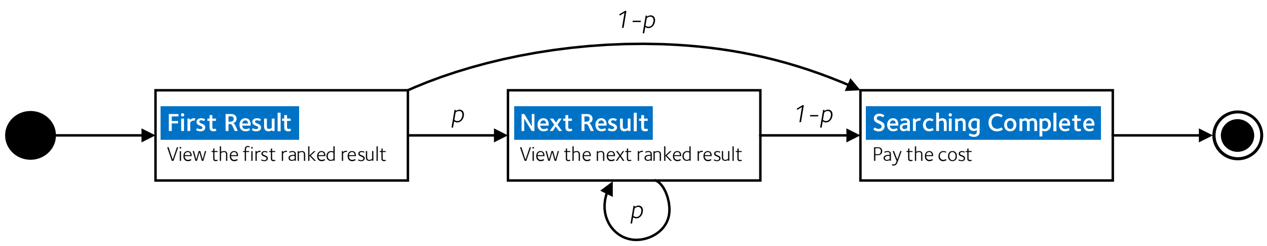


Figure 2.16 An illustration of the simple searcher model encoded within *Rank Biased Precision (RBP)*. A user following this model will *always* examine the first result presented, and will examine subsequent results with decreasing probability (from *p*, the *patience* parameter). Stopping is determined with probability *1-p*. Figure used with support from Moffat and Zobel (2008).

with rel_k once again denoting an assessor's relevance judgement for the document at rank k, and p denoting the *patience parameter.* It is this parameter that provides a decaying probability that a searcher will continue to examine a ranked list of results the further down the rankings they go. A searcher subscribing to the RBP searcher model will stop examining results with probability $1 - p$. The patience parameter allows for a very flexible measure; one can model both very persistent searchers (when p tends towards 1.0) and impatient searchers (where $p \approx 0.5$) (Moffat and Zobel, 2008). When $p = 0.0$, a searcher will examine only the first presented result for relevance, and then stop.

RBP has been shown to fit well with actual searcher data extracted from click logs, as demonstrated by Chapelle et al. (2009) and Zhang et al. (2010). RBP is used within this thesis as a means for attempting to decide when a simulated searcher should stop examining a list of ranked results. By incorporating an additional probability in conjunction with the calculated RBP score at some rank k, we can determine whether the searcher should stop.

2.4.2 User-Based Evaluation

With basic evaluation measures above considering the system-sided aspects (i.e. the quality of the retrieved ranked list), this section considers a number of IIR evaluation measures related to the searcher and their interactions. It may be interesting to note that a large number of measures for IIR have been proposed over the years (and indeed, over the experimental spectrum as shown in Figure 2.10); however, only a small number of measures have been regularly used in the literature. A more in-depth summary of IIR measures can be found in works by Su (1992) and Kelly (2009).

Indeed, Kelly (2009) provides a taxonomy of the different types of measures used within IIR experimentation. The taxonomy consists of four main categories.

- **Contextual Factors** These measures are related to the *subject* (or participant), of an IIR experiment. Factors include those commonly gathered from forms of demographics surveys, such as the subject's age, sex, and prior search experience. In an information seeking context, one will also be able to gather information about the prior knowledge of the searcher and their knowledge. For example, questions can be asked about their knowledge of the topic they are to find information about.
- **Interactions** These measures characterise the interactions that take place between the system and the searcher – including their behavioural characteristics. Examples of such interactions include: the number of queries that the searcher issues; the number of documents that they examine; the depth to which they examine results (stopping depth); and the mean length (in terms) of the queries that they issue. Time-based measures are also included in this category – both at a gross level (i.e. the total session time), and at a more specific level (i.e. the mean time spent entering queries). These measures can be usually computed by extracting and parsing interaction *log data*.[20]
- **Performance Factors** As the name suggests, these measures are related to the outcome of the interactions that take place between the searcher and the retrieval system. These measures can be considered analogous to the system-sided measures that we examined previously in Section 2.4.1, but with an emphasis on a searcher's interac-

[20] By *log data,* we refer to the file that is created by an experimental retrieval system as a subject conducts a search task. Typically, a series of different actions (i.e. issuing a query, or clicking a document link) are logged, and post-hoc log analysis can interpret the logged events, computing the requested measure.

tions (i.e. considering only documents explicitly identified by the searcher). As such, these measures can again be extracted from interaction log data.

- **Usability Measures** These measures are typically a series of qualitative and quantitative approaches for capturing a subject's feelings and attitudes towards a search system that they have used. Common measures in this category include the subject's view of the system's effectiveness and their overall *satisfaction* with how they performed when undertaking the search task in question (Hornbæk, 2006).

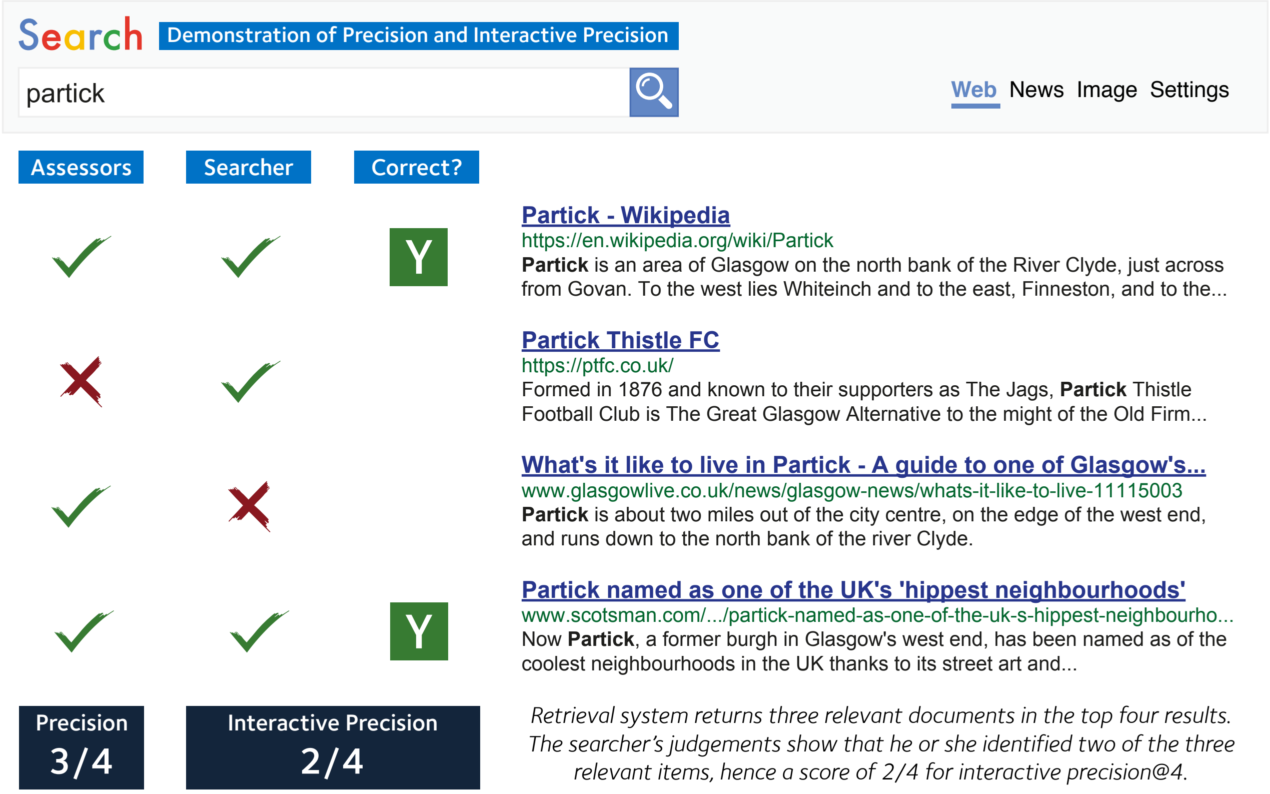
In this thesis, we employ measures from all four of the above – including the extraction of interaction data from user studies, which is subsequently employed to *ground* simulations of interaction. For example, measures include a variant of CG that considers the documents a searcher identifies as relevant during a search session. These are used to compute the level of gain that he or she experienced, rather than the actual CG of the ranked list.[21] This section also discusses an additional measure, considering again the interactions that take place between the searcher and retrieval system.

2.4.2.1 Interactive Precision and Recall

During the IIR process, searchers examine a number of different result summaries (and their associated documents), making individual judgements as to the relevance of each. Some may not be relevant and are disregarded by the searcher. This differs from the abstracted TREC-style searcher model that assumes that all documents returned are assessed in their entirety by the searcher.

Subjects of IIR studies are typically instructed to *save* documents that they consider relevant to a provided information need. As such, the judgements created by subjects of IIR studies may not match with those made by the assessors of relevance judgements provided as part

[21] Of course, the CG of a ranked list will equal a searcher's computed CG if he or she identifies all relevant documents correctly.



partick

Assessors	Searcher	Correct?
✓	✓	Y
✗	✓	
✓	✗	
✓	✓	Y

Partick - Wikipedia
https://en.wikipedia.org/wiki/Partick
Partick is an area of Glasgow on the north bank of the River Clyde, just across from Govan. To the west lies Whiteinch and to the east, Finneston, and to the...

Partick Thistle FC
https://ptfc.co.uk/
Formed in 1876 and known to their supporters as The Jags, **Partick** Thistle Football Club is The Great Glasgow Alternative to the might of the Old Firm...

What's it like to live in Partick - A guide to one of Glasgow's...
www.glasgowlive.co.uk/news/glasgow-news/whats-it-like-to-live-11115003
Partick is about two miles out of the city centre, on the edge of the west end, and runs down to the north bank of the river Clyde.

Partick named as one of the UK's 'hippest neighbourhoods'
www.scotsman.com/.../partick-named-as-one-of-the-uk-s-hippest-neighbourho...
Now **Partick**, a former burgh in Glasgow's west end, has been named as of the coolest neighbourhoods in the UK thanks to its street art and...

Precision
3/4

Interactive Precision
2/4

Retrieval system returns three relevant documents in the top four results. The searcher's judgements show that he or she identified two of the three relevant items, hence a score of 2/4 for interactive precision@4.

Figure 2.17 A graphical example of *interactive precision* alongside traditional, system-sided precision. In the example ranking, the retrieval system returned three relevant documents, yet the searcher only 'correctly' identified two of them, hence an interactive precision *(iP)* score of *2/4=0.5*. This is in contrast to the system's precision score of *3/4=0.75*.

of a test collection. Indeed, some TREC topics that are widely used contain hundreds of documents marked as relevant by assessors. It is unlikely that a subject would be able to find *all* of these documents or agree with an assessor's judgement.

Interactive Precision (iP) and *interactive recall* were defined by Veerasamy and Belkin (1996) and Veerasamy and Heikes (1997). Instead of purely considering precision and recall as measures exclusively utilising the relevance judgements provided as part of a test collection, one would also consider the number of documents considered relevant by a subject of an IIR study that were also TREC relevant. This means that a document judged relevant by an assessor may not be retrieved, viewed and subsequently judged by the subject of an IIR study. We demonstrate this with a visual example in Figure 2.17. The **Assessors** column denotes the judgement from the relevance assessors, while the **Searcher** column denotes the judgement of the searcher. In the example illustration, the retrieval system's *P@4* score is $3/4 = 0.75$, while the searcher's *iP@4* $= 2/4 = 0.5$. Despite the searcher saving three documents as relevant, the document at rank two was not assessed as such, resulting in the iP score being 0.5. Therefore, only two of the three saved documents contribute.

2.4.3 Evaluation Measures and Stopping

Throughout this section, we have outlined a number of different evaluation measures commonly employed in both IR and IIR studies. Common with these measures are implicit models encoded within them that in turn provide a *stopping point* in the ranked list.

These models vary from the simplistic to more complex, with the more complex approaches providing a more realistic rationale of the stopping behaviour of real-world searchers. Regarding stopping behaviours, the most simplistic approach discussed in this chapter is *P@k* – or *stop at rank k.* This is agnostic of the relevance of the results presented, and is often described as a fixed-depth assumption, something that we discuss later in Section 5.1. Other measures such as RBP (Moffat and Zobel, 2008) and nDCG (Järvelin and Kekäläinen, 2002) offer of a more complex stopping model. These measures consider a decreasing likelihood of continuation the deeper a list of results is traversed.

Evaluation measures are however only a small part of the work undertaken in order to understand the different *stopping behaviours* exhibited by searchers during the IIR process. Chapter 3 provides an in-depth overview of prior work examining this area.

2.5 Chapter Summary

In this chapter, we have introduced some of the key constructs and components of an IR system – from the document indexing process, to the retrieval models that are used to return a (typically) ranked list of results to the searcher. We also briefly touched on the history of the field, discussing some of the key manual and mechanised systems that were commonplace before the advent of computers and the WWW, giving rise to contemporary IR systems.

The focus of this chapter has however mainly been on the disconnect between traditional IR research, and the reality of what searchers actually do during the IIR process. In particular, we identified a number of limitations within the TREC-style searcher model that is commonly employed, and discussed a number of more advanced searcher models that reduce the assumptions (and therefore limitations) that the models provide, particularly in terms of a searcher's *interactions.* This led to a discussion of how IR systems and the searchers that use them can be evaluated, along with a discussion of the different categories of study in IR and IIR – from *system-focused* to *user-focused.*

The next chapter continues the focus towards the searcher, considering previous work that has been undertaken to examine *stopping in IIR.*

Chapter 3

Stopping in Interactive Information Retrieval

Towards the end of Chapter 2, we examined a number of different evaluation measures typically employed in IR and IIR studies. In particular, we emphasised the notion of different *stopping models* that are implicitly encoded within these measures, ranging from the naïve to more representative approaches of a real-world searcher's *stopping behaviours*.

Actually, please continue reading...

In this chapter, we provide an overview of work undertaken in the field of IIR that explicitly examine the stopping behaviours of searchers. We enumerate on a number of different *stopping heuristics* that attempt to quantify when searchers should stop examining results, before examining *theoretical frameworks* that provide insight and explanation into why and when searchers stop. We then examine a number of different *user studies* that have examined stopping behaviours. Before examining these prior works, we first consider why examining the stopping behaviour of searchers is important to the field, and to the future development of the retrieval systems and their interfaces that we use extensively today.

3.1 Why Stopping?

Knowing when to stop is a fundamental aspect of human thinking and behaviour. Humans and other animals, when interacting with the world, will employ some form of *stopping criterion* (or criteria) to decide when they should stop (Nickles, 1995). As an example, a shopper who is looking to purchase a new smartphone will stop shopping around once he or she has obtained sufficient information on which new smartphone to purchase. Once their case notes for a patient have been compiled, a medical doctor will then be in a position to diagnose the patient's ailment. In the context of search, numerous reasons exist why searchers stop. Perhaps searchers stop because they have satisfied their information need, have become frustrated with the lack of potentially relevant information – or because of some external factor, such as a time constraint that has been imposed upon them.

The decision of when to stop is not exclusively due to such external factors to the decision maker, but rather from a series of *internal, cognitive factors* of their thinking process (Nickles, 1995). For example, an individual who is hungry will stop eating once he or she feels full, rather than stopping when all of the food presented to them has been consumed. Empirical research has over the years demonstrated that individuals, regardless of the task presented to them, will frequently stop prematurely. Indeed, this naïve behaviour demonstrates that individuals may be willing to go with what *"sounds right"* to them – often minimising the cognitive effort that is required at the expense of greater accuracy (Perkins et al., 1983). However, when searching, this lower level of potential accuracy does lead to individuals making a greater number of errors in their decision making (Baron et al., 1988). Searchers overlook important elements, and potentially miss out useful information (Fischhoff, 1977; Fischhoff et al., 1978; Shafir and Tversky, 1992), with the individual then failing to consider alternatives (Farquhar and Pratkanis, 1993).

Based upon prior research into stopping behaviours, it is clear that such a decision is driven primarily from internal factors. As such, we then consider: *what aspects of the decision maker's*

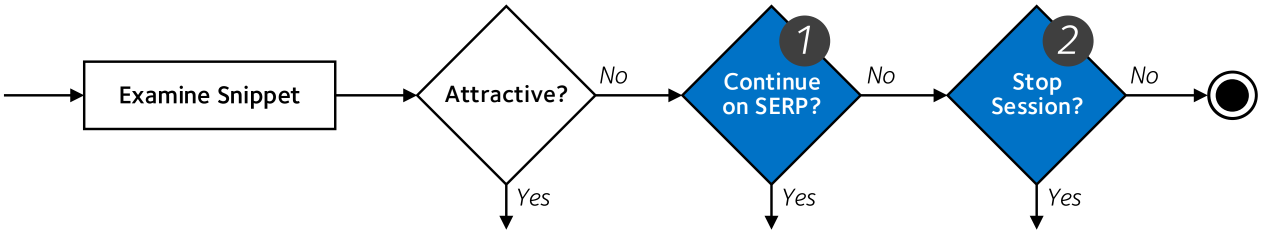


Figure 3.1 Excerpts from various searcher models, highlighting two established *stopping decision points*. These are modelled as points at which searchers can stop performing a given action. These are illustrated as blue diamonds. The two points consider **1** *result summary level stopping* (often called *snippet level* or *query level stopping*), and **2** *session level stopping*.

thinking processes prompt him or her to stop assessing the information provided? Knowing when to stop requires that the individual in question makes a *judgement* regarding the sufficiency of the information obtained, and whether or not additional information is required (Browne and Pitts, 2004). This judgement is normally characterised by both the completeness and correctness of the information obtained thus far (Smith et al., 1991). These claims can be mirrored by qualitative studies on examining stopping behaviour. Here, researchers have found that searchers stop examining search results simply because what they have found previously is *"good enough"* (Zach, 2005) to satisfy their underlying information need. This finding echoes the reasoning that individuals would be happy to stop when what they have found *"sounds right"* (Perkins et al., 1983).

3.1.1 Stopping Decision Points

In Section 2.3.5, we discussed a number of searcher models that are considered to be an improvement over the traditional TREC-style searcher model, a model that is agnostic of a searcher's interactions. The more advanced models considered two distinct *stopping decision points* that capture specific points during the interaction process where a searcher can stop their current activity, and move onto the next step in the process. These stopping decision points are illustrated in an excerpt of a typical searcher model flowchart, as shown in Figure 3.1. Both established stopping decision points are discussed below.

1 **Result Summary Level Stopping** Traditionally called *snippet level stopping* or *query level stopping* in the literature[1], this stopping decision point considers the depth at which a searcher will stop examining a list of ranked results. After stopping at this point, the searcher can continue the search session by issuing a further query.

2 **Session Level Stopping** This second stopping decision point considers the point at which a searcher will stop their search session in its entirety. As such, this stopping decision point is regarded as terminal to the search session.

In particular, session level stopping is considered when a searcher must decide, for example, if they have met their overall search goal, have run out of time or queries, or simply have become so frustrated with a lack of relevant content that they would rather abandon their search. These stopping decision points can be operationalised in a variety of different ways, as we explore in the remainder of this chapter. Chapter 4 also proposes an additional, third stopping decision point that we will consider in a later contributory chapter of this thesis.

3.2 Stopping Heuristics

Considering the above, researchers have over several decades devised a number of different high level *stopping rules* – hereafter referred to as **stopping heuristics** – as a means of encoding a searcher's aforementioned sense of what is *"good enough"* (Zach, 2005) – or even what can be considered as *not good enough*.

Stopping heuristics have been investigated in *decision-making* research. A number of normative stopping heuristics have been identified. As examples, Busemeyer and Rapoport (1988) considered the expected loss from terminating information acquisition. Kogut (1990) examined the expected value of additional information. Other examples of normative stopping heuristics are demonstrated by Pitz et al. (1969) and Busemeyer and Rapoport (1988).

[1] The phrase *result summary level stopping* is used in this thesis to avoid confusion with a new SERP level stopping decision point, discussed in depth in Section 4.3.1 on page 113.

However, as outlined by Browne and Pitts (2004), these heuristics usually fail to describe the actual cognitive behaviours of the decision makers. Such heuristics often assume that the decision maker must *think ahead* to the final decision of when to stop, enabling them to assess the value of additional information (Busemeyer and Rapoport, 1988). This is an inherently difficult task for decision makers to undertake, due to the limited working memory capacity of a human. We are simply unable to cognitively process all of the information attained to make a decision considering all possible outcomes (Browne and Pitts, 2004). Nickles (1995) agreed, stating that normative stopping heuristics made implicit assumptions about the mental activities of the decision maker, especially in terms of mental scaling and weighting. No clear cognitive perspective has yet been provided to address the cognitive mechanisms and/or assumptions of the decision maker's thinking.

Nickles (1995) identified two distinct approaches to considering the cognitive processes involved in decision making: judgement, where an individual assesses a context to choose a course of action; and reasoning, where an individual convinces himself or herself that a particular understanding of the scenario is correct. Research in this area of decision making typically assumed that when assessing a particular scenario (or presented information), a decision maker would draw upon the available evidence and use their judgement to make a decision as to how to proceed (Reisberg, 1997). This assumption was implicitly used in the normative stopping heuristics outlined above. An alternative way to consider what a decision maker undertakes revolves around the notion that taking a decision is dominated by his or her ability to reason. Drawing on the available evidence, *arguments* can be constructed to reach an overall conclusion. This is known as *belief assessment* (Benson et al., 1995), where the individual determines their degree of belief in the conclusion that has been reached.

With the inherent limitations of the normative stopping rules in mind – and the two categorisations defined above, we now enumerate a number of different stopping heuristics that are better able to represent a searcher's cognitive processes, considered as either *judgement-based* or *reasoning-based.* We enumerate these heuristics below in their two classifications.

- **Judgement-based Heuristics** These heuristics are defined as when a decision maker maintains some mental threshold along a key dimension and a running total of the number of occurrences of this measure. More details are provided in Section 3.2.1.
  - **Satisfaction and Frustration** These heuristics consider a decision maker's satisfaction with what they have found during the course of their search *(satisfaction* or *satiation),* or their tolerance to non-relevance *(frustration* or *disgust).*
  - **Difference Criterion** This heuristic concerns the notion of whether a decision maker is learning anything new by examining more documents.
  - **Magnitude Threshold** This heuristic concerns a decision maker's belief that the information that they have found provides sufficient evidence to prompt him or her to stop searching for further information.
  - **Single Criterion** A *single criterion* to the decision maker's information need is considered in this stopping heuristic.
- **Reasoning-based Heuristics** This classification concerns the *mental representation* of the given topic for which a searcher is seeking information. In other words, the mental representation is formed from a series of (perhaps contrasting) points. When combined together, a decision can be made as to the suitability of the information found.
  - **Representational Stability** This stopping heuristic concerns the notion of the decision maker's mental model of the topic and the *stabilisation point.*
  - **Propositional Stability** Here, a series of potential conclusions regarding the underlying information need are formed, with these arguments needing to be satisfied to feel sufficiently satisfied to stop.
  - **The Mental List** A mental list of aspects is constructed, with each item on the list needing to be addressed by the decision maker before stopping occurs.

These were devised largely as ways of modelling the *Expected Search Length (ESL)* (Cooper, 1968), as briefly discussed in Section 2.4.1.3. Nickles (1995) also proposed a number of stopping heuristics that are discussed in subsequent sections, with discussion expanded to include additional heuristics defined by other researchers. We now address each of the two classifications, explaining each of the different heuristics enumerated above in detail.

3.2.1 Judgement-Based Heuristics

As discussed previously, a judgement-based stopping heuristic is defined as when a decision maker is assumed to set and consistently maintain a mental threshold along some form of key dimension (e.g. determining the seemingly relevant from non-relevant), and to keep a running total of the measure relative to the dimension in question (Gettys and Fisher, 1979; Nickles, 1995). When the measure meets or exceeds this set threshold, the searcher will then stop searching for information. Each of the judgement-based heuristics we consider in this thesis are discussed in turn below.

3.2.1.1 Satisfaction and Frustration

Two of the earliest stopping heuristics defined in the literature are by Cooper (1973b), who consider a searcher's tolerance encountering non-relevant material, and how satisfied they become when encountering relevant material. The heuristics were originally defined as a means for estimating the utility a searcher can attain when interacting with a retrieval system. While the means of which Cooper (1973b) estimated the utility of search are not of key relevance to this thesis, the work on stopping heuristics is. The *satisfaction point* and *frustration point* stopping heuristics are considered to be judgement-based heuristics, as they rely solely on the searcher's notion of what constitutes a relevant document. Both consider counts of the number of (non-)relevant documents observed.

Satisfaction Point The *satisfaction point* heuristic considers the point at which a searcher has found enough material to consider his or her search a success. This is achieved by considering the amount of material found that has been judged to be relevant. It can easily be imagined that such a heuristic would apply directly to both result summary level stopping (i.e. *find x relevant documents on this SERP*) and session level stopping (i.e. *find x relevant documents*). This heuristic is also called the *satiation heuristic* (Simon, 1955) (see below). This heuristic can be considered as a decision making process...

> *"[...]through which an individual decides when an alternative approach or solution is sufficient to meet the individuals' desired goals rather than a perfect approach."*
>
> **Simon (1971)**

This suggests that searchers employing the satisfaction heuristic would stop searching as soon as certain conditions arise, instead of after they have exhaustively considered all available information (March, 1994). Conditions could include acceptance of the results; discomfort; boredom; time limits; and the *snowballing* of information (Mansourian and Ford, 2007), where the repetition or saturation of information occurs.

Frustration Point In a converse fashion to the satisfaction point heuristic, the *frustration point* heuristic considers a searcher's overall *tolerance to non-relevance* by stopping after being sufficiently frustrated by the results presented to the searcher. This heuristic is also called the *disgust heuristic* in the literature (see below).

The two relatively straightforward heuristics defined above makes a searcher's interactions with a ranked list of results *inherently adaptive.* In other words, given a set of results, his or her behaviour will change with respect to the perceived quality of the ranked list. As a reminder, this would not necessarily mean considering the system's effectiveness measures, but rather user-focused measures such as interactive precision and recall, as discussed previously in Section 2.4.2.1.

Combining Satisfaction and Frustration Perhaps due to the relative simplicity of the two aforementioned heuristics, identical approaches have been defined elsewhere in the literature. Kraft and Lee (1979) later defined three further stopping heuristics, two of which are the *satiation* (as per Simon (1955)) and *disgust* heuristics. In essence, the rules defined by Kraft and Lee (1979) are the same satisfaction and frustration heuristics as previously defined by Cooper (1973b). Within the satiation rule, a searcher will stop after becoming *satiated* by finding a number of documents considered to be relevant, while the disgust rule considers a searcher's disgust at finding a number of non-relevant documents.

Kraft and Lee (1979) also proposed a third heuristic that combines both satisfaction/satiation and frustration/disgust together into a single heuristic. Here, a searcher following such an approach would be inclined to stop examining content if they were either satisfied with what had been found, or frustrated by having to trawl through material judged to be non-relevant (thus considering multiple criteria). The stopping point would be whatever of the two conditions are met first. Indeed, Kraft and Lee (1979) demonstrated that the ESL of a searcher could be approximated using each of the two stopping heuristics by considering the size of the retrieval set, the number of relevant documents a searcher wished to obtain, and the number of non-relevant documents a searcher would be willing to tolerate. The number of documents required to consider a search as successful is dependent upon whether the search task is high precision (where one would stop comparatively early), or high recall (where one would stop comparatively later), as hypothesised by Bates (1984).

3.2.1.2 Difference Threshold

The *difference threshold heuristic* (Nickles, 1995) concerns whether a new document is providing a searcher with additional, useful content about their information need. Here, the searcher is assumed to keep an internal record of the information that has been consumed along some key dimension. The searcher is also assumed to use this internal record of what has been assessed to compare a new document with previously examined content. When the difference between the new and existing information falls below some internal difference threshold, the searcher stops as nothing new is being learnt.

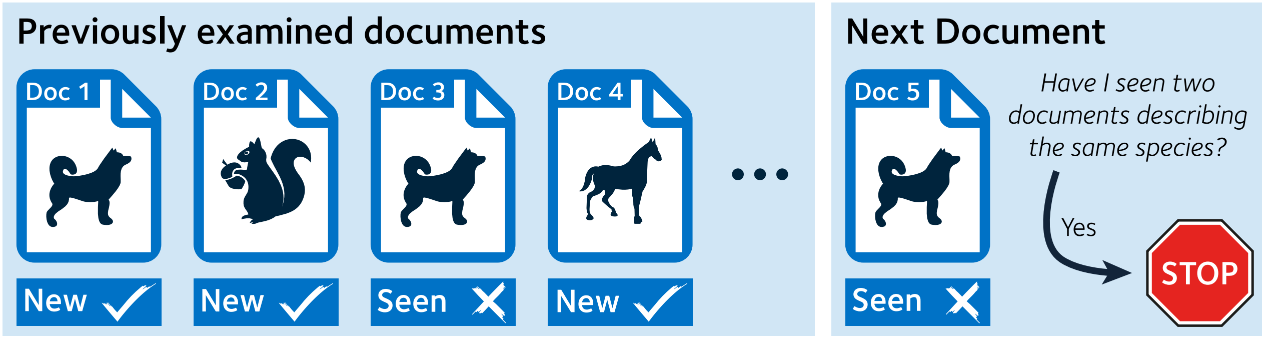

Previously examined documents
Doc 1
New ✓
Doc 2
New ✓
Doc 3
Seen ✗
Doc 4
New ✓
...
Next Document
Doc 5
Seen ✗
Have I seen two documents describing the same species?
Yes
STOP

Figure 3.2 A simplified example of the difference threshold heuristic. Given the information need of finding different species of animal, a searcher issues a query, and examines a number of documents. The third document offers information on dogs, which has been already observed in **Doc 1.** Using the stopping criterion that once the same species has been observed twice, **Doc 5** satisfies it. This then means that the threshold has been met, and the searcher then stops.

As a simplistic example of this heuristic, a searcher is provided with an information need to find as many different species of animal as possible. Once a query has been issued, the searcher begins to examine documents on the SERP. This is illustrated in Figure 3.2, where the first document considers dogs. A simple criterion is employed whereby the searcher stops after encountering the same animal twice, illustrating that nothing new is being learnt from the list of results presented. Once this is met, the searcher abandons the SERP, and can then perform a query reformulation to discover different species of animal.

3.2.1.3 Magnitude Threshold

The magnitude threshold heuristic (Nickles, 1995) considers an individual's belief that the information accrued during the search process provides *sufficient evidence* to prompt him or her to stop searching for further information. The point at which the searcher would decide to stop (stopping criterion) is determined by some predetermined, internal threshold that must be reached (Wald, 1948; Nickles, 1995). Gettys and Fisher (1979) hypothesised that the searcher *"mentally tabulates"* the cumulative impact of the evidence that he or she has uncovered. When the tabulation crosses the predetermined threshold, he or she stops.

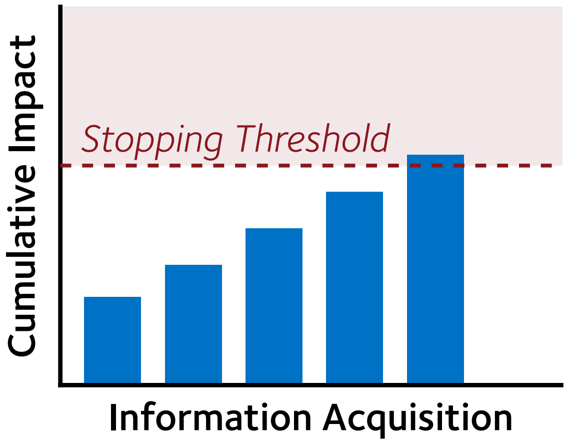


Figure 3.3 The magnitude threshold heuristic. Once a searcher accrues a predetermined level of impact, he or she stops. Adapted from Browne and Pitts (2004).

Determining what exactly this threshold should be before commencing a task has attracted research from several perspectives. This decision can be left open to interpretation by the individuals who choose to operationalise such a heuristic. However, research has shown that under different tasks, varying the criteria by which an individual bases their initial threshold value differs. For example, Busemeyer (1982) demonstrated this for decision making under uncertainty. Saad and Russo (1996) demonstrated the usefulness of this heuristic under common choice tasks. Considering prior knowledge of a topic may also impact upon the threshold chosen – a topic where a searcher has limited knowledge may mean a lower stopping threshold, for example.

An abstract representation of the stopping heuristic is provided in Figure 3.3. From the figure, we can see that a searcher accrues information through each document that is examined. This is combined together to form a *cumulative impact* of the information. For each document examined, the current cumulative impact value is compared against a predetermined threshold value. If the cumulative impact is above this threshold, the searcher then assumes that enough supporting evidence has been collected, and stops.

3.2.1.4 Single Criterion

The *single criterion heuristic* was later defined by Browne et al. (2005). As the name suggests, this heuristic considers a searcher examining information for a *single criterion* related to their information need, typically assumed to be the most important one. The searcher then stops examining content once he or she has deduced that enough information about said criterion has been accumulated for them to be satisfied. The concept of a stopping threshold can be borrowed from the magnitude threshold heuristic, discussed in Section 3.2.1.3. This considers that a searcher will stop once they have accumulated enough impactful information to satisfy their information need.

Browne et al. (2005) go on to outline an example search task where the single criterion threshold would be directly applicable. Their example considers purchasing a mortgage for a new house. Here, a searcher will explore the websites of various mortgage lenders in order to find the best deal for them. Given a mortgage deal, the most obvious criterion that an individual would look for would be interest rates. More attractive deals would be associated with lower interest rates. Of course, other factors may influence the decision, but this example ultimately demonstrates how the heuristic works in simplistic terms.

3.2.2 Reasoning-Based Heuristics

The second category of stopping heuristics as defined by Nickles (1995) are *reasoning based*. While searching and accruing information about a particular topic, a searcher is essentially developing a mental representation of the topic (Yates, 1990). As highlighted by Nickles (1995), these elements can include arguments constructed during informal reasoning, previously constructed arguments, or information evoked from the searcher's long-term memory. As such, Nickles (1995) devised a category of stopping heuristics that are dominated by the searcher's reasoning processes.

3.2.2.1 Representational Stability

The representational stability heuristic (Nickles, 1995) (with the phenomenon initially discussed by Yates and Carlson (1982)) concerns the notion that as a searcher acquires new information, his or her mental model of the underlying information need shifts and develops – but only up to a certain point. From this point, their mental model *stabilises,* and the searcher is said to have accrued enough information to satisfy or understand the (sub)topic.

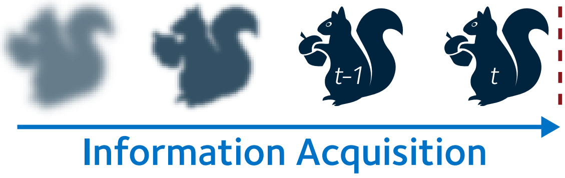


Figure 3.4 Example illustration of the representational stability stopping heuristic. The searcher's model of the given information need begins to stabilise at *t*-1, meaning that a searcher would stop at *t*.

It is stated by Nickles (1995) that while a searcher examines content, he or she generates arguments that serve to develop and elaborate his or her conception of the decision(s) that they are tasked to make. As the searcher continues to reason, certain arguments may be relegated to long-term memory due to the limited size of the searcher's working memory. Searchers will accrue new information, with some perhaps returning to the original subset of arguments. As mentioned previously, it is this point that can be interpreted as a form of stability regarding the searcher's mental model of their information need. This is depicted in Figure 3.4, where given a vague information need, a searcher will trawl a series of documents in order to develop their mental model of the given problem, turning their understanding of the topic from an initial *fuzzy* state to *crystal clear.*

3.2.2.2 Propositional Stability

Similar to the representational stability heuristic, Nickles (1995) also defined the *propositional stability* heuristic which again focuses on the concept of a stabilising mental model of the given information need. Here, a searcher when examining content will form a series of arguments from the information he or she is observing. These arguments can lead to *tentative conclusions*, from which at some point stability is achieved – and the conclusion does not change. Therefore, this heuristic suggests that the stabilised nature of the decision maker's conclusion from the information observed prompts him or her to stop.

3.2.2.3 The Mental List

The mental list stopping heuristic considers a mental list of aspects of some phenomenon. Each of the different aspects within the mental list must be *'checked off'* to a satisfactory

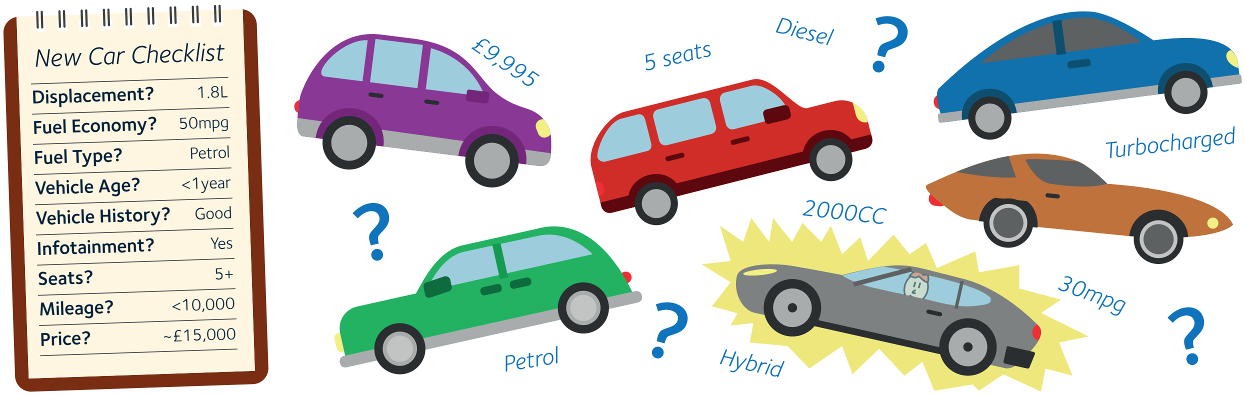


Figure 3.5 Given a well defined information need, Nickles (1995) outlined the mental list heuristic, where a number of different criteria must be satisfied before stopping. In the illustration above, car shopping is used as an example. Here, certain criteria for a new car (as shown on the notepad) must be met before a searcher is satisfied with what they have found.

level before the searcher then decides to stop examining content. This mental list can typically be constructed from a searcher's long-term memory, meaning that they will likely have *a priori* knowledge of the particular information need. So-called belief structures such as *schemas* (Bartlett and Burt, 1933) or *scripts* (Schank and Abelson, 1977) may assist the searcher in organising the construction of the mental list that forms the set of criteria that determines when they stop.

Figure 3.5 provides a graphical illustration of the mental list heuristic. When looking for a new car, a searcher will construct a mental list of different aspects of a car which are essentially the minimum requirements (e.g. a minimum engine displacement of 1.8 litres). Searching is then conducted, with the searcher narrowing down the potential choices available to them to those that satisfy their mental list.

3.2.3 Summary of Heuristics

In this section, a number of different stopping heuristics have been discussed from a number of seminal papers in the literature. While a much larger number of normative stopping heuristics have been defined in prior works, these have been omitted from the review as they do not adequately describe the cognitive behaviours of a searcher, often assuming a searcher has to *think ahead* to make a decision to stop or continue (Browne and Pitts, 2004). In contrast, the heuristics that are enumerated above do not make this assumption, making more realistic assumptions about the searcher's cognitive abilities.

Of course, the different stopping heuristics discussed above are likely to behave differently under different search contexts. As an example, the mental list heuristic might be impossible to use given a searcher with a very limited knowledge of a topic. He or she simply would not know enough information to ascertain key aspects of the topic and construct a set of criteria that must be met (Browne et al., 2005) – Gigerenzer and Goldstein (1999) also discuss this reasoning for the single criterion stopping heuristic. As such, it is hypothesised that the aforementioned stopping heuristics would likely work better with a searcher who is more knowledgeable.

Browne et al. (2005) also discuss the so-called *"structuredness"* of a given search task. If the task has well-defined inputs and outputs – or the goals and operations are clear and easily understood (Simon, 1996) – then it is hypothesised that searchers will employ more precise stopping heuristics for deducing when to stop. For example, the mental list and single criterion stopping heuristics might offer greater degrees of precision than (for example) the frustration and satisfaction heuristics, although the frustration and satisfaction heuristics may perform well for any given search task. Altogether, the heuristics discussed in this section would be applicable for informational search tasks (Browne et al., 2005) such as ad-hoc retrieval (refer to Section 2.3.1.1).

With the heuristics now enumerated, we later in this thesis discuss how we take these stopping heuristics and consider how to *operationalise* them, such that they can be subsequently implemented and compared against each other empirically. This also involves which of the two stopping decision points we discussed in Section 3.1.1 these operationalised heuristics can be used in. Chapter 5 provides explanations of the twelve *stopping strategies* that we employ in the contributory work in this thesis.

3.3 Theoretical Models

In addition to the stopping heuristics above, mathematically grounded, theoretical models have been defined that allow us to describe, predict and explain *how* and *why* searchers behave in the way they do. Crucially for this thesis, such models provide an explanation of their stopping behaviour. As discussed in Section 2.3.4 however, such models also have limitations, ranging from the low-level assumptions engaged by the different models, the variables that are considered or excluded from the models, and the difficulties arising from the complexities of human behaviour (Fishwick, 1995; Azzopardi and Zuccon, 2015).

Despite the limitations of such an approach, such *formal models* also permit the generation of different hypotheses regarding search behaviours. These can subsequently be empirically tested and validated – with examples of such studies including Azzopardi et al. (2013) and Pirolli et al. (1996). Three examples of such theories include *Information Foraging Theory (IFT)* (Pirolli and Card, 1999), *Search Economic Theory (SET)* (Azzopardi, 2011) and the *Interactive Probability Ranking Principle (iPRP)* (Fuhr, 2008). Central to the work in this thesis is IFT that we discuss in detail in the following subsection. As shown by Azzopardi and Zuccon (2015) however, the three theories are all mathematically equivalent, with all ultimately leading to the same understanding. As such, we do not discuss *Search Economic Theory (SET)* and *Interactive Probability Ranking Principle (iPRP)* in detail.

3.3.1 Information Foraging Theory

A well known conceptual model in the field of information seeking is the *berry picking model*, as proposed by Bates (1989a). As shown in Figure 3.6, this model considers searchers looking for information to be analogous to *foragers* scavenging for food in the wild. In the model, foragers are looking for the juiciest and ripest berries on a number of different bushes (or *patches*). The juiciest and ripest berries offer the highest levels of gain. Picking these berries

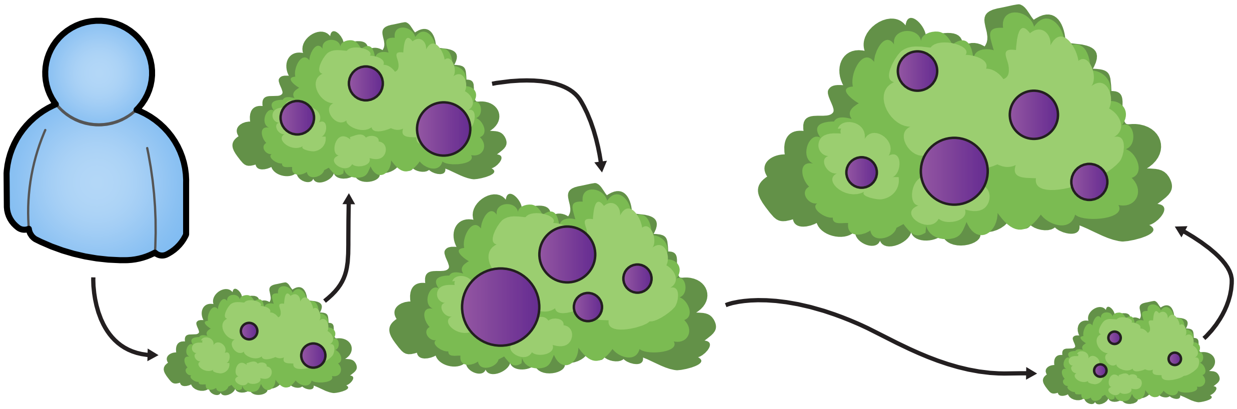


Figure 3.6 The *Berry Picking Model* Bates (1989a). A forager traverses through bushes to pick the juiciest berries for consumption. The model is high level and conceptual in nature, and thus does not provide any justification for *how* or *why* foragers search for the juiciest berries.

helps the forager maximise their level of gain. Applied to search, this construct means that a searcher *forages for information*, picking the most relevant (or juiciest!) documents that help them maximise their level of gain.

While the berry picking model is an intuitive and simple model to understand, its highly descriptive nature does not provide an explanation regarding the behaviour of the forager. *How long should a forager spend examining this berry bush?* This question cannot be answered as such by the model, but Bates (1989b) in a later publication does allude to the fact that searchers could weigh up the costs and benefits in order to decide what to do next.

Theories do however exist that attempt to explain the behaviour of a searcher when *foraging* for information. Initial attempts by Russell et al. (1993) and Sandstrom (1994) demonstrated that *Optimal Foraging Theory (OFT)* (Stephens and Krebs, 1986) could be potentially used to model the search process. This led to the development of *Information Foraging Theory (IFT)*, proposed by Pirolli and Card (1999). The theory provides an explanation as to how information foragers will behave, and as such, also provides a rationale as to how they will stop. IFT is extensively used in this thesis as a theoretical underpinning to several hypotheses. We also outline an optimal stopping heuristic, as well as several other time-based heuristics that derive from work associated with OFT.

3.3.1.1 Patches and Scent

IFT is comprised of three main models: the *information diet model*, concerning *what* information is consumed; the *information patch model*; and the *information scent model.* With the information diet model not considered in this thesis, we focus in this section on discussing the information patch and information scent models.

Central to IFT is the notion of a *patch*, as per the patch model. In the wild, a patch is modelled as an area of land with a degree of potential gain (food) that can be acquired by foraging through the said patch. The *between patch time* is the amount of time a forager spends moving towards a patch, and the *within patch time* is the time spent within the patch, examining its contents for potential gain.

With IFT, a patch can be modelled in a variety of ways. However, as outlined by Azzopardi and Zuccon (2015), the generally agreed approach to model a patch in terms of information seeking is to consider it as a SERP. With this representation, moving between a patch is akin to *issuing a query,* and thus incurs a cost. This is called the between patch time. Staying within a patch is the same as assessing result summaries on the presented SERP and their associated documents, with each summary and/or document taking a certain amount of time to process, or the within patch time. The patch model essentially predicts how long an information forager should stay in a patch (or SERP) before abandoning it and moving to the next patch.

However, given a series of patches (or potential queries), how does a forager deduce which one they should *enter next,* and examine in closer detail? This is described by the information scent model and encapsulates a currently active area of research. Figure 3.7 graphically illustrates the scent of a patch in action – given two patches as depicted in the illustration, which patch will the forager travel to next? Following the scent or *cues* on the ground next to him, the forager observes that the paw prints to patch ❶ are more prevalent, and thus will venture to that patch first. Like foragers in the wild, information foragers will observe

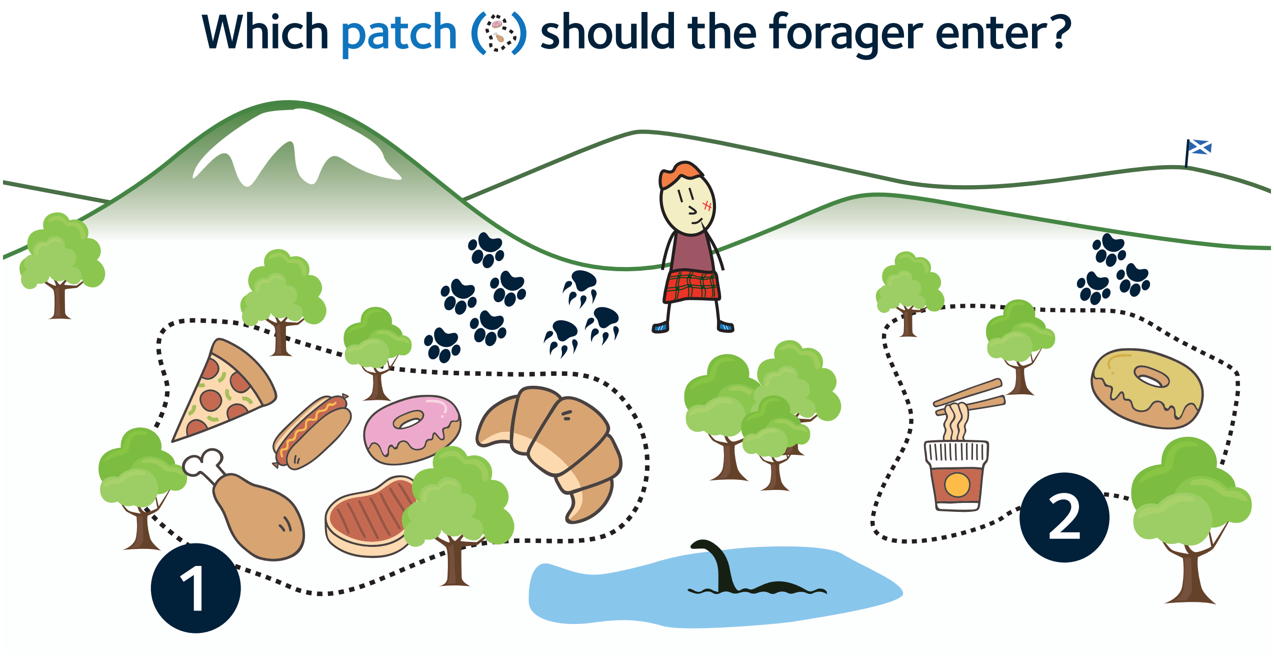


Figure 3.7 A graphical depiction of the *patch model*, part of *Information Foraging Theory (IFT)*. When presented with two patches, each containing food that can be represented as gain, what patch should our forager choose first — patch ❶ or ❷?

a series of *proximal cues* presented to them on a SERP, such as hypertext links, document titles, snippet text and thumbnails to locate information (Pirolli and Card, 1995, 1999; Chi et al., 2001; Olston and Chi, 2003; Pirolli, 2007). In the context of news search, cues were examined by Sundar et al. (2007). Here, cues such as the source of an article (its scent) were shown to have a powerful effect on the perception of the article, and influenced whether the said article was clicked on.

If these cues provide a rationale as to what leads to a promising scent trail, it follows that scent, in combination with patches, provides a rationale as to when a searcher will stop examining a set of results (Pirolli and Card, 1999; Wu, 2012; Wu et al., 2014). For example, a user study by Wu et al. (2014) demonstrated that a searcher would forage to greater depths if the SERP appeared to contain many relevant items. Card et al. (2001) also observed this trend. They found that when navigating through pages, searchers were more likely to leave when the information scent began to decline. Section 3.4 provides more details on these user studies, along with others considering stopping behaviours.

3.3.1.2 Stopping Heuristics

Given a patch with a scent, how can one deduce *when they should stop?* Like all theories, IFT makes some key assumptions from which we can deduce behaviours of a forager. The assumptions are that a forager will: enter a patch with what appears to be the highest yield first; and attempt to maximise their gain per unit of time. Given these assumptions, one would now be able to answer the question posed in Figure 3.7. With a better scent and greater volume of potential energy to be gained, patch ❶ is the answer that a forager would provide to the question *which patch should I explore first?*

In addition, the assumptions provided above allow us to begin formulating a stopping heuristic based on the *optimal behaviour* of a forager. The *Marginal Value Theorem (MVT)* by Charnov (1976) states...

> *"...that a forager should remain in a patch so long as the slope of the gain function is greater than the average rate of gain in the environment."*
>
> **Pirolli and Card (1999)**

The MVT implies that if a forager is within a patch that initially looked promising, yet is yielding a rate of gain less than the *average rate of gain expected within the patch,* he or she should then abandon the patch and then move to another one. This phenomenon is often called the *instantaneous intake* theorem (Stephens and Krebs, 1986). In the context of information seeking, this would imply a query reformulation. Conversely, a forager who has found himself or herself in a patch yielding gain at a rate *greater*

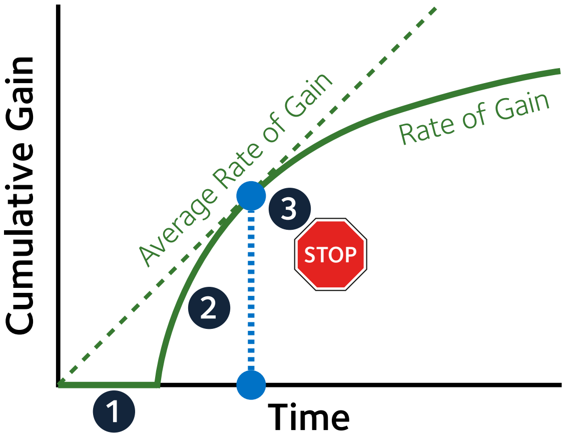


Figure 3.8 The IFT stopping heuristic. The searcher should stop when the rate of gain (solid **green** line) no longer outperforms the average rate of gain (dotted **green** line).

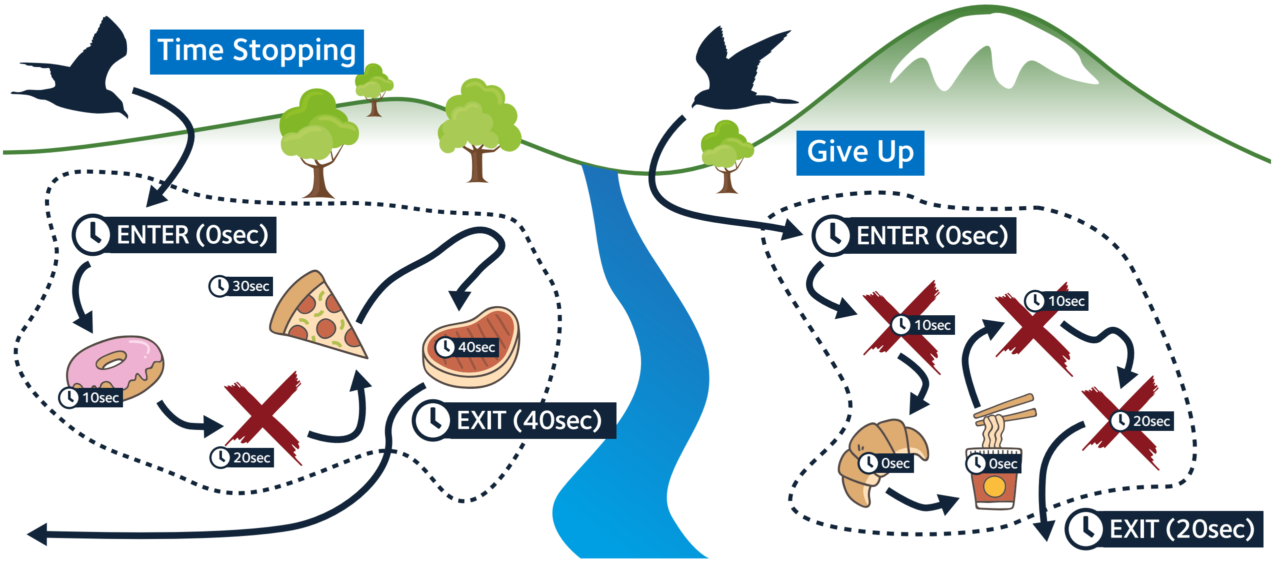


Figure 3.9 Illustrations of the *time stopping heuristic* (left), and the *give up heuristic* (right). On the left, a forager will stop after a time limit has been reached (40 seconds in this example) from the point at which they enter a patch. On the right, a forager will *reset* their timer when they encounter something gainful, but will grow increasingly impatient the poorer the results of their foraging, and eventually stop too (a 20 second limit is shown here).

than the average rate of gain would be best advised to stay within that patch. This is graphically illustrated in Figure 3.8, where the *gain curve* for a forager in a patch is highlighted in green. In addition, the plot illustrates: ❶ the between patch time, where the forager is not acquiring any gain; ❷ the within patch time, where the forager is examining the SERP and associated documents; and ❸ the optimal stopping point, based upon the MVT. Graphically, this is best described as the point at which the tangent to the curve (from the origin) touches the gain curve. From this point onwards, the rate of gain decreases and is less than the average rate of gain, meaning that the forager receives diminishing returns for the investment in examining content within the current patch (or SERP).

Operationalising the instantaneous intake theorem is often difficult to do in practice. How would one measure, for example, the expected rate of gain? Instead, several other stopping heuristics that influence patch leaving have been developed as part of OFT (Stephens and Krebs, 1986). These attempt to approximate the instantaneous intake theorem. Such heuristics include, but are not limited to:

- the so-called number heuristic (Gibb, 1958), where a forager would leave a patch after finding *n* prey;[2]
- the time heuristic (Charles-Dominique and Martin, 1972; Krebs, 1973), where a forager stops after spending *x* seconds within a patch; and
- the give up heuristic (Krebs et al., 1974), where a forager would stop and leave a patch after *x* seconds have elapsed since last finding something of use.

The time-based stopping heuristics are illustrated in Figure 3.9. A further study of different *patch types* (i.e. where the density of prey varies) was also undertaken by McNair (1982). They found that across different patch types, different stopping heuristics worked better in different environments – also demonstrated in works by Iwasa et al. (1981), McNair (1982) and Green (1984). Consequently, a further combination heuristic was devised. For a patch that appears to be fruitful early on, a satisfaction-based heuristic would perform well. Otherwise, employing the give up time-based heuristic (Krebs et al., 1974) would work best. This intuitively makes sense. A searcher, when presented with a SERP of high quality with many relevant results would be prudent to continue examining it for more content if the initial set of results are promising. However, if initial results are not promising, the searcher should be more sceptical, and be prepared to abandon it if, after examination, relevant content was not forthcoming as the results are traversed.

3.4 User Studies

While stopping heuristics provide a means for quantitatively characterising and predicting stopping behaviour (Wu et al., 2014), only a handful of user studies have been undertaken that attempt to understand when enough information is enough (Zach, 2005). As

[2] This stopping heuristic is analogous to the satisfaction and satiation stopping heuristics, defined by Cooper (1973a) and Simon (1955) respectively.

we have already discussed, stopping is an inherently difficult phenomenon to model effectively. This is because it is instrumented by a series of internal factors to the decision maker's thinking (Nickles, 1995). In this section, we detail a number of different user studies that have attempted to provide an explanation for a searcher's stopping behaviours.

3.4.1 Understanding Stopping Behaviours

Two user studies by Zach (2005) and Berryman (2006) have examined searcher stopping behaviours through a series of interviews with subjects. These studies primarily focused on the notion of *why* searchers stopped when they did, with both considering subjects seeking information in an academic work environment.

Zach (2005) considered how senior art administrators determined when to stop searching in their daily jobs, and found that they mostly stopped either because they:

- felt satisfied with the information that they had obtained during their search; or
- stopped because of time constraints.

The study by Berryman (2006) was conducted in a similar approach. Public sector policy workers reported finding it difficult to work out how much information would be enough to satisfy their tasks when initiating them. However, once the structure of what they needed to find had been established, the point at which they felt they should stop became clearer. The findings from this second study provide evidence that the assessments of what constitutes as *enough* can be difficult and complex to deduce. This finding also provides evidence of the development of an underlying mental model of the given information need and provides justification for the representational stability, propositional stability, and mental list stopping heuristics (as discussed in Section 3.2.2).

A number of user studies have also examined stopping behaviours in relation to the concept of satisfaction or satiation (Simon, 1955). As previously discussed in Section 3.2.1.1, this concept suggests that a searcher will cease searching as soon as conditions arise, instead of after they have exhaustively considered all available information (March, 1994).

Considering this approach, Agosto (2002) examined the decision-making abilities of young people when searching on the WWW. In this study, 22 9th and 10th grade students from a U.S. high school demonstrated limitations which affected their decision making, including time constraints that were imposed externally and internally, information overload, and other physical constraints. In order to find websites to help in satisfying their information need, the students used reductive approaches to decrease the amount of information presented on the WWW, and used this to work out when to stop. How students perceived the websites were also largely down to personal preference.

With a completely different set of subjects, Mansourian and Ford (2007) conducted a study where they analysed the stopping behaviours of 37 staff and students from four university biology departments, and classified their stopping behaviours by search depth and search impact. Qualitative results showed that subjects indicated that missing potentially important information in the course of their searching was a matter of concern. The authors reported that the estimations and importance of information missed likely would affect their stopping behaviour. From this, classifications of the perceptions of missing information ranged from *inconsequential* to *disastrous*, and search strategies classified as *perfunctory* to *extensive*, with the information need dictating what category the searcher would have considered appropriate.

A similar study by Prabha et al. (2007) considered searchers in a further academic library setting, with one key finding from their study showing that time constraints led to a decrease in the number of documents that searchers examined. Again, the specific information need and the searcher's role in academia affects every stage of their search processes – which includes affecting what they have found to be enough.

These findings were further demonstrated by Wu et al. (2014), who undertook a study where subjects performed a series of different search tasks. Subjects were then interviewed about their result summary level stopping and session stopping behaviours. Results from this study showed that result summary level stopping decisions were taken primarily on the face of search results, queries and search tasks. Session stopping decisions were determined by the subject's overall goal for each task, the content examined (and their subjective perceptions of the examined content) and the study constraints imposed upon them, such as time constraints and search interface restrictions. Further empirical evidence to this study was later provided by Wu and Kelly (2014). They reported that some subjects discussed *"forced stopping"* (stopping when no more information could be found), and *"voluntary"* stopping that stemmed from the feeling of securing enough information.

Wu et al. (2014) also discussed how information scent affects the stopping behaviour of a searcher. Constituting part of IFT, it is important to note that user studies have been conducted using this model. For example, Card et al. (2001) observed that if a person started with a high information scent web page, he or she would be inclined to visit more web pages on the high scented page's domain. They also found that as the information scent of web pages declined, there was a tendency for the person to leave the site or return to a previously visited page. Loumakis et al. (2011) examined scent that was associated with images presented on SERPs, and how these impacted on the evaluation behaviour of searchers. They found that when images were added to text snippets, participants reported increased confidence that they could find an appropriate result.[3]

Central to the findings of all of the above studies – regardless of the group of subjects or contexts in which the searches were conducted – is the idea that searchers stop when they are *satisfied*. Even though subjects of these studies were acutely aware of the fact they had not found *all* relevant information to their given information need, they were nevertheless satisfied with what they had found, and subsequently decided to stop. While the results from these studies may be underwhelming in terms of concrete explanations as to why people stop, they do provide invaluable insights, and demonstrate just how difficult it is to encapsulate or create descriptive parameters of such behaviour. Indeed, factors such as

[3] *A picture is worth a thousand words.*

time constraints, a searcher's information seeking ability and other factors all influence the internal stopping rules of a searcher, as was discussed by Marchionini (1995).

3.4.2 Quantifying Stopping Behaviours

With the above studies examining *why* people decide to stop, a very limited number of studies have attempted to quantify *when* a searcher should stop searching – something the stopping heuristics presented in Section 3.2 attempt to do. Toms and Freund (2009) studied the actions preceding the endpoints in information seeking to predict what actions would lead a searcher to stop. The most prevalent pattern they observed that matches the searcher models outlined in Section 2.3.5 consisted of a searcher:

1. issuing a query;
2. examining results presented to them on a SERP; and
3. viewing a document.

Interestingly, the authors observed that searchers appeared to be more engaged in page content and in revisiting and assessing pages that had already been found. They hypothesised that this again may be due to the satiation heuristic, where the searchers would purposefully go back to reassess if what they had found was *enough.*

A further study by Dostert and Kelly (2009) examined the stopping behaviours of 23 undergraduate students. Subjects, in parallel to other studies, reported that the primary factor for deciding to stop was their intuition. Like in the study reported by Prabha et al. (2007), the subjects were time constrained. Dostert and Kelly (2009) reasoned that subjects could not adequately articulate this intuition, but hypothesised that they simply felt that given their perception of how much time had elapsed, the number of documents that they had located felt sufficient. However, the authors report a number of additional reasons (as shown in Figure 3.10) why subjects decided to stop, with the reasons providing links back to the stopping heuristics defined in Section 3.2.

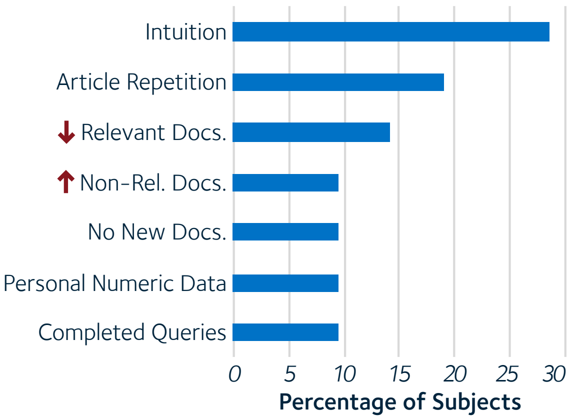


Figure 3.10 Responses of the survey on why subjects stopped by Dostert and Kelly (2009). Like most studies examining stopping behaviour, most subjects stopped because of their *intuition* — or what felt like *enough* to them.

Indeed, the unarticulated notion of finding *"enough"* (Zach, 2005) information links neatly back to the idea of the magnitude threshold stopping heuristic, proposed by Nickles (1995) and detailed in Section 3.2.1.3. With this heuristic, a searcher would stop once they have accumulated a certain predetermined amount of information. From the results of their study, Dostert and Kelly (2009) hypothesised that the threshold was reached once their subjects felt they had correctly identified *half* of the relevant documents available to them. In reality, the searchers had on average only managed to correctly identify 7.35%. In addition to comparisons to the magnitude threshold stopping heuristic, Dostert and Kelly (2009) also drew comparisons from their results to the difference threshold stopping heuristic, as outlined in Section 3.2.1.2. To recap, this heuristic considered a searcher's tolerance to not learning anything new. This is argued by the authors as a reason for respondents citing repetition in the documents found, or a lack of new documents. Lastly, the representational stability stopping heuristic as detailed in Section 3.2.2.1 was also noted by the authors. With this heuristic concerning the stabilisation of the searcher's underlying mental model of the topic, the authors noted that supporting evidence was obtained by subjects responding to a decrease in the number of relevant, and/or an increase in the number of non-relevant documents.

These stopping heuristics were also investigated by Browne and Pitts (2004) and Pitts and Browne (2004) with systems analysts during the process of information requirements determination. The analysts were required to gather a series of information requirements that would allow them to generate diagrams to represent an online grocery shopping system. Browne and Pitts (2004) found that more experienced analysts tended to use the mental list and magnitude threshold stopping heuristics, while less experienced analysts utilised the difference threshold and representational stability stopping heuristics. In addition to these findings, the authors noted that the applicability of different stopping heuristics resulted in varying degrees of quantity, depth and the quality of information obtained.

3.4.3 Considering Search Depths

A number of additional user studies have also considered the so-called *search depth* – that is, the depth on a list of ranked results that searchers stop clicking (the *click depth*). Studies such as the seminal work by Cutrell and Guan (2007) undertook an eye-tracking study and reported that subjects examined the first eight results before deciding to carry out a query reformulation. Lorigo et al. (2008) also examined their subjects' scan paths as they undertook search tasks. On average, subjects scanned only 3.2 distinct search results per query. This work was supplemented by Huang et al. (2011), where they found that subjects proceeded to issue a new query after inspecting the top four results of the presented SERP.

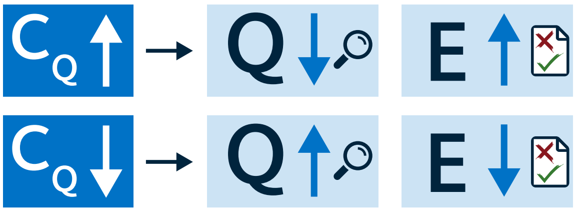


Figure 3.11 The *cost-interaction hypothesis* (Azzopardi, 2011). As the cost of querying increases (C_Q), searchers will issue fewer **Q**ueries and **E**xamine more documents per query.

A study by Azzopardi et al. (2013) also found that the depth to which subjects examined content was affected by the *cost* of entering a query (as illustrated in Figure 3.11). With a search interface where subjects were required to invest more effort to enter a query, significantly fewer queries were issued, with the results for these queries examined to greater depths. This was in contrast to subjects who used a standard search interface, where more queries were issued with subjects examining the content to a shallower depth. These findings comply with the *query-cost hypothesis* (Azzopardi, 2011), that states: *as the cost of querying increases, searchers will pose fewer queries and examine more documents per query.*

This is illustrated in Figure 3.11. Evidence from this study also demonstrated that the search interface individuals are subjected to impacts upon their stopping decision making.

3.5 Chapter Summary

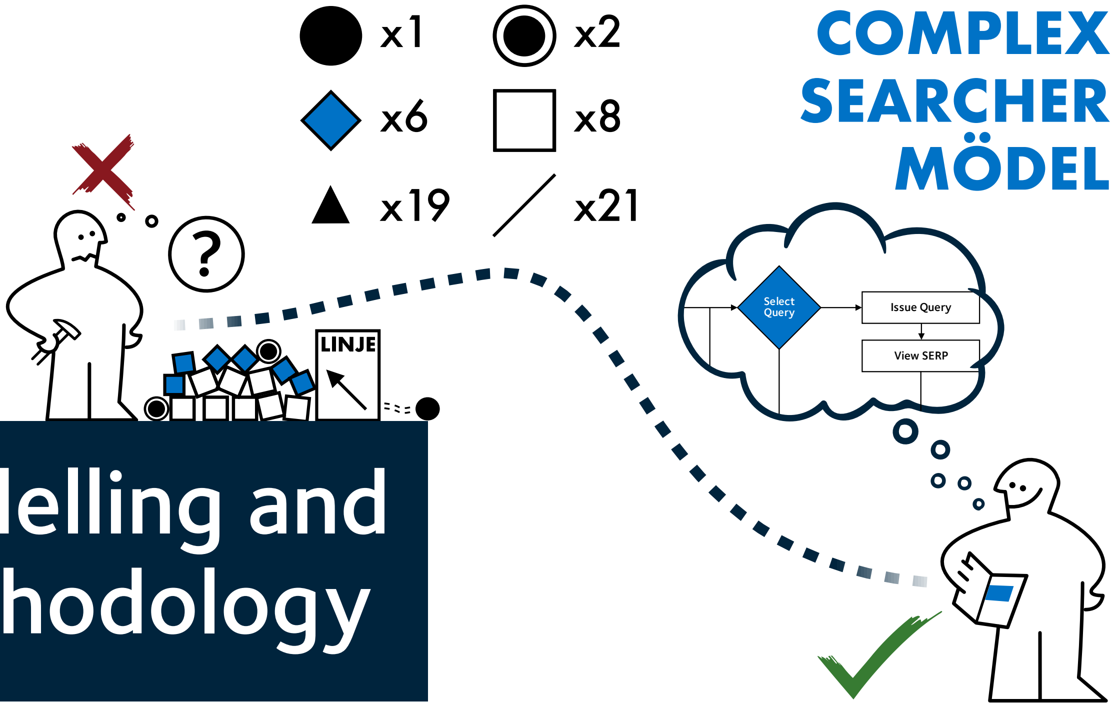
This chapter has provided an extensive overview of how stopping has been examined in the context of IIR. In particular, we have detailed a number of different *stopping heuristics* that have been proposed in the literature. These heuristics represent the attempts of researchers to capture a searcher's feeling of what is *"good enough"* (Zach, 2005). We also discussed theoretical models of search, examining in particular *Information Foraging Theory (IFT)*. This theory provides an explanation as to why and when searchers should stop, and extensive work in the literature based upon *Optimal Foraging Theory (OFT)* has also yielded a series of additional stopping heuristics.

We also provided an overview of the literature concerning user studies and searcher stopping behaviours. Many of these studies showed that searchers are simply unable to articulate why they stopped when they did, with internal heuristics causing them to stop when they simply felt satisfied, perhaps complying with the *satisfaction/satiation stopping heuristics* (Cooper, 1973b; Kraft and Lee, 1979) (or the number heuristic (Gibb, 1958)). However, these different internal stopping heuristics vary from person to person, with factors such as domain knowledge – and external factors such as time constraints – affecting their behaviours (Marchionini, 1995). As such, stopping behaviours are an intrinsically difficult phenomenon to capture and understand effectively.

With the scope and background of this thesis now outlined, we now move towards Part II. We begin to introduce the contributions that the work undertaken within this thesis provides, beginning with an updated searcher model. We will also discuss how we operationalised the *stopping heuristics* outlined in this chapter, turning them into a series of programmable *stopping strategies.*

Part II

Modelling and Methodology

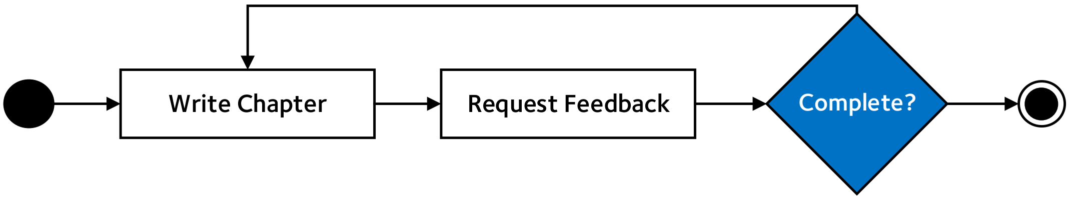


In the second part of the thesis, we introduce the Complex Searcher Model (CSM) that provides a more realistic, conceptual model of the information seeking process. We also detail a series of operationalised stopping strategies that we will apply as part of our general methodology. This details how we instantiate the CSM and other components in subsequent chapters.

Chapter 4

The Complex Searcher Model

In this chapter, we present the *Complex Searcher Model (CSM)*. The CSM is an updated, conceptual searcher model[1] that is one of the major contributions of this thesis. It is an amalgamation and development of prior, established searcher models. These models capture the complex sequence of interactions that take place between a searcher and a retrieval system over the course of a search session. As such, this chapter provides a partial answer to our first high-level research question, **HL-RQ1**.



As discussed in Section 2.3.5, earlier examples of searcher models include the Markov-based approach presented by Baskaya et al. (2013), and the model proposed by Thomas et al. (2014). These searcher models (along with others) are in broad agreement with the general sequence of events that take place within the IIR process – from issuing a query to examining documents for relevance.

[1] The CSM can also be considered as a *browsing model*, as per Carterette et al. (2011).

Given the aforementioned searcher models outlined in Section 2.3.5, the CSM offers a number of advancements in modelling searcher and retrieval system interactions. In this chapter, we provide:

- the flow of the proposed CSM, outlining the different steps and decisions that those subscribing to it undertake (Section 4.1);
- a discussion of the stopping decision points that the CSM considers (Section 4.2);
- a summary of the key advancements that the CSM provides (Section 4.3); and
- an outline of the key assumptions that we consider as part of the CSM (Section 4.4).

We also briefly outline the specifics for evaluating the CSM as a viable searcher model (Section 4.5). Specific details of the implementation of the CSM are discussed in our general methodology (Section 6.4, page 157). We begin this chapter with a discussion of the flow of the CSM, discussing the different steps and decisions that searchers will make.

4.1 Model Flow

The CSM is illustrated as a flowchart in Figure 4.1. It is comprised of a number of different activities denoted by boxes, and decisions represented as blue diamonds. The flowchart is divided up into a number of different blocks, labelled A to F. Each of the blocks denotes a logical set of interactions – block B, for example, considers all of the actions and decisions a searcher is likely to consider in relation to querying. In this section, we outline the flow of the CSM, discussing the key activities and decisions that searchers would undertake when subscribing to it. This is done in relation to the six labelled blocks that are discussed below.

A **Topic Examination** A searcher subscribing to the CSM would begin the search process with some information need. This would typically be provided as a *topic*, with

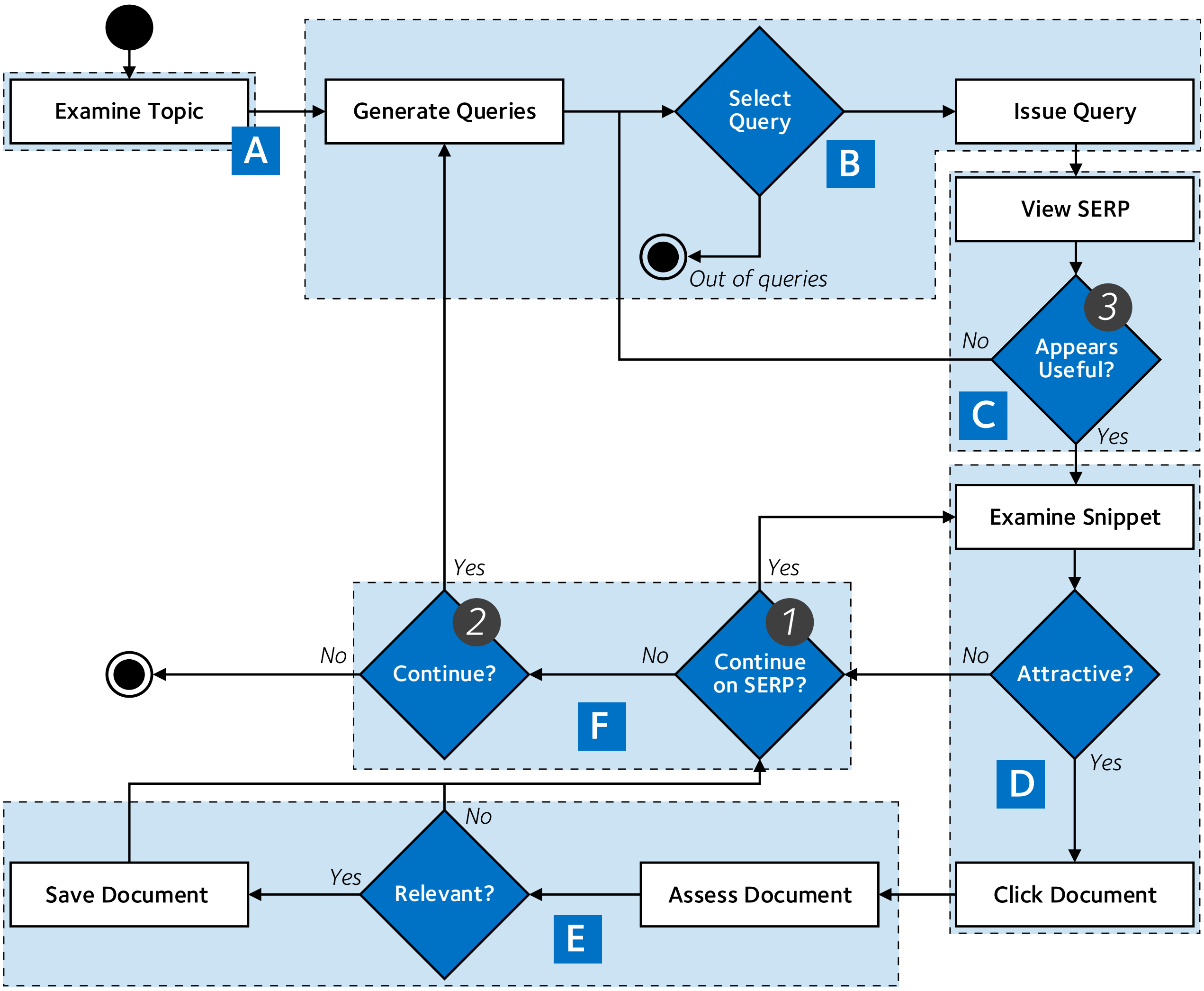


Figure 4.1 A flowchart of the *Complex Searcher Model (CSM)*. A cornerstone of this thesis, the CSM is extensively used as the grounding model for simulations of interaction that we report on in subsequent chapters. The main logical components of the CSM as discussed in Section 4.1 are labelled A to F, complete with surrounding boxes. The *three* stopping decision points are highlighted with numbers ❶, ❷ and ❸ (refer to Section 4.2).

a topic description outlining said information need. From this topic description, various entities can be extracted and used for the generation of queries, as described in block B.

B **Querying** Once the information need has been established, the searcher will then move onto the *querying* block. Here, a number of different activities and a decision point are considered. Within the block, the first activity that the searcher will undertake is the **generation of queries**. Given the information need from block **A**, a searcher will formulate a number of *candidate queries* that they could issue to the underlying retrieval system. This is achieved through the use of some form of *querying strategy* that generates the said candidate queries. The searcher then must make a decision as to what query they should issue. A query is **selected** from the candidate queries list by some process (e.g. some form of ranking). This query is the one the searcher believes is most likely to return relevant documents. The query is then **issued** to the underlying retrieval system[2], with the searcher proceeding to **C**.

As stated previously, IIR is an iterative process where multiple queries can be issued in a single search session. The CSM provides support for this – as can be seen from the flowchart line from block **F** to querying block **B**. At each point, the list of candidate queries generated could theoretically be regenerated, thus supporting query reformulation. If a searcher then finds that the candidate queries list has been exhausted, a stopping point is provided for this scenario.

C **SERP Examination** With the query now issued, the retrieval system will then return a SERP for the searcher to examine. From here, the searcher is able to **view the SERP** – that is, to obtain an *initial impression* of the SERP by examining the various *proximal cues* (Chi et al., 2001) presented. If the SERP does not appear to look promising, or gives the answer straight away, the searcher will abandon the SERP and proceed to issue a further query from the list of candidate queries as described in block **B**. If the SERP however does look **useful**, the searcher will then *enter* the SERP and proceed to examine individual result summaries in detail.

D **Result Summary Examination** Result summaries are presented to the searcher within the SERP. Searchers can take individual result summaries in turn, examining the title and snippet text provided for **attractiveness**. If deemed to be sufficiently attractive to warrant further examination, the searcher will then click on the provided link. This

[2] As the CSM considers interactions with a retrieval system only, we assume that a searcher will have already selected a retrieval system to use beforehand as discussed in Section 4.4.2.

link will then take the searcher to the associated document for further examination. If the searcher does not deem the summary to be sufficiently attractive to warrant further examination, he or she will then move to block F. As described below, the searcher in this block must decide whether to continue examining results on the SERP – and if not, whether to continue with the search session.

E **Document Examination** Once a searcher clicks on an attractive result summary, he or she will then assess the associated document for relevance, after which a further decision must be made. *Is this document relevant to the information need?* If so, the document is saved, meaning that it is added to a list of saved documents, as we describe below. The searcher then proceeds to block F. If not considered relevant, the searcher then proceeds directly F.

F **Deciding to Stop** Regardless of how the searcher reaches this block (either from block D or E), a searcher here can make two key stopping decisions. The first considers whether he or she should remain on the present SERP. If this is decided to be the case, the searcher will then move to the next result summary presented within it, and begin to examine that for attractiveness. If it is decided not to remain on the SERP, the searcher will then move to a final decision – *should I stop this search session, or continue?* If the searcher decides to continue the search session, he or she will then move back to the query generation activity in block B, beginning the cycle again.

Of particular interest to the work in this thesis are the *stopping decision points*, as discussed above – and shown in blocks C and F in Figure 4.1.

4.2 Stopping Decision Points

Outlined previously in Section 3.1.1, established searcher models consider stopping from two key perspectives: *result summary level stopping*, and *session level stopping*. The two established stopping decision points are included within the CSM, and are labelled ❶ and ❷ in Figure 4.1 respectively. They are also briefly outlined below.

- ❶ **Result Summary Level Stopping** This stopping decision point concerns the depth at which a searcher will stop examining a list of ranked results for a given query, assuming that results are ranked in a particular order. After stopping at this point, the searcher can continue the search session by issuing a further query.
- ❷ **Session Level Stopping** This second stopping decision point considers the point at which a searcher will stop their search session in its entirety. As an example, a searcher will stop their search session when they believe that they have satisfied their search goal, for example.

The CSM however includes a third, *SERP level stopping decision point*, highlighted as stopping decision point ❸ within block C of Figure 4.1.

- ❸ **SERP Level Stopping** With this new stopping decision point, a searcher can abandon a SERP before *entering* it to examine result summaries in detail.

This new stopping decision point permits searchers subscribing to the CSM to become savvier with their interactions. By gauging the SERP, a searcher can make an informed decision as to the quality of said SERP before making the decision to invest more time examining its contents, or simply cutting their losses and abandoning it – for better or for worse. The new stopping decision point is one of the key advancements that the CSM provides, and is discussed in more detail in Section 4.3.1.

4.3 Model Advancements

The CSM provides two novel advancements in modelling interactions between a searcher and retrieval system. These are highlighted as blocks B and C in Figure 4.1, and advances our understanding of the *querying* process – as well as including the aforementioned third stopping decision point. In this section, we discuss each in turn. While the advances to querying are novel, they are not the core focus of this work, and thus discussion of the new SERP level stopping decision point will be in greater depth.

4.3.1 SERP Level Stopping

This new stopping decision point – illustrated in block C of the CSM (Figure 4.1) – is motivated by the idea of the information scent (refer to Section 3.3.1.1 on page 92) present on a given SERP. This section also introduced the concept of *proximal cues* (Chi et al., 2001), cues that provide insights into whether the presented SERP will yield information that will aid the searcher in satisfying their underlying information need. This has been demonstrated in prior studies (Wu et al., 2014; Ong et al., 2017) – and in Chapter 7 of this thesis.

By operationalising information scent as the perceived performance of a given SERP, we allow a searcher to obtain an *impression* of the SERP before deciding to *enter* it and examine presented content in detail – or *abandon* the SERP altogether, and move to the next activity. The notion of forming an impression is similar to the summary impressions formed by searchers subscribing to the model defined by Thomas et al. (2014), as detailed in Section 2.3.5. In their model, a searcher would not form an overview of the SERP, but rather an impression of each individual result summary. The impression can then be used as a means of gauging whether further examination would be worthwhile.

This new stopping decision point is analogous to the well-studied phenomenon of *SERP abandonment* in which limited interaction occurs with the searcher. This has been typically assumed to provide an indication of the searcher's *dissatisfaction* with the presented results (Das Sarma et al., 2008; Chuklin and Serdyukov, 2012; Kiseleva et al., 2015), or *satisfaction* (through the concept of *good abandonment*) (Loumakis et al., 2011; Wu et al., 2014).[3]

[3] We discuss this in more detail in Section 4.4.4.

Thus, we provide, to the best of our knowledge the first searcher model that incorporates a path for a searcher to leave a SERP that appears to be of poor quality (or a *low scent*), or even satisfies their information need outright.

4.3.2 The Querying Process

Outlined previously, search sessions are inherently interactive (Ingwersen and Järvelin, 2005). During a session, a searcher's underlying mental model of a given information need can adapt and is likely to change as he or she examines new content for relevance (Borlund, 2003). Searchers may find more descriptive terms associated with the said information need, and incorporate these terms in a subsequent query reformulation.

From block B in Figure 4.1, the querying process has been broken up into two distinct activities and decisions: query generation (thinking of queries that could be issued) and query selection (selecting a query to issue). A searcher subscribing to the model will have the capability of revising their generated query list at each query reformulation, thus supporting the concept of the dynamic information need. Updated terms can in theory be selected from newly examined documents and incorporated within the query generation process for future reformulations.

Query selection then determines which one of the generated queries are to be issued to the retrieval system. Of course, the potential exists whereby all generated queries have been exhausted. This scenario would thus provide a natural stopping point for the searcher, as included in Figure 4.1.

4.4 Model Assumptions

When modelling a real-world phenomenon, a number of different assumptions are made about said phenomenon's exhibited behaviours (Fishwick, 1995). The CSM is no exception from this rule, and we make a number of different assumptions about a searcher's behaviours and the presentation of the retrieval system's results. This section details five key assumptions that we consider as part of the CSM.

4.4.1 Search Task

In this thesis, we are interested in the wider IIR process, considering all of the activities and decisions involved. We are particularly interested in improving our understanding of complex retrieval tasks.

The CSM provides scope for the modelling of a variety of different *interactive search tasks.* Examples of these include the aforementioned *ad-hoc,* exploratory, and *diversity tasks.* These tasks can be undertaken in different search *domains,* such as informational (refer to Section 2.3.2) or patent searching. As discussed in Section 1.2, we exclusively consider informational search in the domain of news. Tasks we consider include both ad-hoc and diversity, such that we can then examine how behaviours vary under each task. This is because while the CSM is able to model other search tasks, the selected search tasks provide for more interesting task types to examine, and consider a greater depth of activities and decisions that would not otherwise be examined by the more simplistic approach.

These tasks are interesting to examine for two key reasons:

- the search goals between each task vary; and
- from an examination of the literature (refer to Section 3.4), it is not clear when *enough information is enough.*

These reasons will undoubtedly produce interesting results between the two tasks. As the tasks are not simple lookups, a searcher will not stop once a single relevant page has been found. Instead, he or she will stop once *enough* (Zach, 2005) information has been found to satisfy their goal, or other constraints are imposed (e.g. time constraints).

4.4.2 Retrieval System Tool Choice

The searcher model proposed by Thomas et al. (2014) provides those subscribing to it with a choice as to which retrieval system they should use. As discussed earlier, we assume with the CSM that a searcher uses a single retrieval system. Our focus is therefore with the interactions that take place between the searcher and the said retrieval system.

Of course, the inclusion of such a decision point would be interesting to examine within the wider IIR process. Different retrieval systems will have benefits and drawbacks for particular domain types (e.g. a patent retrieval system would perform better for patent searching tasks). It would be interesting to examine this kind of *tool switching behaviour* – and could be considered as a further stopping decision point, or *retrieval system stopping*.[4]

4.4.3 Simple SERPs

When considering the SERP presented to the searcher as a whole, we make three simplifying assumptions within the CSM. These are enumerated and detailed below.

- **Ten Blue Links** Under the CSM, a SERP will consist purely of a set of result summaries, coined in IR literature as the *ten blue links*. Of course, we acknowledge that additional components are present in contemporary SERPs, such as multimedia content in federated search (Chen et al., 2012). These are however not considered to simplify the CSM.
- **Linear Examination Order** Once a searcher has decided to examine a SERP in detail, the result summaries presented to the searcher will be examined in a linear order. There is evidence to suggest that real-world searchers examine results from top to bottom, as demonstrated by Joachims (2002) and Joachims et al. (2005), for example.

[4] Refer to Section 10.3.1 on page 349 for a more in-depth discussion on *tool switching*.

Click models, such as the *cascade model* (Craswell et al., 2008), have been developed that utilise this assumption. Such approaches are subject to *positional bias,* where the searcher implicitly trusts the results of the retrieval model and assumes that the first result presented is the most relevant to their information need.

- **No Explicit Pagination** The CSM also assumes that the SERP presented to a searcher is of a single page, with no pagination of results. This does simplify the modelling process, with pagination activities and costs also not considered in earlier searcher models that consider the session as a whole.

4.4.4 Good and Bad SERP Abandonment

As previously mentioned, the CSM provides a third SERP level stopping decision point. Associated literature considers the notion of bad SERP abandonment, where a searcher is dissatisfied with the presented results. More contemporary research has introduced the notion of good SERP abandonment (Khabsa et al., 2016), where a searcher satisfies his or her information need by examining the SERP, requiring no further interactions with it. This is more prevalent on small-screen devices, where an information card presented to the searcher on the SERP may provide all the information required to satisfy the searcher on a simple lookup task, for example.

The CSM does not explicitly consider the notion of good or bad SERP abandonment; the provision exists however for both to be modelled effectively within the scope of the new stopping decision point. Good abandonment can be for example catered for with the inclusion of an additional decision point after determining a result summary to be attractive; the searcher could then make the decision to abandon the SERP if they feel satisfied with the result obtained. This is illustrated as an excerpt of a searcher model flowchart in Figure 4.2. The excerpt demonstrates the result summary **Attractiveness** decision point, the additional decision point determining **Satisfaction** with the result, and the final decision point that determines whether the searcher should abandon the SERP.

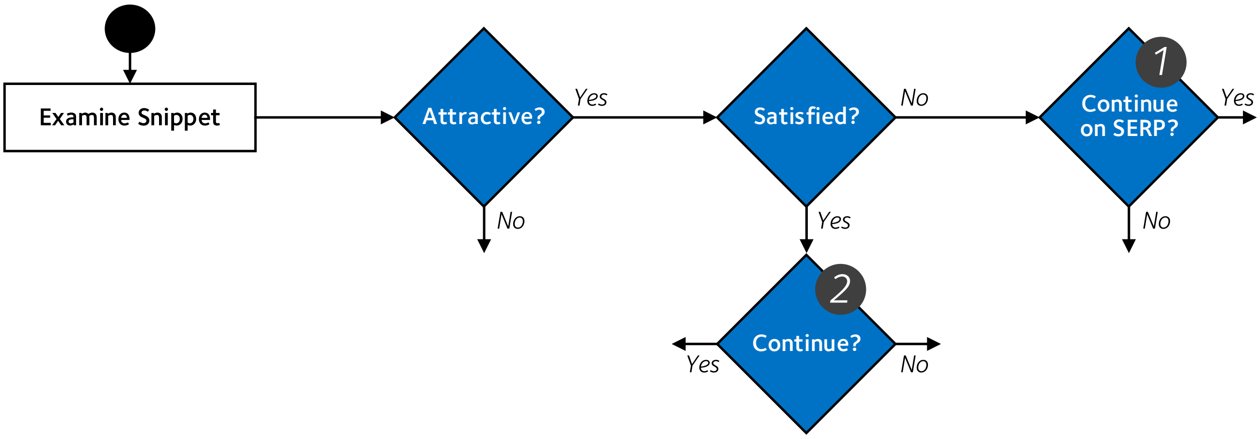


Figure 4.2 The interaction processes that can provide for incorporating *good SERP abandonment*, where a searcher satisfies his or her information need by simply examining a presented result summary. This is opposed to bad SERP abandonment, where the searcher will abandon a SERP if he or she feels the presented results are not of good quality. Upon examining a result summary **(Attractive?)**, a searcher will then determine if the summary addresses their information need **(Satisfied?)**. If so, they reach the session level stopping decision point ❷. Otherwise, they reach the result summary level stopping decision point ❶.

However, for the work in this thesis, we assume a simple SERP consisting only of a ranked list of results. We also assume that searchers subscribing to the CSM will have complex information needs, as discussed in Section 4.4.1 above. As such, we assume that the elements provided as part of the simplistic SERP are unlikely to fully satisfy their information need, and thus we consider SERP abandonment in this thesis exclusively from the perspective of bad abandonment. This is further discussed in Section 4.5.

4.4.5 External Factors

Given the flowchart of the CSM in Figure 4.1, it is clear that the model is completely agnostic of *external factors* that could influence how an individual behaves when searching. Kelly (2009), for example, cited that:

> *"searcher behavior [sic] can be governed by a number of external factors. For instance,*

> *the occurrences of a holiday or a project deadline will likely change the kinds of behaviors users exhibit and these behaviors may not represent their typical behaviors."*
>
> Kelly (2009)

These examples allude to time pressures, but there are a virtually unlimited number of other external reasons that may influence a searcher's behaviour. Even simple everyday occurrences such as a phone call or an incoming e-mail can sufficiently distract the searcher to the point that their behaviours are altered. Our assumption is that an individual searches in a more controlled environment, where they are exclusively tasked to search.

4.5 Evaluating the CSM

The CSM is presented as a generalised, conceptual model of the search process. It captures the key activities and decisions that a searcher must undertake. Given the current searcher models presented in Section 2.3.5, the CSM introduces further levels of complexity and realism into searcher models. Given our choice of search tasks, types, and assumptions, four key assumptions are made for the evaluation of the model.

- **Costs** We assume that a searcher will incur some cost when performing an individual activity within the CSM. For example, a document examination cost will be incurred when a searcher decides to examine a document for relevance.
- **Bad Abandonment** As described previously, a searcher subscribing to the CSM will only abandon a SERP if they consider it to be of poor quality. Given the complex information needs we consider in this work, this is a reasonable assumption to make.
- **Saving Documents** Documents that a searcher subscribing to the CSM will be saved to a list. This provides us with a mechanism of identifying content the searcher deems relevant, which can be used in calculating performance measures (see below).

- **Accruing Gain** Following on from the above, we assume that searchers only gain from documents they examine and save. We do not assume that a searcher will be able to gain from the examination of result summaries, for instance – the information need is complex, and short result summaries would be unlikely to provide an answer to their information need.

4.6 Chapter Summary

This chapter has proposed the *Complex Searcher Model (CSM)*, our solution to partially addressing **HL-RQ1**. Building on prior searcher models, the CSM proposes different advancements to modelling a searcher's interactions, the main development being the inclusion of a new, SERP level stopping decision point. We have outlined a number of different assumptions that we make in the CSM, and also discussed some evaluation considerations related to the work in this thesis. Empirical work presented later in Chapter 9 tests the CSM, providing evidence to support **HL-RQ1** in that the CSM does provide improvements over current searcher models.

In the next chapter, we turn our attention to the twelve stopping strategies that we operationalise and subsequently implement. These then allow us to operationalise the stopping decision points of the CSM.

Chapter 5

Operationalised Stopping Strategies

In Section 3.2, we discussed a number of different stopping heuristics that have been defined in the literature. In this chapter, we take a number of these stopping heuristics forward to produce a series of different **stopping strategies**, providing an answer that addresses **HL-RQ2**.[1]

```
SET counter TO 0
SET k TO 3
WHILE counter <= k
    IF non-relevant THEN
        SET counter TO counter + 1
```

These stopping strategies are operationalised versions of their corresponding heuristics. This means that we can subsequently implement and evaluate their effectiveness. We consider twelve different stopping strategies across seven different categories, the categories being:

- **fixed depth**, which assumes a searcher examines to a fixed depth – and is also considered to be our baseline approach;

[1] Refer to Section 1.2 on page 10 for the definition of the research question.

- **frustration**, considering a searcher's *tolerance to non-relevance;*
- **satisfaction**, taking into consideration how satisfied a searcher feels with what they have found;
- **difference**, which operationalises how *different* new content appears to previously observed content;
- **IFT**, which considers a searcher's *instantaneous intake;*
- **time-based**, which utilise time as a measure for stopping; and
- **measure-based**, considering an established IR measure as a stopping strategy.

In the remainder of this chapter, we consider each of the seven categories enumerated above. For each category, we discuss the different operationalised stopping strategies that we use for the empirical work reported later in this thesis. Before this, we begin with a brief discussion about the different stopping decision points that were outlined in Section 4.2, and the notation used herein when describing the different stopping strategies.

Stopping Decision Points An open question that we have not yet addressed is that of what stopping decision points (of three presented in Section 4.2 on page 111) we will operationalise with the stopping strategies presented in this chapter.

For the purposes of this thesis, we consider the twelve operationalised stopping strategies purely in the context of **result summary stopping** – or considering the depth to which a searcher will examine a list of ranked results. The stopping strategies will be examined in tandem with SERP and session level stopping. These are left for implementation decisions as outlined in later chapters.

Selecting Stopping Heuristics Given all of the different stopping heuristics proposed in Section 3.2 beginning on page 78, a further open question about this work is: *how do you choose what heuristics to operationalise?* Stopping heuristics were selected that we believed would offer good levels of performance for complex search tasks, where the onus was on the searcher to find and learn about a particular topic. Several of the reasoning-based stopping heuristics (such as the mental list heuristic, presented in Section 3.2.2.3) were not selected as operationalising them would have been too prohibitively complex.

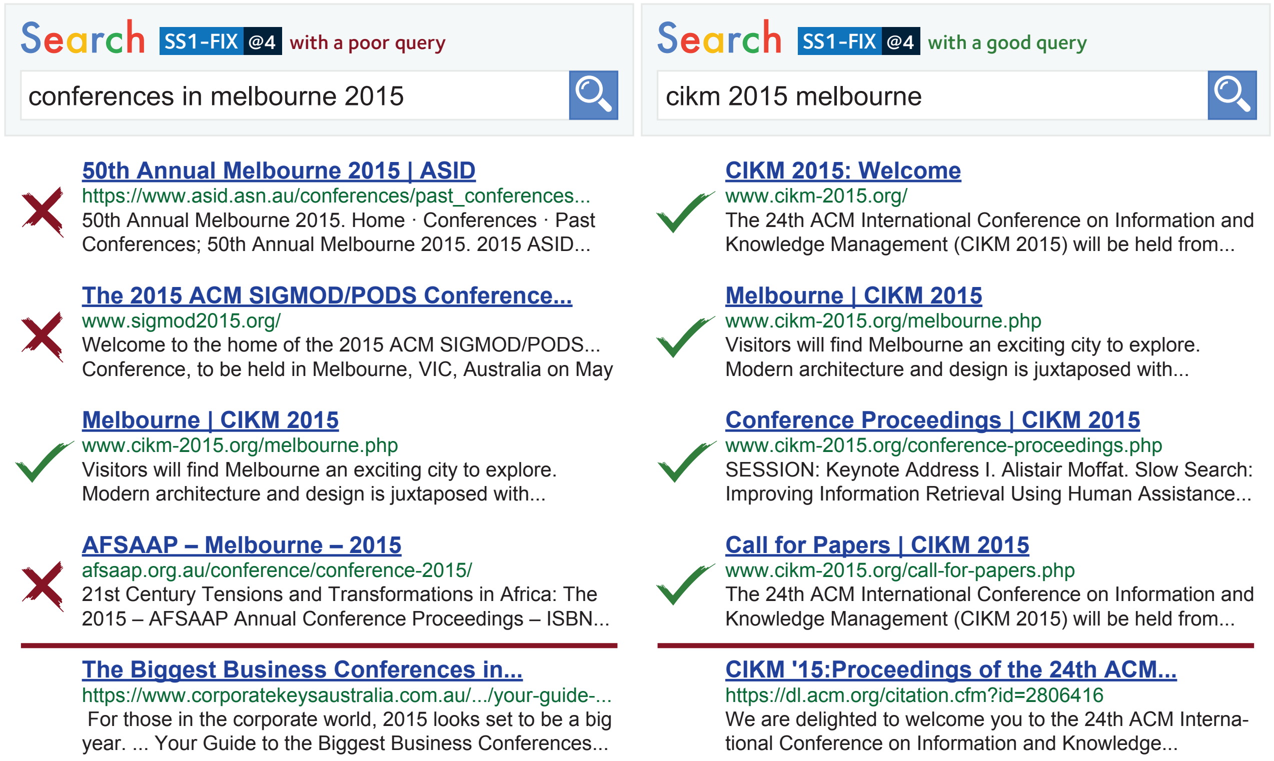
A Note on Notation Each of the operationalised stopping strategies that are introduced in this chapter comes complete with at least one *stopping threshold* variable, allowing one to customise the point at which a searcher subscribing to a given stopping strategy should stop. As demonstrated in the **Presentational Conventions** front matter, the notation we use to illustrate a stopping strategy and its threshold(s) is NAME @THRESHOLD. As an example, SS1-FIX @3 denotes the fixed depth stopping strategy SS1-FIX, set to a threshold of 3. This stopping strategy is outlined below in Section 5.1.

5.1 Fixed Depth

The fixed depth stopping strategy is based upon an assumption held across many of the models and measures widely used throughout the IR community. The assumption is that a searcher will browse to a *fixed depth* before stopping when examining a list of ranked results. *P@k*, defined in Section 2.4.1.1, is a prime example of this, and has been used in many different studies examining the simulation of interaction. Given the wide use of this fixed depth approach in historical and contemporary IR and IIR research, we consider this stopping strategy as the baseline approach to which we will compare more advanced (and *adaptive*) stopping strategies.

- SS1-FIX Fixed Depth A searcher employing this stopping strategy will stop searching once they have observed x_1 result summaries (i.e. SS1-FIX @x1), regardless of the relevance of each judged result summary.

SS1-FIX is a naïve stopping strategy as it assumes that all documents up to rank x_1 are



Search SS1-FIX @4 with a poor query

conferences in melbourne 2015

✗ **50th Annual Melbourne 2015 | ASID**
https://www.asid.asn.au/conferences/past_conferences...
50th Annual Melbourne 2015. Home · Conferences · Past Conferences; 50th Annual Melbourne 2015. 2015 ASID...

✗ **The 2015 ACM SIGMOD/PODS Conference...**
www.sigmod2015.org/
Welcome to the home of the 2015 ACM SIGMOD/PODS... Conference, to be held in Melbourne, VIC, Australia on May

✓ **Melbourne | CIKM 2015**
www.cikm-2015.org/melbourne.php
Visitors will find Melbourne an exciting city to explore. Modern architecture and design is juxtaposed with...

✗ **AFSAAP – Melbourne – 2015**
afsaap.org.au/conference/conference-2015/
21st Century Tensions and Transformations in Africa: The 2015 – AFSAAP Annual Conference Proceedings – ISBN...

The Biggest Business Conferences in...
https://www.corporatekeysaustralia.com.au/.../your-guide-...
For those in the corporate world, 2015 looks set to be a big year. ... Your Guide to the Biggest Business Conferences...

Search SS1-FIX @4 with a good query

cikm 2015 melbourne

✓ **CIKM 2015: Welcome**
www.cikm-2015.org/
The 24th ACM International Conference on Information and Knowledge Management (CIKM 2015) will be held from...

✓ **Melbourne | CIKM 2015**
www.cikm-2015.org/melbourne.php
Visitors will find Melbourne an exciting city to explore. Modern architecture and design is juxtaposed with...

✓ **Conference Proceedings | CIKM 2015**
www.cikm-2015.org/conference-proceedings.php
SESSION: Keynote Address I. Alistair Moffat. Slow Search: Improving Information Retrieval Using Human Assistance...

✓ **Call for Papers | CIKM 2015**
www.cikm-2015.org/call-for-papers.php
The 24th ACM International Conference on Information and Knowledge Management (CIKM 2015) will be held from...

CIKM '15:Proceedings of the 24th ACM...
https://dl.acm.org/citation.cfm?id=2806416
We are delighted to welcome you to the 24th ACM International Conference on Information and Knowledge...

Figure 5.1 An example of the fixed depth stopping strategy, stylised in this thesis as SS1-FIX. Here, a searcher has an information need for the conference *CIKM 2015* in Melbourne, VIC, Australia. The left example shows the top five results for a poor performing query, with few unattractive results (denoted by ✗); conversely, the right shows results for a query performing well, with many attractive results (denoted by ✓). With SS1-FIX @4, the searcher will stop at a depth of *4*, regardless of the perceived relevance of the content provided.

considered attractive enough for a searcher to consider examining in closer detail. On average, this strategy does make sense. However, on a per-query basis, this strategy appears counter-intuitive and would be a waste of the searcher's time.

For example, Figure 5.1 demonstrates two SERPs side by side. Given a searcher's desire to find pages providing information to *CIKM 2015*[2], two queries are issued. The query on the left yields poorer results than the query on the right, denoted by the ✓ and ✗ that denote relevant and non-relevant results respectively. With SS1-FIX @4, four result

[2] CIKM 2015 was a conference held in Melbourne, VIC, Australia in October 2015. The paper that initially presented many of the different stopping strategies outlined in this chapter was presented by the author at that conference. Refer to Maxwell et al. (2015b).

summaries are always examined before stopping, regardless of the perceived quality of the results. Examining four documents for the query on the results list on the left is a waste of the searcher's time, with lots of non-relevant material. A searcher would be better *adapting* his or her behaviour depending upon the perceived quality of the ranked list.

5.2 Frustration and Satisfaction

We referred to SS1-FIX as a *fixed* stopping strategy, as it is not *adaptive*. The remaining stopping strategies presented in this chapter (with the exception of SS9-TIME) are considered to be adaptive as they permit a searcher to adapt their stopping depth depending upon the result summaries that they observe in a ranked list. In this section, we propose three adaptive stopping strategies that are based upon a searcher's *tolerance to non-relevance* and a simple *goal-based* approach.

5.2.1 Searcher Frustration

We first discuss how the frustration stopping heuristics are operationalised, as outlined in Section 3.2.1.1. Given a set of result summaries presented on a SERP, how many unattractive summaries would a searcher be prepared to examine before becoming frustrated with the SERP, and abandoning it? This stopping heuristic attempts to address this question. Indeed, as detailed in Section 3.2, a number of researchers have proposed stopping heuristics that consider unattractiveness.

The frustration heuristic intuitively makes sense for exhaustive searchers (Kraft and Lee, 1979). As an example, when tasked to find as many documents as possible related to different species of animals that are endangered, becoming disgusted with the presented results when a lack of unseen animal species are shown would be a suitable point at which to break and reformulate a new query, or abandon the search session altogether.

From the heuristics defined by Cooper (1973b) and Kraft and Lee (1979), we propose two variants of the frustration and disgust heuristics, SS2-NT and SS3-NC.

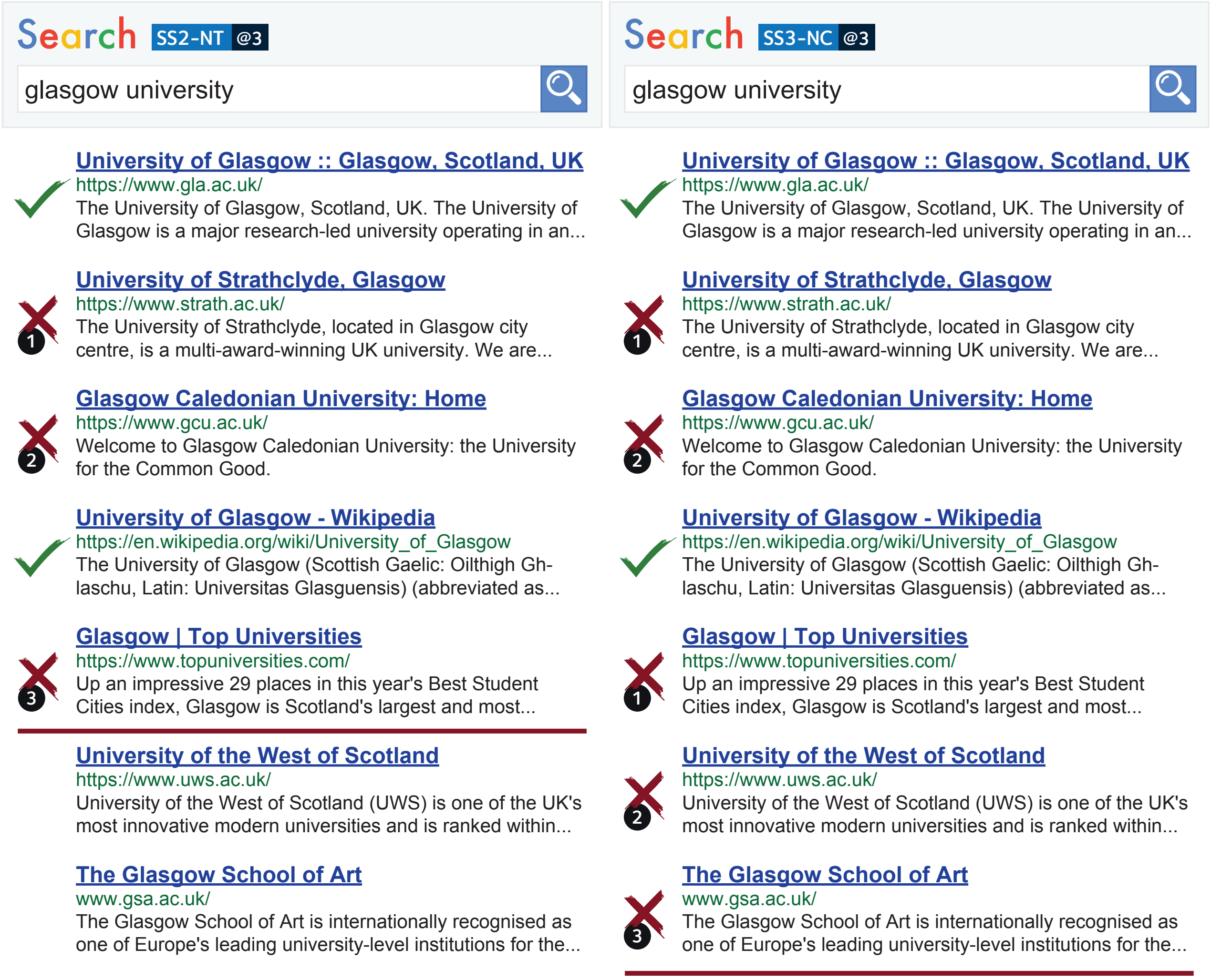
- SS2-NT Non-relevant, Total Under this stopping strategy, the searcher will stop once they have observed x_2 unattractive result summaries.
- SS3-NC Non-relevant, Contiguous Similar to the stopping strategy defined above, a searcher employing this stopping strategy will stop once they have observed x_3 unattractive result summaries *in a row (contiguously)*.

With these stopping strategies adaptable to the presented results, this inherently makes the strategies more realistic (Moffat et al., 2013). Figure 5.2 illustrates both strategies in action across the same query and associated results. On the left of the figure is an illustration of when a searcher employing SS2-NT would stop, and on the right, an example of SS3-NC. We use SS2-NT @3 and SS3-NC @3. Under SS2-NT, a searcher would stop at rank 5, while a searcher would stop at rank 7 when employing SS3-NC.

5.2.2 Goal/Satisfaction-Based

Analogous to frustration and disgust are the satisfaction, satiation and number-based stopping heuristics (Cooper, 1973b; Simon, 1955; Gibb, 1958). Rather than focus upon the frustration or disgust that a searcher might experience when confronted with unattractive result summaries, satisfaction-based stopping heuristics – explained in Section 3.2.1.1 – consider a searcher encountering a number of *attractive* result summaries before becoming sufficiently satisfied with what they have found before stopping.

- SS4-SAT Satiation A searcher using this stopping strategy will stop examining content after encountering x_4 attractive result summaries.



glasgow university

✓ **University of Glasgow :: Glasgow, Scotland, UK**
https://www.gla.ac.uk/
The University of Glasgow, Scotland, UK. The University of Glasgow is a major research-led university operating in an...

✗ 1 **University of Strathclyde, Glasgow**
https://www.strath.ac.uk/
The University of Strathclyde, located in Glasgow city centre, is a multi-award-winning UK university. We are...

✗ 2 **Glasgow Caledonian University: Home**
https://www.gcu.ac.uk/
Welcome to Glasgow Caledonian University: the University for the Common Good.

✓ **University of Glasgow - Wikipedia**
https://en.wikipedia.org/wiki/University_of_Glasgow
The University of Glasgow (Scottish Gaelic: Oilthigh Ghlaschu, Latin: Universitas Glasguensis) (abbreviated as...

✗ 3 **Glasgow | Top Universities**
https://www.topuniversities.com/
Up an impressive 29 places in this year's Best Student Cities index, Glasgow is Scotland's largest and most...

University of the West of Scotland
https://www.uws.ac.uk/
University of the West of Scotland (UWS) is one of the UK's most innovative modern universities and is ranked within...

The Glasgow School of Art
www.gsa.ac.uk/
The Glasgow School of Art is internationally recognised as one of Europe's leading university-level institutions for the...

glasgow university

✓ **University of Glasgow :: Glasgow, Scotland, UK**
https://www.gla.ac.uk/
The University of Glasgow, Scotland, UK. The University of Glasgow is a major research-led university operating in an...

✗ 1 **University of Strathclyde, Glasgow**
https://www.strath.ac.uk/
The University of Strathclyde, located in Glasgow city centre, is a multi-award-winning UK university. We are...

✗ 2 **Glasgow Caledonian University: Home**
https://www.gcu.ac.uk/
Welcome to Glasgow Caledonian University: the University for the Common Good.

✓ **University of Glasgow - Wikipedia**
https://en.wikipedia.org/wiki/University_of_Glasgow
The University of Glasgow (Scottish Gaelic: Oilthigh Ghlaschu, Latin: Universitas Glasguensis) (abbreviated as...

✗ 1 **Glasgow | Top Universities**
https://www.topuniversities.com/
Up an impressive 29 places in this year's Best Student Cities index, Glasgow is Scotland's largest and most...

✗ 2 **University of the West of Scotland**
https://www.uws.ac.uk/
University of the West of Scotland (UWS) is one of the UK's most innovative modern universities and is ranked within...

✗ 3 **The Glasgow School of Art**
www.gsa.ac.uk/
The Glasgow School of Art is internationally recognised as one of Europe's leading university-level institutions for the...

Figure 5.2 An example of the two frustration rules, SS2-NT (left) and SS3-NC (right), both three unhelpful (non-relevant) result summaries, under the same query and results. Given that SS2-NT considers the total number of result summaries judged to be unhelpful, a searcher employing this stopping strategy would stop at rank *5* in the example above. Considering a set of contiguous unhelpful summaries, a searcher using SS3-NC would stop at rank seven.

While we consider this stopping strategy in the context of result summary level stopping, such a stopping strategy may not be particularly useful when operationalised at this stopping decision point. Consider the scenario where a searcher issues a poor query, yielding next to no summaries deemed to be worthy of further examination. In this scenario, a searcher fully complying with SS4-SAT may struggle to find enough documents to reach their goal. This will mean that the searcher wastes time examining poor results. Such a stopping strategy may be better suited to an overall search goal – or at the session level stopping decision point. As a means of potentially avoiding a searcher becoming *'trapped'* in an examination of a fruitless set of results, time limits could be imposed. We also consider an additional stopping strategy to alleviate this issue, as discussed below.

5.2.3 Combining Frustration and Satisfaction

The next stopping strategy proposed considered a combination of both the frustration/disgust and satisfaction/satiation stopping heuristics. This was named the *combination heuristic* by Kraft and Lee (1979). Employing this stopping strategy, a searcher would stop either when they became frustrated or were satisfied with the number of attractive summaries that they had observed – whichever of the two were met first. As such, we can convert this into a fifth stopping strategy, defined below.

- SS5-COMB Combination — Frustration/Satiation A searcher using this stopping strategy will employ both frustration (disgust) and satisfaction (satiation) stopping heuristics to determine when to stop, ceasing their examination of the SERPs contents for the first stopping heuristic whose criterion is met.

While SS4-SAT can be selected as the operationalised satisfaction/satiation component, one of either SS2-NT or SS3-NC can be selected for the frustration/disgust component of this fifth stopping strategy. We discuss this in our general methodology in Section 6.4.2.6 on page 173. Note that like SS2-NT and SS3-NC, we include items issued from previous queries of the same search session.

5.3 Difference Threshold

The next set of stopping strategies are based upon the difference threshold heuristic, as outlined in Section 3.2.1.2 on page 83. To operationalise this stopping heuristic, we considered the difference between a given result summary's snippet text and the snippet texts of previously examined result summaries. Here, the idea was that as a searcher examined result summaries on a SERP, summaries may be encountered that are not *sufficiently different* from what had already been observed.[3] When encountering a result summary that is not sufficiently different, a searcher subscribing to the difference threshold stopping heuristic will then decide to stop examining results.

From this stopping heuristic, we devised two separate stopping strategies where the difference between snippet texts was computed in different ways. The first approach considered the *term overlap difference.*

- **SS6-DT** **Difference, Terms** This stopping strategy compares occurrences of terms in a given result summary's snippet text against all terms in previously examined result summary snippets. If $\frac{|s_{curr} \cup s_{prev}|}{|s_{curr}|} > x_6$, the new snippet text is then considered as too similar to previously examined result summaries. The searcher then stops examining result summaries on the present SERP.

Essentially, **SS6-DT** considers that if more terms overlap between old and new, the greater the chance that the new result summary would not contain any new information. In the definition above, s_{curr} denotes the terms of the currently examined result summary snippet, s_{prev} denotes terms from all previously observed result summary snippets[4], and x_6 is the threshold at which the searcher will stop.

The second difference-based stopping strategy utilised *Kullback–Leibler Divergence* (Kullback and Leibler, 1951) to determine how different a given result summary is from result summaries that have been previously examined.

[3] This means that searchers wouldn't be learning anything new (Nickles, 1995), and thus, under the eyes of such a strategy, would be wasting their time.

[4] All previously result summaries could be either session-based or query-based. This is an implementation decision, which we discuss in Section 6.4.2.6 on page 173.

- **SS7-DKL** **Difference, KL-Divergence** Here, KL-divergence is used as a means for comparing a given result summary (represented as a *bag of words*) against those previously observed. If the resulting value is less than threshold x_7, the present result summary is considered to be too similar, and the searcher stops. The searcher then abandons the present SERP.

Details related to the implementation of the difference heuristic stopping strategies, as well as the parameter threshold settings used, can be found in Section 6.4.2.6.

5.4 Instantaneous Intake

In Section 3.3.1.2, we discussed several stopping heuristics that were derived from OFT and IFT. The IFT-based heuristic considers a searcher's *optimal stopping point* at which a forager[5] should stop, as suggested by the underlying models of IFT. This is calculated by observing a searcher's *average rate of gain.* If the value of knowledge gained drops below this threshold, the searcher should stop, as graphically illustrated in Figure 3.8 on page 94.

We now propose an eighth stopping strategy, this time based upon the notion of the average rate of gain accrued by a searcher.

- **SS8-IFT** **Optimal Stopping** With this stopping strategy, a searcher is assumed to have some idea of the average rate of gain (denoted as x_8). If the rate of gain from the observed documents thus far does not exceed x_8, the searcher then stops and proceeds to undertake the next action as dictated by the CSM.

Computing the average rate of gain is a non-trivial problem. We leave specific implementation details of how this was achieved – along with other implementation details of the stopping strategy – to our methodology, reported in Section 6.4.2.6 on 173.

[5] As we discussed in Section 3.3.1, a *forager* can be considered analogous to a searcher seeking information.

5.5 Time-Based

In addition to the optimal stopping point approach discussed above, Section 3.3.1.2 also outlined a number of different OFT-inspired stopping heuristics that primarily used time as a measure of determining when to stop. From these approaches, we create two further time-based stopping strategies.

- **SS9-TIME** **Time-based** Based upon the *time heuristic* (Charles-Dominique and Martin, 1972; Krebs, 1973), a simulated searcher using this stopping strategy will abandon a SERP after x_9 seconds have elapsed since they entered it.
- **SS10-RELTIME** **Time, Give-Up** Using the *give-up heuristic* as defined by Krebs et al. (1974), a searcher will abandon a presented SERP x_{10} seconds after the last document that was found and considered relevant/useful (saved) to the given information need.

Given these stopping strategy definitions, **SS9-TIME** performs akin to **SS1-FIX**, in the sense it offers a fixed interaction time on each SERP, and is agnostic of the quality of the presented ranked list. Conversely, **SS10-RELTIME** offers a more adaptive solution similar to **SS2-NT** and **SS3-NC**, basing the time at which the searcher stops x_{10} seconds after a relevant document was last saved.

For this thesis, we also consider the *combination heuristic* proposed by McNair (1982). The stopping strategy that we propose based upon this heuristic assumes that a searcher has been able to acquire an idea of how potentially relevant summaries are *distributed* across the results presented within the SERP.

- **SS11-COMB** **Combination — Time and satiation** Encountering a SERP expected to yield a high volume of relevant content early on (high scent), a searcher will employ the satisfaction/satiation stopping heuristic. However, if the SERP is judged to yield

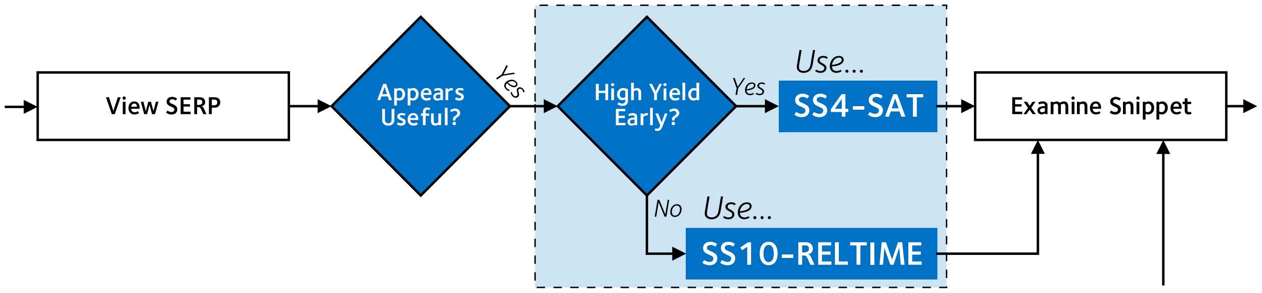


Figure 5.3 An excerpt of the CSM with the additional decision point that SS11-COMB incorporates within the searcher model. After deciding that individual result summaries within a SERP are worth examining in more detail, a searcher will then also have to decide whether the presented SERP will yield a high number of fruitful results *early* in the rankings, or trickle relevant material over greater depths (or not at all). The additional decision point and selected stopping strategies are highlighted within a blue box.

relevant items over greater depths or is judged to be of poor quality (low scent), the give-up time-based heuristic is used instead.

From our definitions above, SS4-SAT is used for the satisfaction/satiation component, and SS10-RELTIME is used for the give-up time heuristic component. The combination stopping strategy attempts to ensure that a searcher does not waste time on a SERP that appears to offer a low yield, but conversely capitalises upon patches that present a high yield. Of course, determining the perceived yield is a question of implementation; refer to Section 6.4.2.6 for more information on how we implemented this particular stopping strategy. Essentially, this combination stopping strategy incorporates an additional decision point within the searcher model, where one must determine if the presented SERP is high yield early on or not. This is illustrated as an excerpt of a flowchart in Figure 5.3.

5.6 Measure-Based

The final proposed stopping strategy is based upon an established IR measure. *Rank Biased Precision (RBP)* – as discussed in Section 2.4.1.5 – is utilised as the basis of our final stopping strategy. Under RBP, the decision to continue to the next result in a ranked list is based upon a patience parameter or the probability of continuing. Essentially, RBP states that the probability of continuing decreases as a searcher progresses further down a ranked list.

- **SS12-RBP** **Rank-Biased Precision** With this stopping strategy, a searcher will stop examining a SERP when the likelihood of continuing falls below the RBP probability computed at that rank, given a patience parameter x_{12}.

By including such a measure, we provide a platform for which contemporary IR measures can be compared against the performance of other stopping heuristics defined in the literature. Implementation details, such as how we implemented the probabilistic component, can be found in Section 6.4.2.6.

5.7 Chapter Summary

This chapter has outlined 12 different stopping strategies, all of which are based upon prior stopping heuristics and an established IR measure. As such, this chapter provides a possible answer to **HL-RQ2**. In subsequent chapters of this thesis, we take these stopping strategies forward, discuss the specifics of how they were implemented in Section 6.4.2.6 (page 173), and how they were employed in our empirical experimentation.

Chapter 6

General Methodology

In this chapter, we provide an overview of the *general methodology* we used in this thesis. Part III reports on a number of empirical contributions, where we explore how stopping behaviours vary under different search contexts. This chapter is broken down into six main components that we summarise below.

- **Context, Data, Tasks and Retrieval System** This involves the search context, document corpus, topics, tasks and retrieval system used throughout this thesis.
- **User Studies** We discuss the general methodology behind two user studies designed to examine how different factors affect stopping behaviours.
- **Interaction and Performance Data** We discuss how we extracted key measures from the interaction data obtained from the two user studies.
- **Simulations of Interaction** Making use of the aforementioned interaction data, we outline the methodology of an extensive series of *simulations of interaction* designed to replicate the user studies.

- **Examining Performance** We evaluate the performance of simulated searchers, allowing us to examine what stopping strategies offer the best levels of performance under different search contexts.
- **Comparing Searchers** Finally, we outline the approach for comparing the performance of real-world searchers against simulated counterparts, meaning we can determine what configurations offers the closest approximations to real-world behaviours.

The remainder of this chapter is devoted to a discussion of each of these components. We first begin with a discussion of the basic setup of our experiments, considering the retrieval system and document collection used. We also consider the basic instructions that we issued to subjects of our user studies, such as the simulated search task that we employed.

6.1 Context, Data, Tasks and Retrieval System

The context for all experiments reported in this thesis is **news search.** We employ a widely used corpus of news articles and associated topics, with queries issued against the retrieval system described in Section 6.1.2.

Given the context of news search, we employ a *simulated work context* with which subjects – both real-world and simulated – conform to. As outlined by Borlund and Schneider (2010) and Li and Hu (2013), simulated work contexts are designed as close as possible to the situations facing real searchers, and thus provide the context that elicits a searcher's interactions with a retrieval system. Subjects who participated in our user studies were instructed to imagine that they were newspaper reporters. They were required to identify articles to write stories about topics provided to them (refer to Section 6.1.3). Depending upon the search task given to the subjects, subjects would then save articles that they believed were relevant to a given topic – or were relevant, and discussed a new aspect of the said topic.[1]

[1] Refer to Section 6.2 for further details on the different goals that we imposed upon searchers.

6.1.1 Document Corpus

Under the context of news search, we employed a corpus of newspaper articles. The *TREC AQUAINT* corpus was selected for all experimentation work in this thesis. The corpus consists of a total of $1,033,461$ news articles (referred to as *documents)* from the period ranging 1996-2000. All of the documents were collected from three *newswires,* namely: the *Associated Press (AP);* the *New York Times (NYT);* and *Xinhua (XIE).* The AQUAINT corpus was used as it has been extensively used in prior research. Studies include for example: Collins-Thompson et al. (2004); Ofoghi et al. (2006); Baillie et al. (2006); Azzopardi and Vinay (2008); Kelly et al. (2009); Azzopardi et al. (2013); Maxwell and Azzopardi (2014); Harvey and Pointon (2017); Yang and Fang (2017); and Wilkie and Azzopardi (2017). Basic corpus statistics can be found in the illustration below.

Number of Documents in Corpus **1,033,461**

Unique Terms **707,778**

Mean Document Length (terms) **275**

Total Number of Terms in Corpus **284,597,335**

6.1.2 Retrieval System

The AQUAINT corpus was then indexed using the *Whoosh IR Toolkit.*[2] We applied Porter stemming and removed stopwords as per the 421-term classical stopword list by Fox (1992).[3] During the indexing process, we also removed documents with duplicate titles. With documents originating from newswires, we found many occurrences of documents with the same title. Documents discussing ongoing events were continually revised as new information about the event arose. For documents with duplicate titles, we retained the document

[2] *Whoosh* can be freely acquired using the `pip` *Python* package manager (via *PyPi*) – documentation for Whoosh can be accessed at `http://whoosh.readthedocs.io/en/latest/intro.html`. **LA** *2018-05-18* The corpus was indexed with Whoosh `2.7.4`.

[3] More information on the indexing process can be found in Section 2.2.1.

with the latest timestamp. This document represented the final version of the published article, containing the most recent or up-to-date information.

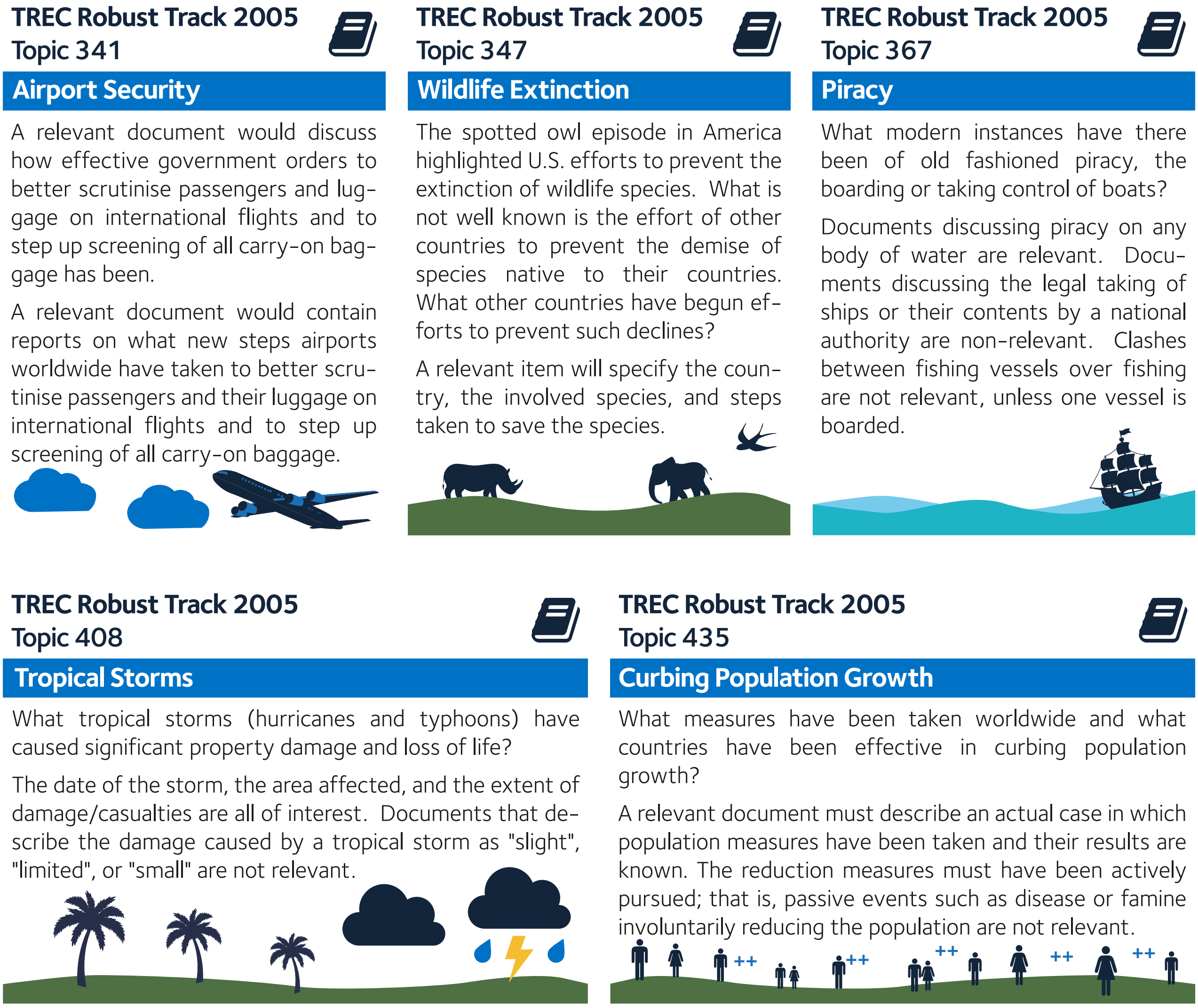
From the indexing process, a 3.8GB index was produced. The index contained a total of 959,678 indexed documents. The Whoosh IR toolkit was employed to issue queries against the index. All ranked results for queries were computed with BM25, where $\beta = 0.75$ and $k_1 = 1.2$ (refer to Section 2.2.2.3 on page 38). Terms in all issued queries were ANDed together to restrict the set of retrieved documents to those that contained all of the query terms. This decision was also taken as many retrieval systems implicitly AND terms together.

6.1.3 Topics

Five topics were also selected from the 50 provided in the *TREC 2005 Robust Track* (as used with the AQUAINT collection) as outlined by Voorhees (2006). These topics were used throughout experimental work reported in this thesis and were selected based on evidence from a previous user study (of similar nature) conducted by Kelly et al. (2009). Evidence showed that the topics offered similar levels of difficulty. The five topics, along with a short description of what constitutes a relevant document, are listed below. These summaries are derived from the TREC topic descriptions that are provided as part of the TREC 2005 Robust Track. Figure 6.1 provides an illustration of the five topics, along with their descriptions. In addition, Table 6.1 provides basic summary statistics of the number of non-relevant and relevant[4] documents that were identified by the TREC assessors. The remaining 45 topic descriptions are not used.

- **Topic 341** **Airport Security** For this topic, relevant documents discuss additional security measures that were taken by international airports around the world. Relevance is only denoted when a document discusses measures that go beyond basic passenger and carry-on luggage screening. For example, AQUAINT document

[4] The TREC 2005 Robust Track uses graded judgements for relevant documents; these are identified as somewhat (1) and definitely (2) relevant.



TREC Robust Track 2005
Topic 341

Airport Security

A relevant document would discuss how effective government orders to better scrutinise passengers and luggage on international flights and to step up screening of all carry-on baggage has been.

A relevant document would contain reports on what new steps airports worldwide have taken to better scrutinise passengers and their luggage on international flights and to step up screening of all carry-on baggage.

TREC Robust Track 2005
Topic 347

Wildlife Extinction

The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines?

A relevant item will specify the country, the involved species, and steps taken to save the species.

TREC Robust Track 2005
Topic 367

Piracy

What modern instances have there been of old fashioned piracy, the boarding or taking control of boats?

Documents discussing piracy on any body of water are relevant. Documents discussing the legal taking of ships or their contents by a national authority are non-relevant. Clashes between fishing vessels over fishing are not relevant, unless one vessel is boarded.

TREC Robust Track 2005
Topic 408

Tropical Storms

What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?

The date of the storm, the area affected, and the extent of damage/casualties are all of interest. Documents that describe the damage caused by a tropical storm as "slight", "limited", or "small" are not relevant.

TREC Robust Track 2005
Topic 435

Curbing Population Growth

What measures have been taken worldwide and what countries have been effective in curbing population growth?

A relevant document must describe an actual case in which population measures have been taken and their results are known. The reduction measures must have been actively pursued; that is, passive events such as disease or famine involuntarily reducing the population are not relevant.

Figure 6.1 The titles and descriptions of the five TREC topics used in experimental work. Topics are extracted from the *TREC 2005 Robust Track*, as outlined by Voorhees (2006). Descriptions provide an explanation as to what constitutes a relevant (and often non-relevant) document.

`NYT19980616.0123` discusses *San Francisco International Airport's* attempts at introducing a *robot sniffer*, attempting to look for nitroglycerine in luggage.

- **Topic 347** **Wildlife Extinction** This topic concerns wildlife extinction, and what efforts have been taken by countries *other than the U.S.* to counter the decline in endangered wildlife. Relevant documents would explicitly mention the country, the species concerned, and the efforts the state or other governmental agency took to prevent a decline in numbers. For example, document `XIE20000531.0205` discusses the breeding programme undertaken by China to bolster the number of Siberian Tigers.

Table 6.1 Basic statistics for the five TREC topics selected, including the number of documents judged to be relevant (with graded judgements) and explicitly non-relevant by the TREC assessors.

		Non-Rel.	TREC Relevant		
	Total	**Total**	**Somewhat**	**Definitely**	**Total**
Topic 341	617	580	33	4	37
Topic 347	665	500	22	143	165
Topic 367	621	526	11	84	95
Topic 408	685	502	57	126	183
Topic 435	707	555	98	54	152

- **Topic 367** **Piracy** Instances of modern piracy are considered relevant to this topic – not in the sense of software piracy, but the act of a water going vessel being boarded by individuals wishing to hijack it. Document `APW19980601.1065` provides an example of this – the *Petro Ranger*, a large fuel tanker, was boarded by pirates in 1998 in the South China Sea. To be relevant to the topic, the name of the vessel and the body of water it was hijacked on must be mentioned – those discussing instances of when states intercepted vessels are not relevant.

- **Topic 408** **Tropical Storms** Documents discussing major tropical storms are relevant, where the storm is reported to have caused significant damage and a large number of casualties. This is a particularly timely topic for the document corpus considered, as the 1998 hurricane season in the Caribbean has been reported to be one of the most costly in history.[5] For example, document `APW19980921.1265` discusses the effects on Puerto Rico of Hurricane Georges in September 1998, leaving three dead, many houses damaged, and thousands homeless.

[5] This is reported by the US *National Oceanic and Atmospheric Administration (NOAA)*, as seen at `http://www.outlook.noaa.gov/98hurricanes/`. **LA** *2018-05-18*

- **Topic 435** **Curbing Population Growth** The final topic considers efforts that have been made by countries around the world to stem the ever-increasing human population. Documents discussing this issue are only relevant to the topic if the results to a case have been made public, and a reduction in a country's population has been actively pursued. The document must mention the country and the measures that the state or governmental agency pursuant to bring about a fall in population. As such, events like famines that resulted in a fall in the population are not relevant. A perhaps well-known example of a country pursuing a reduction in its population is the Chinese one-child policy, enacted in the late 20th century. Document `NYT19981031.0070` discusses the Chinese government's efforts to curb its expanding population at the time, with sexual education and heavy financial penalties for additional children.

For all user studies reported in this thesis, we selected topic **367** as a **practice topic**, permitting participating subjects to familiarise themselves with the experimental system used. We therefore do not report any results from interactions that took place with this topic when reporting the user studies. In the next section, we outline the different search tasks that were undertaken by subjects of the user studies.

6.2 User Study Methodology

Using the aforementioned corpus, retrieval system and topics, we now move onto a discussion of the common methodology employed across the two user studies. These are detailed in Chapters 7 and 8. While intricate details of each study's methodology do indeed vary (such as a summary of the subjects that were employed), these are nevertheless common components. These are discussed in this section. As a reminder, the two studies examine:

- the length (and thus quality) of snippets presented in result summaries are varied (Chapter 7, conducted between July and August 2016); and

- the overall search goal (time constraints vs. relevancy accruement) and task goal (ad-hoc vs. aspectual) are changed (Chapter 8, conducted in January 2018).

Specifically, the methodology used for these studies allowed us to determine how the stopping behaviour of a searcher varies when these conditions and interfaces were varied. We discuss the specific interfaces and conditions that we trialled in subsequent chapters of this thesis.

Both user studies were undertaken using a custom built experimental framework called **TREConomics**.[6] The pure-Python framework has been developed over a number of years. It permits for straightforward deployment of IIR-based studies that have examined a variety of different aspects. It has been successfully used in a number of prior works, including those by: Azzopardi et al. (2013); Maxwell and Azzopardi (2014); Kelly and Azzopardi (2015); Edwards et al. (2015); and Crescenzi et al. (2016).

6.2.1 Experimental Details and Flow

Each user study was designed to last for 45 to 50 minutes, which included the completion of requested search tasks and surveys. Both experiments followed a similar structure, where subjects would complete a number of surveys before beginning a search task, and were then asked to complete a further survey after they had finished their task. These surveys, as discussed in Section 6.3.4, permitted us to gather a series of usability measures (refer to Section 2.4.2) about the perceived experiences that subjects had when interacting with the various interfaces and conditions that were trialled.

The basic structure of both user studies was as follows.

❶ Subjects began by reading the experiment briefing sheet. If they approved of the experimental outline, they then agreeing to continue.

[6] **TREConomics** can be accessed at `https://github.com/leifos/treconomics`. **LA** *2018-05-15*

❷ A demographics survey was then completed.

❸ Subjects then attempted the *practice task*, using the practice topic. This allowed subjects to familiarise themselves with the system and its interface.

❹ Subjects would then complete the various search tasks set out for them. Each task consisted of three steps:

- a pre-task survey, capturing a subject's prior knowledge about the topic;
- the search task itself; and
- a post-task survey, capturing the subject's experiences regarding searching for information about the given topic.

❺ Upon completion of the required search tasks, subjects would then respond to a post-experiment survey. In this survey, subjects were asked general questions about their experience across all the different tasks.

❻ Finally, upon completion, subjects would be presented with a results screen, providing a summary of their performance. Performance for each subject was presented on a per-task basis. When the subject proceeded to the next step, they were presented with a message informing them the experiment has concluded.

Subjects undertook a total of four search tasks. For each of these tasks, different interactions and experiences were captured by the **TREConomics** framework. Including the practice task at the beginning of each experiment, this took the total number of search tasks per subject up to five. Following a **within-subjects study design**, the four search tasks – each using a different topic as described in Section 6.1.3 – permitted us to trial one of the four experimental conditions/interfaces in each study. The topics and interfaces/conditions were assigned to subjects using a Latin-square rotation to minimise ordering effects. A within-subjects design increases the statistical power – the number of 'subjects' is higher than a between-subjects design. Limitations of such a design include issues such as fatigue. Attempts were made to limit this by being mindful of the time required.

6.2.2 Experimental Search Interface

In this section, we discuss the experimental search interface that was used by subjects of the user studies.[7] The interface would be familiar to anyone who had used a web-based retrieval system, meaning that the learning curve for using the interface would most likely be low. Upon commencement of the experiment, the interface would launch in a fixed-size popup window (refer to Section 6.2.4.3) of the web browser being used.

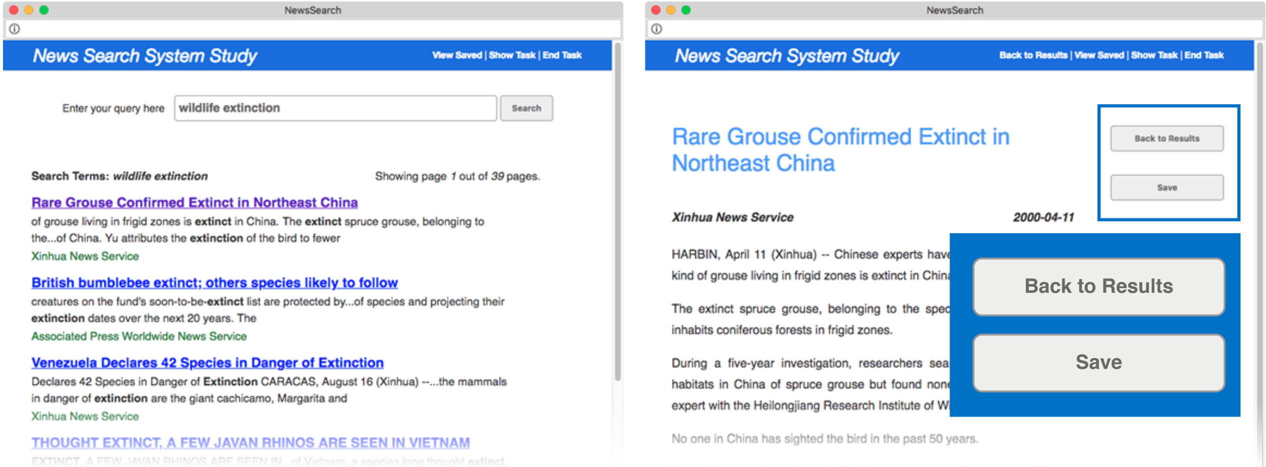
The interface consists of three main views, the two most important being shown in Figure 6.2. The views were:

- the *Search Engine Results Page (SERP)*, presenting the query box and results for an issued query;
- the *document view,* providing the full text of a document; and
- the *saved documents list,* providing a list of the documents that each subject had saved during the search session.

In addition to the three views above, we also provided a *topic view,* which, when requested, would open a further popup window that contained a description of the topic. This was to serve as a reminder. Subjects were provided the topic description in full before the search task began.

Common to all views was the inclusion of the blue navigation bar at the top of the popup window. As we discuss further in Section 6.2.4.3, this bar was included to provide a series of different navigational links. Such an example would be on the document view page, which contained a link to return to the originating SERP. Where applicable, we also provided a link for the subject to end the search task if he or she felt that they were satisfied.

[7] Slight modifications to the search interface were made to the goal-based study, as we discuss in Section 8.2.1 on page 250.



Enter your query here wildlife extinction Search

Search Terms: ***wildlife extinction*** Showing page *1* out of *39* pages.

Rare Grouse Confirmed Extinct in Northeast China
of grouse living in frigid zones is **extinct** in China. The **extinct** spruce grouse, belonging to the...of China. Yu attributes the **extinction** of the bird to fewer
Xinhua News Service

British bumblebee extinct; others species likely to follow
creatures on the fund's soon-to-be-**extinct** list are protected by...of species and projecting their **extinction** dates over the next 20 years. The
Associated Press Worldwide News Service

Venezuela Declares 42 Species in Danger of Extinction
Declares 42 Species in Danger of **Extinction** CARACAS, August 16 (Xinhua) --...the mammals in danger of **extinction** are the giant cachicamo, Margarita and
Xinhua News Service

THOUGHT EXTINCT, A FEW JAVAN RHINOS ARE SEEN IN VIETNAM
EXTINCT, A FEW JAVAN RHINOS ARE SEEN IN...of Vietnam, a species long thought extinct.

Rare Grouse Confirmed Extinct in Northeast China

Xinhua News Service ***2000-04-11***

HARBIN, April 11 (Xinhua) -- Chinese experts have
kind of grouse living in frigid zones is extinct in China

The extinct spruce grouse, belonging to the spec
inhabits coniferous forests in frigid zones.

During a five-year investigation, researchers sea
habitats in China of spruce grouse but found none
expert with the Heilongjiang Research Institute of W

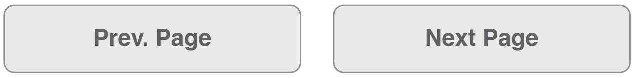
No one in China has sighted the bird in the past 50 years.

Figure 6.2 Example screenshots of the basic search interface used as part of **TREConomics**. On the left is a screenshot of typical experimental SERP for the query `wildlife extinction`. The right shows the document view, showing the option for subjects to **Save** a document that they consider relevant to the given topic. Buttons on the document interface are zoomed.

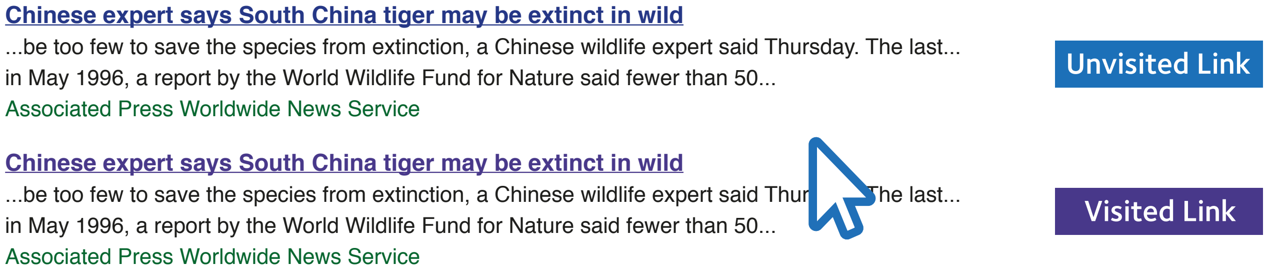
6.2.2.1 The SERP

As can be observed from the left screenshot in Figure 6.2, the SERP does not look all that different from a SERP on a contemporary web retrieval system – sans right rail components, the lack of which we discussed previously in Section 2.3.2.1. The experimental SERP provides the query box at the top, allowing subjects to enter their query term(s), and a button to submit their query, named Search. The ↩ key could also be used to submit a query, as is standard in contemporary retrieval system interfaces.

Once submitted, results were displayed underneath the query box. The issued query was provided, along with an approximation of how many pages of results were provided to the searcher from a given query. This hints that pagination is utilised – with 10 results per page shown. At the bottom of each SERP were links that would allow the searcher to move to the previous and next page of results.



Result summaries were shown as discussed in Section 2.3.2.1. The title, the source, and any snippet text were all provided. Given that the experiments were based on news search, the source is the name of the newswire from which the document originates.



Chinese expert says South China tiger may be extinct in wild
...be too few to save the species from extinction, a Chinese wildlife expert said Thursday. The last...
in May 1996, a report by the World Wildlife Fund for Nature said fewer than 50...
Associated Press Worldwide News Service

Chinese expert says South China tiger may be extinct in wild
...be too few to save the species from extinction, a Chinese wildlife expert said Thur The last...
in May 1996, a report by the World Wildlife Fund for Nature said fewer than 50...
Associated Press Worldwide News Service

When a subject clicked on the link, he or she would then be taken to the document view (discussed below), displaying the associated document in its entirety. Standard hyperlink colours were employed – blue for unvisited, and purple for visited.

6.2.2.2 The Document View

The right screenshot in Figure 6.2 illustrates the document view. The view provides the title, the document source (newswire), the date at which the document was created, and the full text of the document. On the right rail of the page, subjects were provided with two buttons – one to return them to the originating SERP, or another to save the document. The act of saving a document is a crucial component to both studies we discuss in this thesis. It provided us with a mechanism to determine what documents that subjects thought were relevant to the associated topics. From there, we could also use this series of saved documents to calculate a subject's performance.

6.2.2.3 The Saved Documents View

The third key view allowed subjects to view a list of documents that they had previously saved as relevant to the given topic. This list of documents also provided buttons allowing subjects to change their decisions as to what constituted a relevant document.

```
2016-07-07 19:03:02,512 INFO <USERID> 0 4 4 408 SEARCH_TASK_COMMENCED
2016-07-07 19:03:02,545 INFO <USERID> 0 4 4 408 VIEW_SEARCH_BOX 1
2016-07-07 19:03:03,131 INFO <USERID> 0 4 4 408 QUERY_FOCUS
2016-07-07 19:03:07,106 INFO <USERID> 0 4 4 408 QUERY_ISSUED 'tropical storm damage'
2016-07-07 19:03:07,115 INFO <USERID> 0 4 4 408 QUERY_START 'tropical storm damage'
2016-07-07 19:03:07,318 INFO <USERID> 0 4 4 408 QUERY_END 'tropical storm damage'
2016-07-07 19:03:07,326 INFO <USERID> 0 4 4 408 QUERY_COMPLETE 'tropical storm damage'
2016-07-07 19:03:07,334 INFO <USERID> 0 4 4 408 SEARCH_RESULTS_PAGE_QUALITY 1  0 10 6
2016-07-07 19:03:07,342 INFO <USERID> 0 4 4 408 VIEW_SEARCH_RESULTS_PAGE 1
2016-07-07 19:03:10,355 INFO <USERID> 0 4 4 408 DOCUMENT_HOVER_IN 7309 XIE19960728.0162 129 -2 2
2016-07-07 19:03:11,333 INFO <USERID> 0 4 4 408 DOC_CLICKED 7309  0 -2 -2
2016-07-07 19:03:11,353 INFO <USERID> 0 4 4 408 DOC_MARKED_VIEWED 7309 XIE19960728.0162 129 -2 2
2016-07-07 19:03:18,020 INFO <USERID> 0 4 4 408 DOC_MARKED_RELEVANT 7309 XIE19960728.0162 129 1 2
```

Figure 6.3 An excerpt from the interaction log of the user study presented in Chapter 7. A sequence of interactions are shown that were logged by the **TREConomics** framework.

6.2.3 Capturing Interactions and Survey Responses

In addition to the interface, the **TREConomics** framework provided extensive logging capabilities to capture a variety of different events triggered by subjects as they performed search tasks (with survey responses saved separately to a RDBMS). This resulted in the generation of an experiment *log file*, capturing the date, time, searcher and topic for each event that was logged. Figure 6.3 provides an anonymised excerpt from the interaction log of the user study presented in Chapter 7.

The figure illustrates the different actions that were logged from when a searcher begins interactions with the query box (QUERY_FOCUS), to issuing a query (QUERY_ISSUED, complete with the terms of the query), to clicking a document (DOC_CLICKED), and, finally, to saving the document (or considering it relevant to the given topic, DOC_MARKED_RELEVANT). A detailed discussion of the different behavioural measures that we examined from the interaction log is detailed in Section 6.3.

6.2.4 Crowdsourcing Considerations

An important factor in planning any user study is the economics of collecting input from subjects. *Where do the subjects come from? How do we recruit them?* A traditional, lab-based study as discussed in Section 2.3 typically involves a significant investment in time and monetary cost from the researchers conducting the experiment (Spool and Schroeder, 2001). For both user studies previously detailed, we employed a crowdsourced approach to our experimentation. Crowdsourcing is the practice of obtaining input into a task by enlisting the services of a number of people, recruited over the Internet.

As highlighted by Zuccon et al. (2013), crowdsourcing provides an alternative means for capturing user interactions and search behaviours. Greater volumes of data can be obtained from more heterogeneous workers at a lower cost – all within a shorter timeframe. Of course, pitfalls of a crowdsourced approach include the possibility of workers completing tasks as efficiently (but not effectively) as possible, or submitting their tasks without performing the requested operations (Feild et al., 2010).

Despite these issues, it has been shown that there is little difference in the quality between crowdsourced and lab-based studies (Kelly and Gyllstrom, 2011; Zuccon et al., 2013). Nevertheless, quality control is a major component of a well-executed crowdsourced experiment, with examples in a similar research area including work by Kazai et al. (2011) and Crescenzi et al. (2013).

Using crowdsourcing for the two user studies, we detail in the remainder of this section the precautions that were taken, discussing both the requirements for the subjects themselves, and their device's setup. We also provide a discussion of the crowdsourcing platform used.

6.2.4.1 Platform Details

Both studies were run over the *Amazon Mechanical Turk (MTurk)* platform. Workers[8] from the platform each performed a single task (or, to use MTurk terminology, a *Human Intelligence Task (HIT)*, with a single HIT corresponding to the entire experiment. This is in con-

[8] In this section, a *worker* refers to an individual undertaking the experiment on the MTurk platform. This term is considered interchangeable with a *subject*.

trast to many other crowdsourced studies, where workers would typically undertake small (typically decision-based) HIT transactions. This decision was taken so that the experiment would closely resemble a laboratory-based experiment.

6.2.4.2 Subject Requirements

Due to the expected length that workers would take to complete the two studies[9], workers who completed either study in full were reimbursed for their time with US$9 – greater than the hourly US$7.25 minimum wage set by the U.S. federal government.[10] Workers interested in undertaking either of the two studies were required to meet a minimum set of criteria to be eligible to participate. We required that workers were:

- from the U.S.;
- native English speakers;
- possessed a HIT acceptance rate of at least 95% (from prior experiments); and
- had at least 1000 prior HITs approved.

Requiring a high HIT acceptance rate reduced the likelihood of recruiting workers who would not complete the study in a satisfactory manner. Recruits were forewarned about the length of the HIT, providing them with a chance to abandon the experiment if they felt the anticipated experiment time was too long to their liking.

6.2.4.3 Technical Requirements

Given worker limitations, we also enforced a number of technical constraints. Workers attempting each experiment were required to be either using a desktop or laptop computer

[9] Note that two different sets of workers were used – the studies were run at different times.

[10] This was correct at the time of writing; value acquired from the *U.S. Department of Labor* at `https://www.dol.gov/whd/minwage/america.htm` **LA** *2019-02-25* .

with a screen sufficiently large enough to display the experimental interface without having to resort to excessive scrolling. This also ensured a consistent number of result summaries would be present on different worker's screens. As such, we imposed a minimum display resolution of 1024 × 768 for both studies.

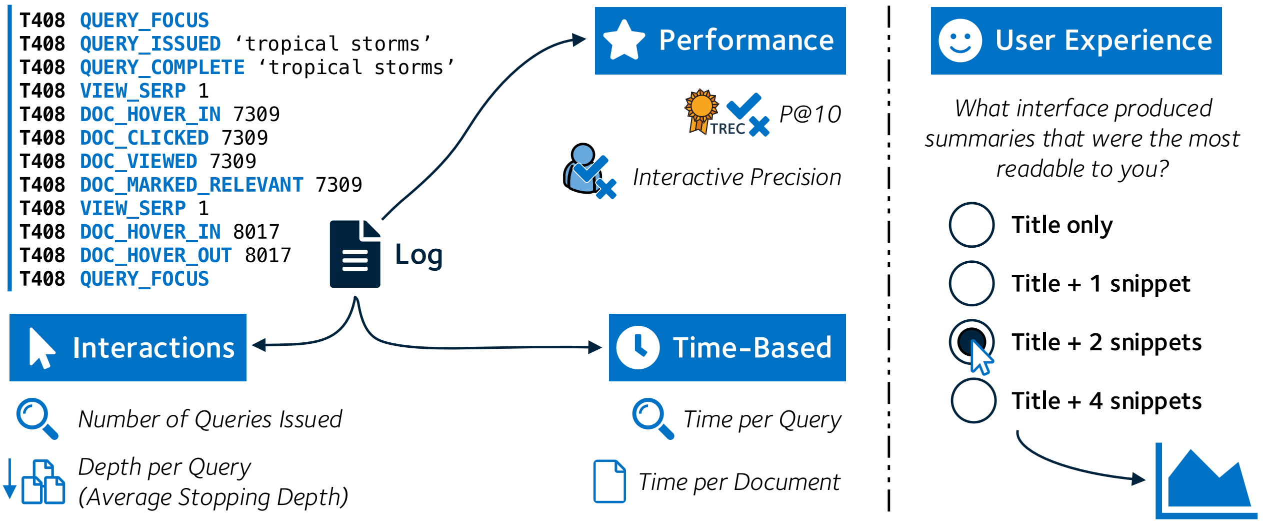
Conducted through a web browser, we wanted to ensure that only the controls provided by the experimental apparatus were used, meaning that the popup window that we presented in Figure 6.2 had all other browser controls disabled to the best of our ability (i.e. browser history navigation, etc.). The experimental system was tested on several major web browsers (including *Google Chrome, Mozilla Firefox, Apple Safari* and *Microsoft Edge)*, across different operating systems (including *Microsoft Windows, Apple macOS* and several *Linux* distributions, focusing on *Ubuntu)*. This gave us confidence that a similar experience would be had across different system configurations.

6.3 Extracting User Study Data

As discussed in Section 6.2.3, the **TREConomics** framework provided the necessary infrastructure for us to log the various interactions and capture survey responses from each individual subject across the two user studies trialled. In this section, we provide details on the different aspects that we subsequently used to evaluate searcher behaviours, performances and user experiences. Figure 6.4 provides a graphical illustration of how we split these various aspects into four distinct categories.

The first three categories can be extracted directly from the interaction log that recorded different interactions by each subject as they progressed throughout each experiment. The categories we considered are listed below.

- **Behavioural** measures capture the broad interactions that take place, such as the number of documents that a searcher examined in detail.



```
T408 QUERY_FOCUS
T408 QUERY_ISSUED 'tropical storms'
T408 QUERY_COMPLETE 'tropical storms'
T408 VIEW_SERP 1
T408 DOC_HOVER_IN 7309
T408 DOC_CLICKED 7309
T408 DOC_VIEWED 7309
T408 DOC_MARKED_RELEVANT 7309
T408 VIEW_SERP 1
T408 DOC_HOVER_IN 8017
T408 DOC_HOVER_OUT 8017
T408 QUERY_FOCUS
```

Figure 6.4 An illustration of the different types of measures that are captured, and from what sources. Interaction, time-based and performance measures are derived from the user study experiment log (with TREC QRELs used in conjunction with the interaction log to compute a subject's performance). User experience metrics are collated from a number of different surveys.

- Performance measures could then be extrapolated, with aid of TREC QRELs to ascertain the performance of subjects.
- Time-Based measures can also be derived from directly examining the interaction log, measuring the time spent between different logged interactions.

In addition to these categories, we also considered a number of user experience measures that were derived from a series of surveys. As highlighted in Section 6.2.1, surveys were presented to subjects at a number of different stages throughout the experiment. In conjunction with the three log-based categories defined above, the user experience measures could be used to complement the empirical evidence to test whether the interactions of subjects actually correlated with their perceived experiences.

In all, the interactions (including aspects such as clicks and time-based measures) were used as a *grounding* for our subsequent user simulations of interaction. How we grounded these simulations is discussed in Section 6.4.2. The grounding was undertaken in conjunction with an analysis of the interactions recorded, examining how stopping behaviours varied under different interfaces and conditions.

6.3.1 Behavioural Measures

Recorded solely from interaction log data, the basic interactions covered a large proportion of the aspects we considered in our analyses. The key behavioural measures that we examined are listed below.

- **Queries** The number of queries that had been issued by subjects.
- **Documents** The number of documents that were examined (viewed).
- **SERPs** The number of SERPs that were examined.
- **Examination Depth** The depth to which subjects clicked on (and hovered over) result summaries on the associated SERPs.

From these measures, we could ascertain whether searcher behaviours varied when a certain condition or interface was changed – allowing us to address questions such as: *whether snippet length affects the depth to which subjects examine content?* To compute depths, click and hover depths were used – we however only report click depths in subsequent chapters. The reasoning for this is discussed in Section 6.3.2 below. All of the aforementioned measures were computed on a per-session basis. Means and totals for each measure were computed for each session (where appropriate).

6.3.2 Time-Based Measures

As discussed in Section 6.2.3 – and also illustrated in Figure 6.3 on page 147, each logged interaction was saved with a timestamp which allowed us to determine when each event occurred.[11] With these timestamps, we could measure the time between two associated events, thus yielding the time taken to perform a given activity. We considered five key time-based measures across both user studies. These are enumerated below, along with a description of the log events.

- **Queries** This measure considered the time spent by a searcher issuing queries to the retrieval system. This was captured from the point at which the searcher focuses upon the query box (`QUERY_FOCUS`) to the point at which the query was submitted (via the `QUERY_ISSUED` event, either by pushing the `Search` button, or the subject hitting ↩ on their keyboard).
- **SERP Content** This measure considered the total amount of time that a searcher spent on a given SERP. This was captured as the point at which the subject was presented with the SERP itself (`VIEW_SEARCH_RESULTS_PAGE`) to the point at which they left – either through the issuance of a further query, clicking on a document hyperlink, or navigating to one of the other views of the experimental interface.
  - **Result Summaries** As discussed below, this was the mean time spent by a subject examining individual result summaries on a given SERP. This was included within the SERP content time.
- **Documents** This measure considered the time a subject spent on the document view. This was captured as the point at which the document was presented on the subject's screen (`DOC_MARKED_VIEWED`) to the point at which they left, which, like the SERP content time, could be determined from a number of different events, such as the event logged when returning to a SERP (`VIEW_SEARCH_RESULTS_PAGE`).

The fifth time-based measure that we considered in our reporting of results was an amalgamation of the four listed above.

[11] Timestamps were saved to the nearest thousandth of a second, as per the specification of the standard Python logging framework – refer to `https://docs.python.org/2/howto/logging-cookbook.html` **LA** *2018-05-29* for an example of the framework in action.

- Total Session Time The total session time was the addition of each of the times measured above. This was essentially the same as from the very first QUERY_FOCUS event to the TASK_COMPLETED event, which is either triggered by an interface timeout (Chapter 7), or the subject ending the task herself or himself (Chapter 8).

It should be noted that in this thesis we report all durations in seconds. We considered a number of different options when measuring each of the above. For example, querying time is measured only as the time the searcher spends interacting with the query box. Subjects may well have spent longer considering what terms to enter, perhaps as they were browsing existing content. However, this could not be captured; our logging tools were not capable of capturing this additional time.

A further option used was the time per result summary. This was computed by dividing by the click depth reached on a given SERP by the duration between the first *hover* event, where the subject hovered his or her cursor over the <div> container of a result summary, and the time at which they left the SERP. The first hover event was chosen as it was deemed to be a good indicator of the beginning of interaction with result summaries. The mouse cursor has been shown in prior studies to correlate strongly with the subject's gaze on the screen (Chen et al., 2001; Smucker et al., 2014). However, issues with network latency meant that several of the hover events were logged in the incorrect order, making the approach of measuring each individual HOVER_IN and HOVER_OUT event unreliable. Using the click depth and total SERP time provided us with a value with which to work. The approach also assumes that subjects examined each result summary on a SERP up to a particular depth, spending an equal amount of time examining each. This was sufficient for the work in our studies to ascertain whether or not a variation in the task goal or presentation of results affected the depths to which subjects examined results.

These time calculations and approximations were also used as a means for providing *grounding* to the simulations of interaction, as we discuss later in Section 6.4.2.1. It should also be noted that the time per interaction could also be computed, such as the *time per query*. This was simply considered as the summation of the querying time across the entire session, for example, divided by the total number of queries issued. The same principle could be applied for the *per SERP time* and the *per document time* by substituting querying with SERP and document times respectively.

6.3.3 Performance Measures

In conjunction with behavioural and time-based measures, we were also able to extract a number of different performance measures from the interaction logs.[12] Key performance measures that we captured included:

- **query performance**, primarily measured with *P@*10 (although additional *P@k* values are reported); and
- **interactive precision and recall** (as discussed in Section 2.4.2.1), including:
  - the number of documents saved (identified as relevant); and
  - the number of those documents that were TREC relevant (and vice-versa).

6.3.4 Demographics and User Experience Surveys

A number of surveys were also filled out by subjects. These captured different information about each searcher's individual search experiences. While there are similarities between what is asked (refer to Sections 7.2.1 and 8.2.1.5 for further details), we provide in this section a high-level overview of the different surveys, before examining questions that were common between the two studies. Below, we outline the demographics, pre- and post-task surveys, and post-experiment surveys – provided in the order of the experimental flow detailed in Section 6.2.1.

[12] Some measures were computed with the `trec_eval` evaluation tool, discussed in Section 2.3.1.1.

6.3.4.1 Demographics

Details in keeping with general demographics were attained about the different subjects from this survey. These included: the subject's age; gender; their present occupation; and their highest level of professional qualification (from either high school, Associate's degree, Honours degree, MSc of PhD).

In addition to these basic questions, we also asked several questions pertaining to their perceived search proficiency. Questions included:

- how often they searched for information;
- what pointing device they were using (i.e. mouse, trackpad); and
- their preferred general purpose web retrieval system.

Considering that both of the user studies instructed subjects to imagine they were newspaper reports (and search a collection of news articles), we also asked them how often that they explicitly searched for news articles online.

6.3.4.2 Pre-Task

Between both user studies, we asked the same questions within the pre-task survey. Subjects were provided with a short description of their search task and a topic description, which in turn provided their information need for the said task. After examining the topic description, subjects were then queried on the following:

- how well they knew about the topic prior to this study;
- how relevant the topic was to their life;

- how interested they were to learn more about the topic;
- whether they had searched for information related to the topic before; and
- how difficult they felt it would be to search for information on the topic before commencing the search task.

Responses were provided on a seven-point Likert scale, providing the option for neutrality between the two extremes – extremes being *nothing/not at all/very difficult* to *lots/very much/very easy*. Responses to these questions helped us gauge the perceived difficulty of the task, and ascertain how much background knowledge could potentially affect results.

6.3.4.3 Post-Task and Post-Experiment

Post-task and post-experiment surveys were unique to each of the two user studies. Sections 7.2.1 and 8.2.1.5 provide further information on what questions were asked. However, the post-task surveys focused on how well the subjects thought they (and the retrieval system, under the given condition and/or interface) performed during the search task. Post-task surveys considered the experiment as a whole, asking questions about what condition and/or interface the subjects preferred, or performed better, for example.

6.4 Simulating Searcher Behaviours

With the general layout and components of the two user studies explained, we now consider how we *simulated searcher behaviours*. The simulation of interaction provides a low-cost means of exploring a variety of different searcher strategies and configurations (Azzopardi et al., 2011). In this section, we provide an overview of the general aspects of the stochastic searcher simulations, which are reported in the later chapters of this thesis. In this section, we discuss:

- how our simulations were *grounded;* and
- how we instantiated the different components of the CSM defined in Chapter 4 for our simulation experiments.

We conclude the chapter with a discussion as to how we evaluate the results from our simulations, allowing us to determine what stopping strategies offer the best overall performance and approximations of real-world searcher behaviours. This is done in consideration of the two user studies discussed in Chapters 7 and 8. By grounding our simulations with data derived from the two aforementioned user studies, we can then obtain an insight into how searcher stopping behaviours vary under different contexts.

The Mean Searcher Comparisons between simulated searchers and the results of real-world searchers are made between the *average* behaviours observed. This average behaviour is considered across each of the different experimental interfaces and conditions that we trial across the two user studies, discussed in Chapters 7 and 8. This consideration:

- simplifies and reduces the number of simulations that are required to be run; and
- provides a simple overview of how stopping behaviour varies across each interface and condition, rather than across each individual searcher.

While the simplifications make it easier to report results, we acknowledge that the averaging/aggregation that takes place may hide subtle behavioural differences that can be observed between searchers. We discuss this particular limitation of our work later in Section 10.3.4 on page 355.

Considering Stochastic Simulations We consider a series of different *stochastic* simulations that mimic searcher behaviours. Stochastic simulations of interaction rely on probabilities (typically extracted from real-world log data) to determine the likelihood of a particular action occurring (e.g. clicking on a link presented on a SERP). Taking an example of a prior study using this approach, Yilmaz et al. (2010) used interaction probabilities for deducing the likelihood of clicking on an attractive result summary – something that is extensively used throughout the simulations reported in this thesis. We discuss this further in Section 6.4.2.3. Again, if such probabilities are grounded from real-world interaction data, this increases the realism of the simulations.

These models considered stochastically determining, for instance, the attractiveness of a result summary to the given information need – something that we also utilise. We discuss this further in Section 6.4.2.3.

6.4.1 The SimIIR Framework

All simulations of interaction reported in this thesis were run on the **SimIIR** framework, a custom-built framework for the simulation of interaction. It captures the wider IIR process (Maxwell and Azzopardi, 2016b).[13] The framework consists of a number of individual *components,* each which must be instantiated to yield a *simulation.* Figure 6.5 provides an illustration of the framework's basic architecture, highlighting each of the individual simulator components, and the framework's key outputs.

In this section, we briefly outline each of these components, discussing the need for each. Each of these components can be mapped to one of the individual decision points and/or activities of the CSM, as outlined in Section 4.1 on page 108.

A *simulation* within the **SimIIR** framework consists of the following main components.

- **Topics** One or more topic(s) can be provided, each consisting of a title and topic description (i.e. the TREC topic descriptions, as per Section 6.1.3).

[13] **SimIIR** can be accessed at `https://github.com/leifos/simiir`. **LA** *2018-05-29*

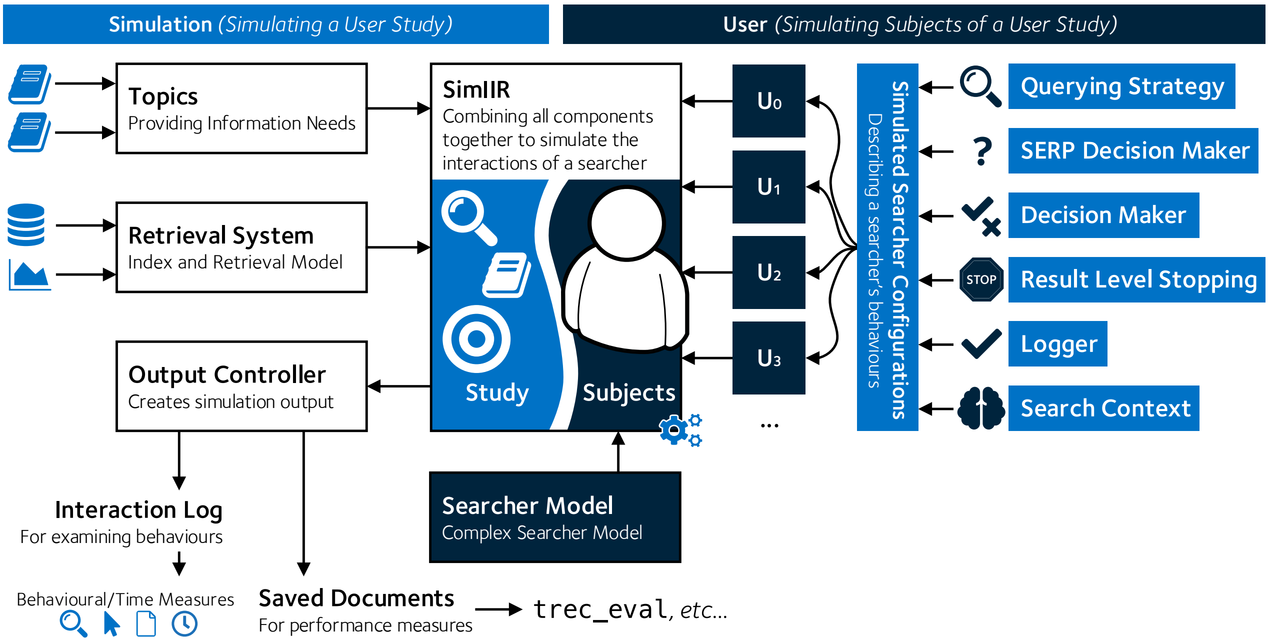

Simulation *(Simulating a User Study)*
User *(Simulating Subjects of a User Study)*
Topics
Providing Information Needs
Retrieval System
Index and Retrieval Model
Output Controller
Creates simulation output
SimIIR
Combining all components together to simulate the interactions of a searcher
Study
Subjects
U0
U1
U2
U3
...
Simulated Searcher Configurations
Describing a searcher's behaviours
Querying Strategy
SERP Decision Maker
Decision Maker
STOP
Result Level Stopping
Logger
Search Context
Searcher Model
Complex Searcher Model
Interaction Log
For examining behaviours
Behavioural/Time Measures
Saved Documents
For performance measures
trec_eval, etc...

Figure 6.5 The architecture of the **SimIIR** framework, with the components split across both *simulation* and *user* categories. Simulation components define the simulation — a representation of some real-world user study, with user components defining the behaviours of simulated searchers.

- **Retrieval System** An interface is provided to link an underlying retrieval system with the simulation. In the case of this thesis, this component links back to the setup described in Section 6.1.2.
- **Output Controller** This component is responsible for generating the output files that can be fed into evaluation programs such as `trec_eval`, as outlined in Section 2.3.1.1.

Simulations also consist of one or more **simulated searchers**. These searchers attempt to complete a given search task, having been instantiated with differing constraints. A simulation is therefore in essence loosely associated with the concept of a real-world *user study*. Each individual simulated searcher can be likened to an individual subject of a user study. In turn, each of the simulated searchers is defined by a series of additional components that *describe their behaviours*.

- **Querying Strategy** The querying strategy determines how queries are generated from topic descriptions, and subsequently selected.

- **SERP Decision Maker** This decision maker determines where a searcher should *enter* a SERP and begin to examine individual result summaries, or abandon the SERP and issue a subsequent query. This corresponds to the new SERP level stopping decision point of the CSM, discussed in Section 4.3.1 on page 113.

- **Decision Maker** These components are responsible for judging the attractiveness and relevance of result summaries and documents respectively. For the work in this thesis, this component is stochastic and grounded on interaction data.[14]

- **Result Summary Level Stopping Strategy** This component, instantiated using one of the stopping strategies outlined in Chapter 5, determines the point at which a simulated searcher will stop interacting with a ranked list of results.

- **Logger** The logger component is responsible for providing *interaction costs* for particular interactions (e.g. issuing a query), keeping track of the combined session time, and determining whether the overall search session goal, time limit – or other session level stopping constraint – has been met.

- **Search Context** This component can be considered as a basic representation of a searcher's memory, keeping track of the different prior interactions. Examples include the result summaries and documents that have been examined, prior queries that have been issued, and what documents that have been saved (considered relevant).

All the above components are underpinned by a **searcher model** component, providing a flow of interactions to the search process undertaken by simulated searchers. In all simulations reported in this thesis, the CSM represents this component, and outlines the different sequence of interactions that can occur between the different components. We do not discuss further technical details about how the **SimIIR** framework can be instantiated here; refer to Maxwell and Azzopardi (2016b) for more information.

[14] Deterministic decision maker components have also been developed – refer to (Maxwell and Azzopardi, 2016a) for more details.

6.4.2 Grounding and Instantiating Simulations

To ensure that the simulations of interaction that we report in this thesis are as realistic as possible, we *grounded* the simulations using real-world observations extracted from interaction log data. Doing so ensured that the results generated from the simulations were credible abstractions of reality (Azzopardi et al., 2011). Given the CSM (and the means by which we can evaluate it, as reported in Section 4.5 on page 119), we considered grounding our simulations from three perspectives.

- **Query Generation** We consider the generation of *psuedo-realistic* queries to issue to the underlying retrieval system. As discussed in Section 6.4.2.2, these queries are generated using *querying strategies* that are created from observing real-world searcher querying behaviours. We also *replay* the queries issued by real-world subjects of the user studies for one set of simulation runs.
- **Interaction Costs** We extract a series of different interaction costs from log data to ensure that the time spent by simulated searchers is an average representation of the time spent by real-world searchers under a particular search context.
- **Interaction Probabilities** As with interaction costs above, we also considered a series of grounded interaction probabilities that determine the likelihood of a simulated searcher determining whether to: *enter* a given SERP (used in Chapter 9 only); the attractiveness of a result summary; or the relevance of a document.

These are considered in conjunction with the twelve stopping strategies (as discussed in Chapter 5), and the various constraints that we imposed upon each searcher, such as a time-limited search session. The remainder of this section is left to a detailed discussion of the key components of our simulations of interaction. In particular, this section focuses upon how we instantiated each of the individual components of the **SimIIR** framework to build realistic, credible simulations of the search process.

6.4.2.1 Interaction Costs

Considering the individual CSM components as illustrated in Figure 4.1 on page 109, a number of different *interaction costs* can be derived. These are costs that must be expended by searchers subscribing to the model in order for them to successfully complete the search process. In conjunction with the time-based measures discussed in Section 6.3.2, we identified five different interaction costs that searchers are faced with and thus use in our simulations. These are listed below, with an illustration of the costs provided in Figure 6.6.

- **Querying** This considers the **Issue Query** activity of the CSM, and considers the time required by a simulated searcher to enter (and subsequently submit) a query into the retrieval system's interface. Again, this is considered as from when the subject focused on the query box, to the point where they submitted the query.
- **SERP Examination** This cost considers the **View SERP** activity, and denotes the time spent by a searcher considering whether the presented SERP is attractive enough to *enter* and examine in more detail. This is considered as the point at which the SERP is rendered on their screen, to the point where they begin interacting with it in any way.
- **Result Summary Examination** The **Examine Snippet** activity is considered here, this being the time required to examine an individual result summary for attractiveness. Estimations for this interaction cost are described in Section 6.3.2.
- **Document Examination** This costs denotes the amount of time required to assess a document for relevance to the given information need. This is the **Assess Document** activity in the CSM.
- **Saving** The **Save Document** activity is considered for this final cost, where a searcher will actively save and identify the document as relevant. This is considered as the time from the point at which a searcher clicked the Save button to when they left the document and returned to the SERP.

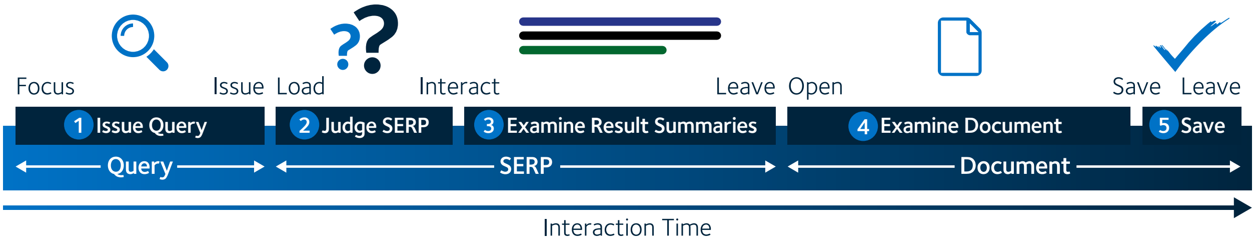


Figure 6.6 Illustration of the five interaction costs paid by searchers subscribing to the CSM. Each cost is shown with the start and end events by which the costs were measured from the user study interaction logs. Time spent on individual interface components is shown in white. Refer to Section 6.4.2.1 for a detailed explanation of each interaction cost considered.

The derived costs are averaged for each of the conditions and interfaces trialled in the studies reported in Chapters 7 and 8. Refer to Table 7.6 on page 217 and Table 8.10 on page 283 for Chapters 7 and 8 respectively. These tables show the actual costs that were extracted from user study interaction data.

Fixed Interaction Costs All simulations of interaction discussed in this thesis rely upon the notion that all interaction costs are *fixed* over each interface and condition trialled. For example, this means that no variation in querying time exists between a searcher who issues single term queries, and another searcher entering terms that consist of three terms. All queries require the same cost to be entered and submitted. This decision was taken to reduce the complexity of our simulations. By including dynamic interaction costs, this would have made the simulations themselves – and the subsequent comparisons – much more complex. Previous work such as time-biased gain (Smucker and Clarke, 2012) has however shown that estimations of dynamic interaction costs can be made.

6.4.2.2 Query Generation Strategies

The generation of queries is an important aspect of any simulation of interaction. Starting from the simplistic TREC-style searcher model where a single query is issued, numerous studies have focused upon the issue of query generation, and how one can generate a series

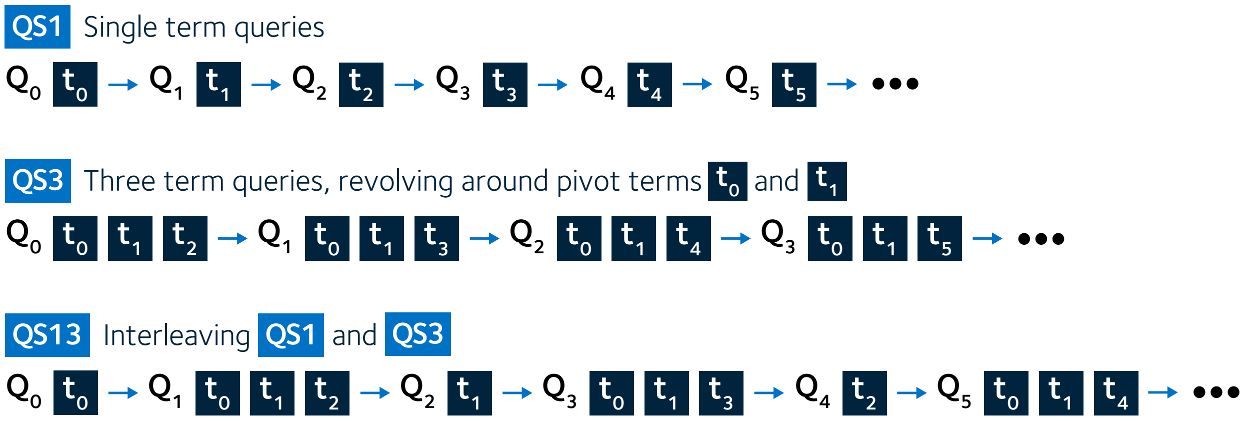

QS1 Single term queries
Q_0 t_0 → Q_1 t_1 → Q_2 t_2 → Q_3 t_3 → Q_4 t_4 → Q_5 t_5 → •••
QS3 Three term queries, revolving around pivot terms t_0 and t_1
Q_0 t_0 t_1 t_2 → Q_1 t_0 t_1 t_3 → Q_2 t_0 t_1 t_4 → Q_3 t_0 t_1 t_5 → •••
QS13 Interleaving QS1 and QS3
Q_0 t_0 → Q_1 t_0 t_1 t_2 → Q_2 t_1 → Q_3 t_0 t_1 t_3 → Q_4 t_2 → Q_5 t_0 t_1 t_4 → •••

Figure 6.7 Extensive examples of the three querying strategies used in this thesis, **QS1**, **QS3**, and **QS13**. Queries are denoted by Q_n, with individual terms denoted by t_n. In these examples, a total of six terms are used (from t_0 to t_5). Queries are separated by arrows (→).

of pseudo-realistic queries based upon prior interaction data. We highlighted some of these prior works in Section 2.3.4 (page 52).

In this thesis, we consider a number of different *querying strategies* as proposed by Keskustalo et al. (2009) and Baskaya et al. (2013) in order to generate queries for our simulations of interaction. These strategies are considered to be *idealised, prototypical* approaches to query generation, themselves being grounded from a prior user study examining the query behaviour of subjects.[15] Of the five strategies identified by the authors, we consider two in this thesis that were shown in simulations by Keskustalo et al. (2009) to yield the *worst* and *best* performance – but also shown to reflect actual searcher queries. The two querying strategies, identified as **QS1** and **QS3**, are briefly explained below. We also provide an illustration of the two strategies in Figure 6.7, where Q_n denotes query n within a search session, and t_n denotes query term n from a list of terms available to formulate queries.

- **QS1 Single Term** This querying strategy generates a series of *single term queries.*
- **QS3 Three Term** This second querying strategy generates queries with two *pivot* terms, and one additional term. Therefore, the first two terms remain constant, with the third term changing for each subsequent query.

[15] Refer to Keskustalo et al. (2009) for further information on the user study undertaken.

Queries generated by QS3 are considered to be realistic in the sense that queries issued in real-life web search sessions consist of three terms on average (Keskustalo et al., 2009).

With these two querying strategies in mind, we then *interleaved* the two strategies together. This ultimately yields a third querying strategy that we identify as QS13.

- QS13 Interleaved With this querying strategy, queries from both QS1 and QS3 are generated and subsequently interleaved between each other, starting with the first query from QS1.

Refer to Figure 6.7 for an example of how this querying strategy works. These querying strategies allowed us to test the *robustness* of each result summary stopping strategy. Recall that Keskustalo et al. (2009) highlighted that QS1 yielded relatively poor performance compared to QS3. Therefore, it follows that a searcher, when issuing a query generated by QS1, will observe that the results presented are of poor quality, and thus stop at a shallow depth when compared to examining results of queries issued by QS3. Examining many results from a poor query is by and large a waste of the searcher's time, so robustness of result summary stopping strategies can be checked to see if queries generated by QS1 are abandoned earlier than those generated by QS3.

Reported Querying Strategy With interleaved querying strategy QS13 allowing us to test the robustness of various simulation configurations, we provide a novel report on simulations of interaction using this interleaved querying strategy. As previously discussed, we also *replayed* real-world queries issued by user study subjects. Refer to Section 6.4.3.2 for further information on how this was achieved.

Term List Generation Given the querying strategies, how did we then generate the ranked list of terms to be used, shown as t_0 to t_5 in Figure 6.7? For all simulations in this thesis, all terms were derived from the given TREC topic title and description. For all queries, stopwords were removed as per the stopword list defined by Fox (1992).

For QS1, we combined the title and description terms together and creating a *Maximum Likelihood Estimate (MLE)* language model, allowing us to create a probability distribution for the likelihood of a term to appear in a topic description, i.e. $p(term|topic)$. A list of single term queries was then ranked in descending order by this probability to yield the set of queries we would use for t_0, t_1, t_2, and so on.

A similar approach was used for QS3. The same MLE approach was used to rank the title and description terms separately, creating two separate rankings of terms. For the *pivot terms* – the two terms that are consistently used as the first two terms of each QS3 query – all possible two term title terms were used, with the highest joint probability being selected as terms t_0 and t_1. Single terms from the topic description were then used as per QS1, with the descending probability ordering used to then determine the order in which the third query term was selected.

6.4.2.3 Summary and Document Decision Making

As discussed earlier in Section 6.4, our simulations were stochastic in nature. Decisions pertaining to the attractiveness of a result summary *(should I click this link and examine it further?)* and the relevance of a document to the given information need *(should I save this document?)* were determined through a series of different *interaction probabilities.* Chapters 7 and 8 present the interaction probabilities used within the simulations. In this section, we describe the approaches used to derive them.

In parallel with earlier studies considering the simulation of interaction – such as those by Yilmaz et al. (2010) and Baskaya et al. (2013), for example – result summary and document decision making both revolve around two key probabilities:

- the probability P(C) of considering a given result summary on a SERP to be sufficiently attractive to *'click'* and load the associated document; and

- the probability **P(S)** of determining the document to be relevant to the given information need after examination, and thus *saving* it.

These are considered separately. The action of requesting a document from clicking on the associated result summary does not necessarily mean that the document *is* relevant; merely, it means it appears attractive enough to examine in more detail (Turpin et al., 2009). The above probabilities are broken down further with regards to TREC relevance. This warrants an examination of the TREC relevance judgements to determine whether the result summary and/or document being clicked and/or saved would be considered to be relevant to the given topic by the TREC assessors. As such, $P(C)$ and $P(S)$ can be split further, such that we can then determine:

- the probability that a result summary that has been clicked is TREC relevant **P(C|R)** or not **P(C|N)**; and
- the probability that a document saved is TREC relevant **P(S|R)** or not **P(S|N)**.

From these definitions, we may take the interaction logs from the two user studies, split the interactions by the interface or condition for which probabilities were derived, and summate the different measures – as shown by the equation for calculating $P(C)$, where:

$$P(C) = \frac{|clicked_{Rel}| + |clicked_{\neg Rel}|}{|examined|}$$ **Equation 6.1**

and $P(S)$, where:

$$P(S) = \frac{|saved_{Rel}| + |saved_{\neg Rel}|}{|examined|}.$$ **Equation 6.2**

In the above equations, *Rel* denotes the count for TREC relevant items, with $\neg Rel$ representing items that were not TREC relevant. Finally, $|clicked|$ represents the number of result

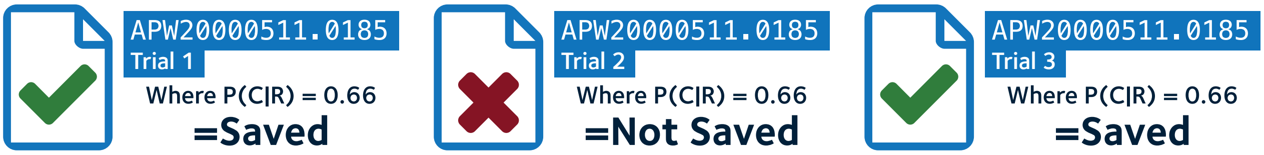


Figure 6.8 An example of the same TREC relevant document being judged differently across multiple trials. This can have a negative effect upon simulation runs, yielding dubious results.

summaries that were clicked (deemed attractive enough to examine further), $|examined|$ denotes the total number of documents that were assessed in full (i.e. presented in the document view as shown in Figure 6.2), with $|saved|$ denoting the number of documents that were identified as relevant, and subsequently *saved*.

To compute probabilities concerning TREC relevance only, $P(C|R)$ and $P(S|R)$ were defined as $P(C|R) = \frac{|clicked_{Rel}|}{|examined_{Rel}|}$ and $P(S|R) = \frac{|saved_{Rel}|}{|examined_{Rel}|}$. These definitions are the same as the above, sans the non-relevant, or $\neg Rel$ values. Conversely, $P(C|N)$ and $P(C|R)$ were defined as $P(C|N) = \frac{|clicked_{\neg Rel}|}{examined_{\neg Rel}}$ and $P(S|N) = \frac{|saved_{\neg Rel}|}{examined_{\neg Rel}}$ – this time, without the TREC relevant judgements included (i.e. a judgement of `0`, or no judgement at all).

Monte-Carlo Simulations Stochastic approaches to modelling interactions provide a simple means of operationalising the components of the simulation. Such an approach judges the attractiveness and relevance of result summaries and documents with a roll of the dice, rather than explicitly examining the content provided to formulate a judgement.

However, such an approach is not without limitation. By their very nature, a stochastic simulation based upon random probabilities will require a large number of different trials to be executed from which an *average* may be computed. Each trial will potentially result in a different outcome, as illustrated in Figure 6.8. With a probability of clicking document `APW20000511.0185` set to 0.66, there is a 66.66% chance that the document would be clicked in each run, resulting in different outcomes across trials.

Different outcomes can lead to a wide variance between different trials, which in turn leads to a requirement of running a more powerful experiment over a larger number of trials, or

Table 6.2 A visual example of how the use of pre-rolled judgements (right) performs when compared without (left). The top five rankings (from **D**$_1$ to **D**$_5$) across four examples are shown. On the left are two simulated searcher configurations that do not use pre-rolled judgements. On the right, the same configurations are shown with pre-rolled judgements. Notice that for the examples with pre-rolled judgements, the same judgements are provided across configurations. On the left, this is not the case — indeed, this phenomenon can have an undue influence upon the behaviour of a simulated searcher, such as affecting their stopping behaviour. This is illustrated below, with the stopping depth for SS2-NT @2 being affected (stopping depths are highlighted in red).

Without pre-rolled (incomparable)				**With pre-rolled (comparable)**			
SS1@5		SS2@2		SS1@5		SS2@2	
D$_1$	R	D$_1$	N	D$_1$	R	D$_1$	R
D$_2$	N	D$_2$	R	D$_2$	R	D$_2$	R
D$_3$	N	D$_3$	N	D$_3$	N	D$_3$	N
D$_4$	R	D$_4$	R	D$_4$	R	D$_4$	R
D$_5$	R	D$_5$	R	D$_5$	N	D$_5$	N

Monte-Carlo style simulations (Benov, 2016). In turn, this leads to an increase in the amount of time (and computational power!) required to execute all simulations.

Pre-Rolled Judgements In Figure 6.8, if the same document is judged differently between individual trials, the results from two different simulated searcher configurations are incomparable. This is demonstrated in the left hand side of Table 6.2. With a given probability, a result summary can be judged to be attractive or not ($P(C)$). Without proper controls, these judgements will vary across trials, meaning that in Table 6.2, SS1-FIX @5 and SS2-NT @2 are incomparable (refer to Section 6.4.2.6 for more information on these configurations) because the judgements for the same ranked list of results differs. As these judgements change, this has an undue influence on the behaviour of the simulated searcher across different experimental configurations.

In order to address this issue and permit for fair comparisons across simulated searcher configurations (and to allow for reproducible results), we used **pre-rolled judgements** that determined the attractiveness or relevancy of result summaries and documents respectively *a priori.* For each document within the AQUAINT collection, we pre-computed outcomes for each individual document using the relevant grounded interaction probability, storing them in *action judgement files.* One related to the action of clicking on result summaries (i.e. $P(C)$), with the other relating to saving documents (i.e. $P(S)$). These judgement files were then used by the decision maker components. All this component had to do was then perform a simple lookup for the corresponding document judgement for a given trial.

By pre-computing these judgements in advance, the same document would therefore be considered relevant in the same trial under a different configuration. This is demonstrated in Table 6.2, this time on the right-hand side. Across this trial, judgements by the simulated searcher have been pre-rolled in advance, and thus the judgements for documents D_1 to D_5 are the same. As they are the same, this permits fairer comparisons between different configurations, with the table showing how the simulated searcher's behaviour varies.

It should also be noted that the generated pre-rolled judgements were seeded to allow for these files to be easily reproduced. This process was repeated 50 times to account for variability between trials. This meant that for every simulated searcher configuration, we ran a total of 50 trials in which result summaries could be considered attractive or not, and documents could be considered relevant or not. As such, all results reported later in this thesis are an average over 50 trials. This value was selected since for each of the two user studies reported in Chapters 7 and 8, approximately 50 subjects took part in each.

6.4.2.4 Computing Gain

As searchers examine information, they *gain knowledge* that helps shape their mental model of the underlying information need (Nickles, 1995). In the simulations reported in this thesis, gain is acquired when a simulated searcher, having examined a document, subsequently considers said document is relevant, and saves it. During post-hoc analysis, we then can compute how many documents were saved, and how many were saved and TREC relevant. This is determined by looking up the TREC relevance judgements. Gain for the document is then simply computed as the relevance judgement score from the TREC QRELs. Given that we utilised the TREC 2005 Robust Track (Voorhees and Harman, 2005), *graded relevance judgements* are used. This value, when summated over all saved documents, is the *Cumulative Gain (CG)* score for a simulated trial. This is discussed in more detail in Section 6.4.3.1.

6.4.2.5 SERP Level Decision Making

Previously outlined in Section 4.2, the CSM includes an additional SERP level stopping decision point. Motivated by the *information scent*[16] offered by a SERP (or *patch*), this stopping decision point joins other established decision points, including result summary and session level stopping. The new decision point permits a searcher subscribing to the CSM to either *enter* the SERP and begin examining result summaries in detail if the SERP appears to offer a good scent, or abandon the SERP if it appears to be poor (and saving time).

For our simulated analyses of the user studies reported in Chapters 7 and 8, the SERP level stopping decision point is not considered. In other words, a simulated searcher will *always* examine a given SERP for content in detail – labelled SERP Always in this thesis. This acts as our baseline for the SERP level stopping decision point. This decision was taken to simplify our results and to provide greater emphasis upon how the different snippet level stopping strategies (as discussed below) affect searcher behaviour and performance.

Chapter 9 provides empirical results for different result summary level stopping strategies when the new SERP level stopping decision point was utilised. This chapter also provides a detailed explanation as to how we operationalised the new stopping decision point, as presented in Section 9.2.1 on page 311.

[16] To recap, we discuss the notion of information scent in Section 3.3.1.1 on page 92.

6.4.2.6 Result Summary Level Stopping Strategies

In Chapter 5, we outlined twelve stopping strategies that had been operationalised from various stopping heuristics and the RBP evaluation measure. In this section, we once again enumerate each of the different stopping strategies, discussing what *stopping threshold values* that we trialled for each. Any specific implementation details to a given stopping strategy are also noted in this section.

For several of the stopping strategies, threshold values need to be approximated from real-world interaction data to provide some grounding. An example of such a threshold is the RBP patience parameter, as we discussed in Section 2.4.1.5 (page 67). As such, some of the threshold value ranges were approximated from the user study interaction data in Chapter 7, and used in all simulation experimentation for consistency. The values that we ascertained were of most interest as they offer close approximations to how real-world searchers actually behaved in consideration of the respective stopping strategies.

Below, we begin our discussion of each of the different stopping strategies, outlining the parameter thresholds used.

- SS1-FIX For our fixed depth stopping strategy, we trialled a range of values, where x_1 was set from 1 to 10 in steps of 1, and then 15 to 24 in steps of 3. This resulted in 14 separate parameter threshold configurations, with enough values such that a searcher would comfortably reach the time limits imposed in the study detailed in Chapter 7.
- Both SS2-NT and SS3-NC used the same range of values for threshold values x_2 and x_3. These stopping strategies focused upon a searcher's tolerance to non-relevance, as discussed in Section 5.2.1.

Note that for both SS2-NT and SS3-NC, any document that has been previously examined during the same search session (returned in the ranked results of a prior query) will be included in the count of non-relevant items. This is opposed to ignoring previously observed items, which has been shown in prior work to offer poorer performance (Maxwell et al., 2015b).

Next, we enumerate the stopping strategy considering the *satiation* of a searcher.

- SS4-SAT Concerning the number of documents that a searcher should save before being satisfied, we examined a range of values for x_4, from 1 to 10 in steps of 1.

Our first combination heuristic then combines the frustration and satisfaction heuristics together. However, two frustration stopping strategies exist – SS2-NT and SS3-NC. To reduce the complexity of the simulations – and to corroborate with empirical evidence suggesting that this was the better performing strategy (refer to Maxwell et al. (2015b)), SS2-NT is assumed to be the non-relevant component of the combination strategy.

- SS5-COMB Here, we utilised the stopping threshold values defined for x_2 and x_4 above. These were for the frustration component SS2-NT and SS4-SAT respectively.

Next, we consider the two stopping strategies that focus on the difference threshold heuristic (Nickles, 1995), where searchers would abandon a set of results if the summaries provided did not appear to yield any new information.

- SS6-DT Considering the term overlap difference between a result summary and prior summaries, a range of values from 0.0 to 1.0 in steps of 0.05 were utilised. This was to explore the entire range of possible values. The smaller the threshold, the less similar the content of the new result summary to those previously examined.
- SS7-DKL Using KL-divergence, a range of values for x_7 were trialled, from 3.0 to 8.0 in steps of 0.5. A small-scale pilot study examining this stopping strategy over the AQUAINT index showed that a majority of values fell within this range.

For both SS6-DT and SS7-DKL, we considered both the *per-query difference* and the *per-session difference.* For the per-query variant, previously observed result summaries consisted of the first result summary, thus meaning that the simulated searcher would always consider at least two result summaries before deciding to stop. For the per-session variant, all previously observed result summaries over the entire search session (i.e. including those from previous queries) were used. In a pilot study, as reported by Maxwell et al. (2015a), we consider in this thesis the per-query variants only. These offered better performance than their per-session variants.

Next, we consider the stopping strategy based upon a searcher's optimal searching behaviour. This concerned computing a searcher's *average rate of gain.* To determine the rate of gain at a given result summary at rank i, we first computed the DCG (discussed in Section 2.4.1.4 on page 65) received from the observed documents, up to the point in the ranked list at position i. We then divided g (the DCG) by the total time taken, yielding $i * t_d + t_q$, where i represented the rank, t_d was the time required to examine a document, and t_q was the time required to issue a query.

- SS8-IFT Computing the searcher's average rate of gain as defined above, we considered a gain threshold (x_8) from 0.002 to 0.03 in steps of 0.002. A minimum of two result summaries were examined before calculating the average rate of gain ($y_8 = 2$).

The estimate computed was very dependent upon the first document in the ranked list. For example, if judged to be non-relevant, the searcher would gain 0 – meaning that the searcher would then immediately stop when $x_8 > 0$. To counter this, we also considered an additional parameter that specified how many result summaries the searcher should *always* examine before making a decision based upon the rate of gain experienced.[17] This would

[17] This second parameter y_8 was set to 2 for all experiments utilising SS8-IFT. A pilot study by Maxwell et al. (2015b) found that a value of 2 proved to be far less sensitive to non-relevant items, and resulted in better performance by the searcher.

essentially mean that a searcher employing SS8-IFT would look at a minimum of y_8 result summaries, and from there begin to make decisions as to whether they should continue.

Next, we consider the time-based stopping strategies as outlined in Section 5.5. These consider the total time spent examining a SERP and its associated documents, and the time since last identifying a relevant document.

- SS9-TIME Considering the total amount of time spent on a SERP and linked documents, we considered for this *total time* stopping strategy values for x_9, from 30 to 150 seconds in steps of 30 seconds.
- SS10-RELTIME Stopping (or *giving up* (Krebs et al., 1974)) after x_{10} seconds have elapsed since saving a relevant document (or the start of the search session if no documents have been saved), we consider for this stopping strategy a smaller range of values, from 10 to 50 seconds in steps of 10 seconds.

The parameter threshold values for x_9 and x_{10} were grounded using interaction data derived from the user study discussed in Chapter 7 only. For x_9, the mean time spent interacting on a SERP and its documents were computed at approximately 90 seconds – a 60 second decrease and increase were selected for the lower and upper bounds respectively. Likewise, a relevant document was on average identified approximately every 30 seconds. This provides motivation behind selecting the range of values chosen for x_{10}.

The second combination rule, based upon the combination heuristic proposed by McNair (1982), considers that a searcher decides whether a given SERP is of high yield at shallow ranks (or not). Depending upon the outcome of this decision, a different stopping strategy will be employed.

- SS11-COMB For a SERP yielding a high scent at shallow ranks, a searcher will employ the satisfaction stopping heuristic, SS4-SAT. The give-up time-based strategy

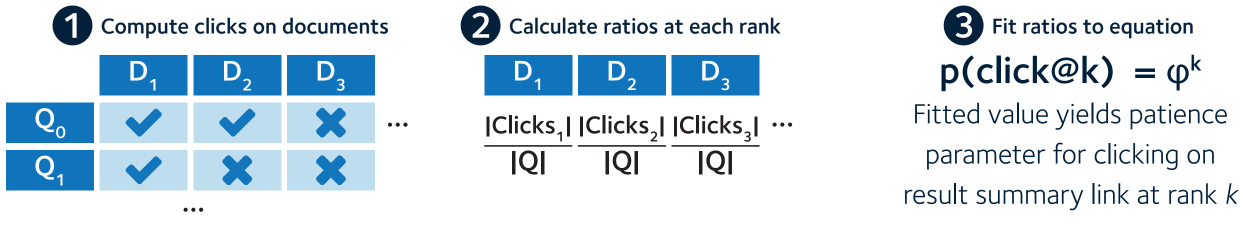
SS10-RELTIME is employed if the scent is low. Parameter threshold values are identical to those defined for x_4 and x_{10} above.

All combinations of values for x_4 and x_{10} were trialled. To determine the scent of a given set of ranked results (and thus the stopping strategy used), we considered the first ranked result, or $P@1$. If the first ranked document was TREC relevant, the SERP was assumed to provide a high yield early on. Conversely, if $P@1 = 0$, the SERP was judged to be low yield at shallow ranks. This matches with definitions of poor scented SERPs by Wu et al. (2014) and Hassan and White (2013)[18]. We also make the further assumption that the document presented at the first rank will be used to judge the SERP yield.

The final result summary level stopping strategy was based upon RBP. Once again, the patience parameter was again grounded from the user study outlined in Chapter 7 only.

- **SS12-RBP** Given a fitted patience parameter of 0.9087, we trialled a range of values around this range, from 0.8 to 0.95 in steps of 0.05. We also trialled 0.99 and 0.9087.

To estimate the patience parameter, we assumed that the patience of a searcher could be determined by considering the depths at which documents were clicked, as per RBP. The deeper the searcher went down a list of ranked results, the more patient they were considered to be. For every query issued, we determined whether the document at each rank of the corresponding set of results was clicked or not, as shown in step ❶ below.



❶ Compute clicks on documents

	D_1	D_2	D_3	...
Q_0	✓	✓	✗	...
Q_1	✓	✗	✗	
...				

❷ Calculate ratios at each rank

D_1	D_2	D_3	
$\frac{\lvert Clicks_1 \rvert}{\lvert Q \rvert}$	$\frac{\lvert Clicks_2 \rvert}{\lvert Q \rvert}$	$\frac{\lvert Clicks_3 \rvert}{\lvert Q \rvert}$	...

❸ Fit ratios to equation

$$p(click@k) = \varphi^k$$

Fitted value yields patience parameter for clicking on result summary link at rank *k*

For each rank, we could then compute the ratio of clicks over each query, as demonstrated at step ❷. This yielded a decreasing ratio with increasing depths, demonstrating that

[18] This is discussed in further detail later in Section 9.2.1 on page 311.

Table 6.3 Summary table of the twelve stopping strategies, along with each of the threshold parameter values trialled. Note that for SS5-COMB and SS11-COMB, thresholds from different stopping strategies are used for the respective components of each combination strategy.

Stopping Strategy	Threshold Parameter Values
SS1-FIX	x_1 = [1, 2, 3, ..., 8, 9, 10, 15, 18, 21, 24]
SS2-NT	x_2 = [1, 2, 3, ..., 8, 9, 10, 15, 18, 21, 24]
SS3-NC	x_3 = [1, 2, 3, ..., 8, 9, 10, 15, 18, 21, 24]
SS4-SAT	x_4 = [1, 2, 3, ..., 8, 9, 10]
SS5-COMB	x_2 = [1, 2, 3, ..., 8, 9, 10, 15, 18, 21, 24] **(SS2-NT)** x_4 = [1, 2, 3, ..., 8, 9, 10] **(SS4-SAT)**
SS6-DT	x_6 = [0.0, 0.05, 0.10, ..., 0.90, 0.95, 1.00]
SS7-DKL	x_7 = [3.0, 3.5, 4.0, ..., 7.0, 7.5, 8.0]
SS8-IFT	x_8 = [0.002, 0.004, 0.006, ..., 0.026, 0.028, 0.03] y_8 = 2
SS9-TIME	x_9 = [30, 60, 90, 120, 150]
SS10-RELTIME	x_{10} = [10, 20, 30, 40, 50]
SS11-COMB	x_4 = [1, 2, 3, ..., 8, 9, 10] **(SS4-SAT)** x_{10} = [10, 20, 30, 40, 50] **(SS10-RELTIME)**
SS12-RBP	x_{12} = [0.80, 0.85, 0.90, 0.9087, 0.95, 0.99]

searchers were less likely to click on results further down the rankings. Finally, $p(click@k) = \phi^k$ was fit to the data. This represents the probability of clicking on a result summary at rank k, with ϕ denoting the RBP patience parameter. When fit, $\phi = 0.9087$.

Table 6.3 on page 178 lists each of the stopping strategies, along with the different threshold parameter values that were trialled for each.

6.4.2.7 Simulated Searcher Constraints and Goals

Like in the corresponding user studies, we imposed different constraints upon the simulated searchers to keep comparisons as fair as possible. These constraints and goals are discussed in depth in the relevant chapter. Refer to Section 7.3.1.2 on page 216 and Section 8.3.1.2 on page 282 for further information for the two studies.

6.4.3 Simulation Runs and Evaluation

Having now discussed how all of the various components of the CSM and the **SimIIR** framework were instantiated for our simulations, we now move onto a discussion of how we actually ran the simulations – and evaluated them.

With two high-level research questions focusing on the empirical work, we designed and executed a set of simulation runs to address both

- **HL-RQ3a** Given the aforementioned operationalised stopping strategies, how well does each one perform?

  To address this research question, we propose a series of **performance runs** that allow us to determine the best overall level of performance that can be attained using a particular configuration of a simulated searcher, via a number of *what-if* scenarios.

- **HL-RQ3b** How closely do the operationalised stopping strategies compare to the actual stopping behaviours of real-world searchers?

  To address this second research question, we also propose a series of **comparison runs** that instead focus upon how closely different configurations of simulated searcher approximate the stopping behaviours of real-world searchers.

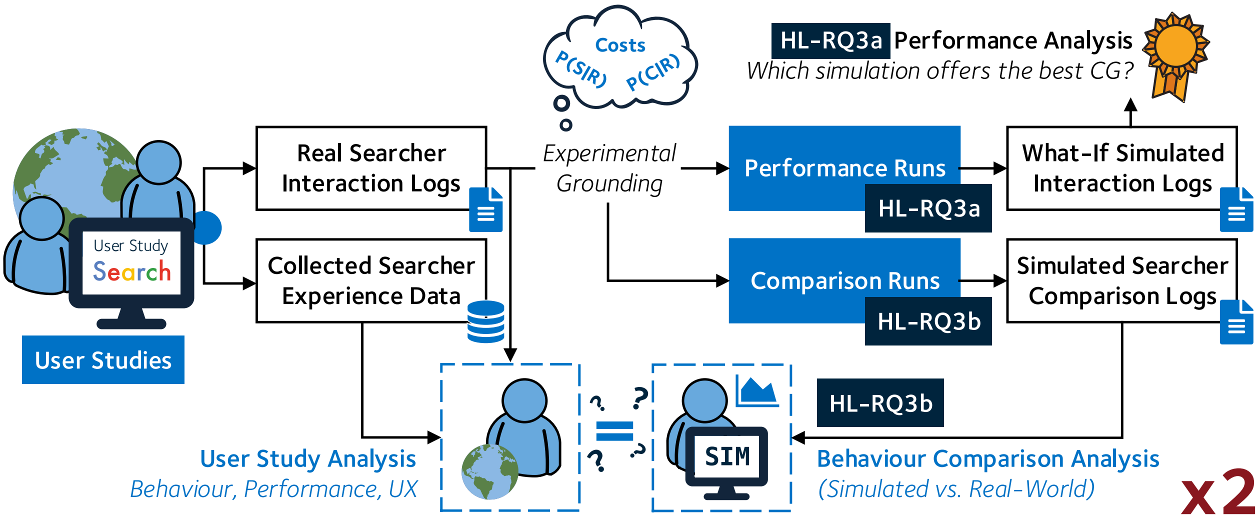
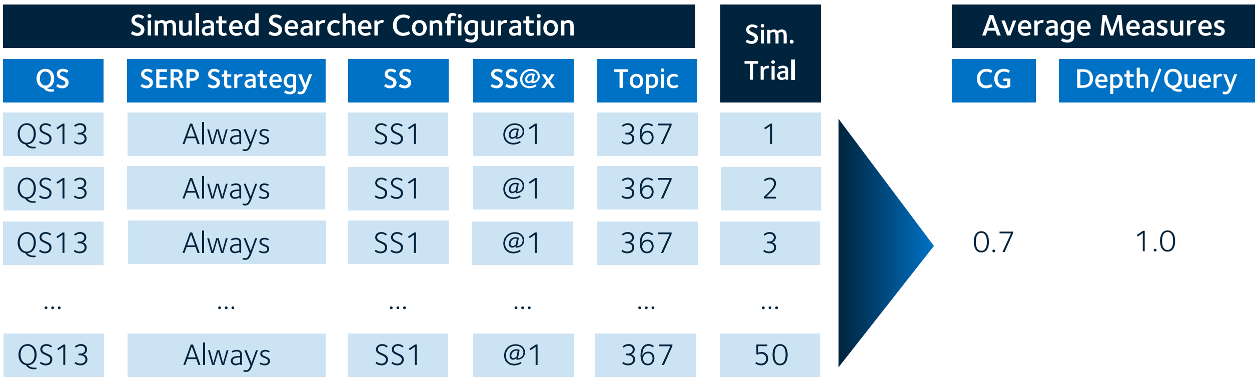

Costs
P(SIR)
P(CIR)
HL-RQ3a Performance Analysis
Which simulation offers the best CG?
User Study
Search
User Studies
Real Searcher Interaction Logs
Collected Searcher Experience Data
Experimental Grounding
Performance Runs
HL-RQ3a
Comparison Runs
HL-RQ3b
What-If Simulated Interaction Logs
Simulated Searcher Comparison Logs
HL-RQ3b
SIM
User Study Analysis
Behaviour, Performance, UX
Behaviour Comparison Analysis
(Simulated vs. Real-World)
x2

Figure 6.9 How the two sets of simulations, represented as blue boxes, fit within the wider experimentation framework as discussed in this chapter. The illustration also shows what components address the two high level research questions, HL-RQ3a and HL-RQ3b.

Simulations: Within the Methodology These simulations of interaction fit within the wider experimental framework discussed in this chapter, illustrated in Figure 6.9. Within the figure, we can see the link between the user studies and the two sets of simulations (highlighted with blue boxes) via the act of *grounding*. The illustration also provides linkage between the simulations, and the two sets of analyses that are undertaken – the performance analysis, addressing HL-RQ3a, and the behaviour comparison analysis that addresses HL-RQ3b. The performance analysis is an examination of the hundreds of different possible simulation configurations, allowing us to explore how performance varies through *what-if* simulations. In all, this process is **repeated twice**, once per user study, as shown in the illustration. We discuss the two different sets of simulations that address research questions HL-RQ3a and HL-RQ3b in Sections 6.4.3.1 and 6.4.3.2 respectively.

6.4.3.1 Performance Runs

Named as a series of *what-if* simulations above, the performance runs instantiate the different components of the CSM and **SimIIR** framework as previously discussed throughout Section 6.4.2. Using the grounded interaction probabilities and costs, these simulations were



Simulated Searcher Configuration					Sim. Trial		Average Measures	
QS	SERP Strategy	SS	SS@x	Topic			CG	Depth/Query
QS13	Always	SS1	@1	367	1			
QS13	Always	SS1	@1	367	2			
QS13	Always	SS1	@1	367	3		0.7	1.0
...	...	...	...	...	...			
QS13	Always	SS1	@1	367	50			

Figure 6.10 An example set of simulation results (each represented on a different row), with each row representing the same configuration, over the 50 different trials. Performance values can then be extracted from each trial, with a mean computed. In this example illustration, the mean CG and depth per query are shown.

trialled over the five selected topics of the TREC 2005 Robust Track (Voorhees, 2006) (as discussed in Section 6.1.3), with queries generated via the querying strategy outlined in Section 6.4.2.2. Altogether, this provided us with a wealth of simulated interaction data from which we could calculate a series of averages over the different trials run. As illustrated in Figure 6.10, we then computed the various performance measures over each simulated searcher configuration, taking an average over each of the five topics.

Figure 6.10 illustrates a simple example configuration of a simulated searcher, using: querying strategy **QS13**; the **SERP Always** (baseline) SERP examination strategy; result summary level stopping strategy **SS1-FIX @1**; and TREC topic № 367. The configuration was also run over 50 separate trials. All performance runs were examined over the same five topics as outlined previously in Section 6.1.3. Performing 50 trials for each individual configuration allowed us to account sufficiently for the variability that would be presented across runs, with further detailed presented in Section 6.4.2.3.

Given the name performance runs, we examined the performance of each simulated searcher trialled. While examining the performance of queries (via the measures outlined previously in Section 6.3.3), we also examined measures illustrated in Figure 6.11: mean levels of CG, and the mean **depth per query (DQ)**.

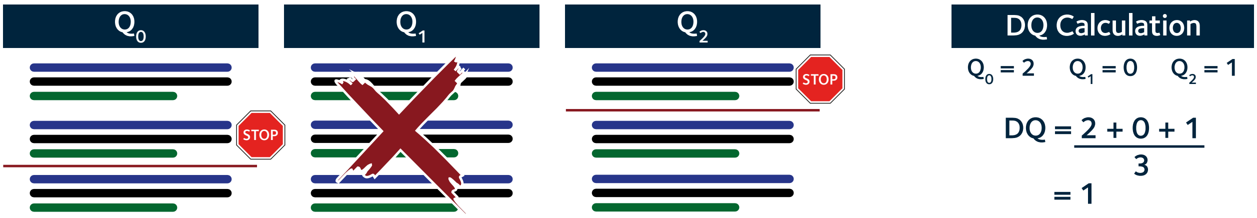


Figure 6.11 An example illustrating how the *mean depth per query* is computed across a search session. In this example, three individual queries are issued, with no results examined for $\mathbf{Q_1}$.

CG was discussed in passing in Section 6.4.2.4. For our simulations, we consider the CG as the amount of gain accrued over the *course of a search session* – which, by definition, can entail more than a single query. A more robust series of stopping strategies that are better at stopping a simulated searcher examining poor quality SERPs to great depths will provide higher levels of CG, but only if the queries issued offer good performance. Similarly, an effective SERP level stopping decision point implementation will stop the searcher from examining a poor SERP in the first instance, leaving more time to examine SERPs that could potentially offer higher quality results.

The other major measure used in our performance measures was the depth per query. With this measure, performance is not measured, but rather the stopping behaviour of the simulated searchers. As shown in Figure 6.11, a fictional search session consists of three queries.

In the example illustration, a simulated searcher examines to a depth of 2 for $\mathbf{Q_0}$, and a depth of 1 for $\mathbf{Q_2}$. The searcher does not even enter the associated SERP for $\mathbf{Q_1}$, as the SERP level stopping decision point prevents the searcher from examining result summaries in detail. The resultant *DQ* for the search session is therefore 1.

6.4.3.2 Comparison Runs

Rather than focus upon the overall performance attained by simulated searchers under different scenarios, the second set of simulations we ran focused on comparing simulated searcher behaviours against their real-world counterparts.

The simulations for this set were configured much like the performance set – save for the difference between the querying component. Rather than considering queries generated by a querying strategy, we instead took the queries from the associated user study and issued each one in turn. In effect, we replayed all real-world queries issued (as discussed in Section 6.4.2.2).

Real-world queries were extracted and grouped by the experimental interface or condition in which it was executed. In order to compute per-session measures (e.g. CG), we could then take the queries belonging to a particular subject and topic combination and group them together, summing or averaging measures where appropriate. With only four topics trialled during the user studies, we considered only the four TREC topics, omitting the practice topic (№ 367). This was due to the fact that we only had real-world query data for the four topics trialled in the user studies. We once again ran a simulated searcher for each different configuration, over every query issued. A total of 50 trials were once again used.

To perform our comparisons between the real-world and simulated searchers, we used the Mean Squared Error (MSE) to compute the difference between the two. For this, our calculations were performed by examining the click depth of the real-world searchers over each query, and taking the simulated click depths. Simulated click depths are defined as the depth of the last document that was considered attractive enough to examine on a given simulated SERP. Considering each configuration of simulated searcher (i.e. considering the different ways in which **SimIIR** components were instantiated), we could then produce a table of click depths, as provided in the example below.

Query	Topic	Real-World	Sim. 1	Sim. 2	Sim. 3	...	MSE
Q_0	408	5	3	5	2	...	4
Q_1	408	1	2	1	2	...	0
Q_2	408	7	1	4	3	...	16
...	...	...	...	...	...	...	...

In the above example, **Sim. x** represents the mean value of a particular simulated searcher configuration, with the mean taken over the 50 simulated trials. For each query, the real-world click depth is shown, along with the simulated click depth from each simulated searcher trialled. Highlighted cells show what is being compared on each row – for instance, for query **Q**$_0$, the real-world click depth of 5 is compared against the **Sim. 1** click depth of 3. Considering the MSE value between the two, using the following formula:

$$MSE = (\theta - \hat{\theta})^2, \qquad \text{Equation 6.3}$$

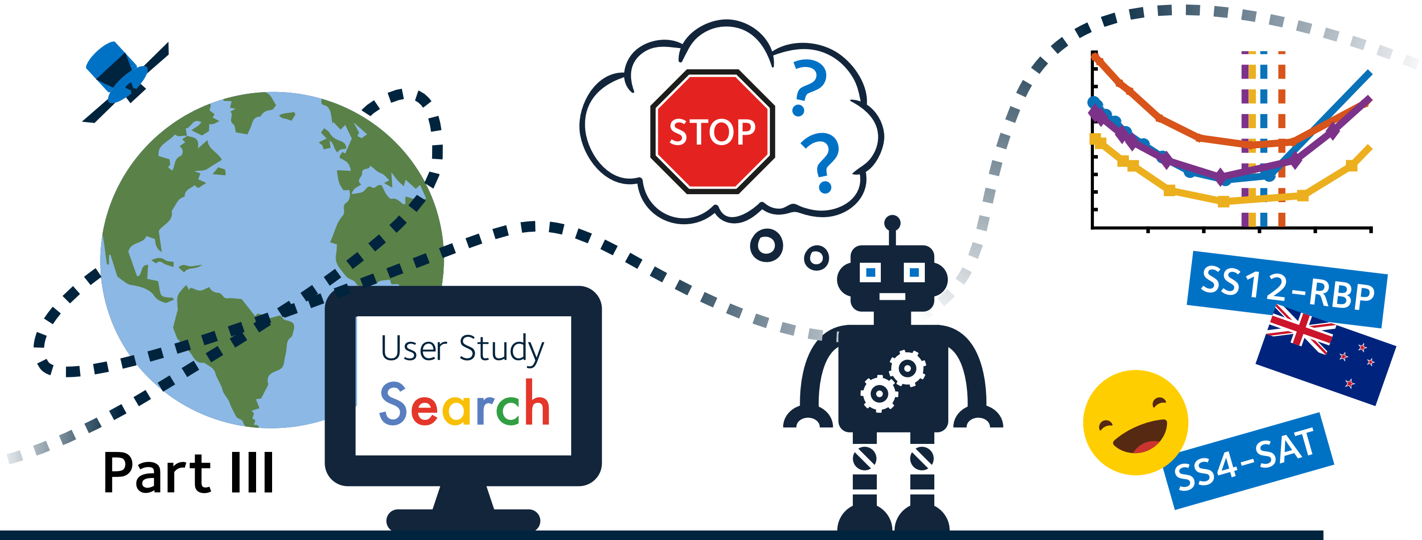
where θ denotes the real-world click depth, and $\hat{\theta}$ denotes the click depth approximation, we arrive at a MSE value of 4. The closer the MSE value is to 0, the better the approximation given. In the above example, the compared values for **Q**$_1$ therefore offer the best approximation of the actual stopping depth of the searcher. After each MSE value had been calculated, this could then be used to plot against the mean depth per query across a variety of different stopping strategies. For example, recall that SS1-FIX considers the stopping depth across a range of threshold parameter values (x_1), with the parameter denoting the stopping depth. The higher this value, the greater the depth per query that will be attained. By computing the MSE at each point, we were able to determine which stopping threshold offered the best approximation of stopping behaviours, across a range of mean depths per query, for that particular stopping strategy.

6.5 Chapter Summary

In this chapter, we have outlined the general methodology that is used throughout the remaining chapters of this thesis. As we report on two separate user studies in Chapters 7 and 8, this chapter provides an overview of the common approaches followed, with unique aspects pertaining to the individual user studies discussed in the relevant chapter.

In order to tackle the high-level research questions of this thesis, our general methodology was to first undertake a user study that captured a variety of different behavioural, performance and user experience measures, as discussed in Section 6.2. The data derived from this user study was then used to ground a series of complex simulations of interaction, attempting to mimic the behaviours exhibited by the real-world user study subjects. After discussing how we instantiated each of the different components of the CSM and **SimIIR** framework, we then concluded the chapter with a discussion on the two sets of simulation runs trialled, allowing us to address research questions HL-RQ3a and HL-RQ3b.

With the conclusion of this chapter, all necessary groundwork has been laid to present the results of our user studies and simulations of interaction, which we present in Part III.



Part III

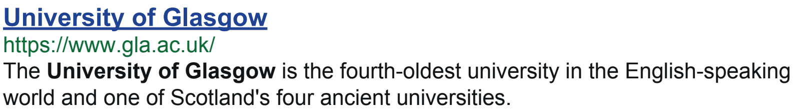
Examining and Simulating Searcher Stopping Behaviours

In this part of the thesis, we present our ***empirical contributions****, examining what happens to searcher stopping behaviours under* ***different search contexts****. We then report on a number of different* ***simulated analyses****, examining our proposed stopping strategies. Finally, we report on an empirical analysis of the new* ***SERP level stopping decision point****.*

Chapter 7

Result Summary Lengths and Stopping Behaviour

The SERP is core to a searcher's experience when using a retrieval system. The presentation and design of the SERP has over the years been subject to much research. Today, more complex components (such as the *information card* (Navalpakkam et al., 2013) or *social annotations* (Muralidharan et al., 2012)) are now becoming commonplace in web retrieval systems. Despite these advancements, much work still remains on examining how more traditional SERP components are designed and presented to searchers. As we will focus on in this chapter, *result summaries* are such a component.



University of Glasgow
https://www.gla.ac.uk/
The **University of Glasgow** is the fourth-oldest university in the English-speaking world and one of Scotland's four ancient universities.

As shown in the above example, result summaries have been traditionally viewed as the *ten blue links*, each with their corresponding title and source (typically a URL) of the associated document. Included with these two components are the textual *snippets* of *keywords-in-context*, derived from the document itself. These snippets are approximately 130-150 characters (or two lines) in length (Hearst, 2009). Researchers have explored result summaries in a variety of different ways, such as: examining their length (Paek et al., 2004; Cutrell and Guan, 2007; Kaisser et al., 2008); the use of thumbnails (Woodruff et al., 2002; Teevan et al., 2009); their attractiveness (Clarke et al., 2007; He et al., 2012); and the generation of *query-biased snippets* (Tombros and Sanderson, 1998; Rose et al., 2007).

In this chapter, we are interested in examining how the length (and subsequently information content) of result summaries affects SERP interactions – specifically examining their stopping behaviours – and a searcher's ability to select relevant over non-relevant items. This is in tandem with an examination of different stopping strategies (outlined in Chapter 5), and how they adapt to increasing result summary lengths. Prior research has demonstrated that longer result summaries tend to lower completion times for informational tasks, where searchers need to find only a single relevant document (Cutrell and Guan, 2007). However, does this finding hold in an ad-hoc context, where searchers need to find *several* relevant items? Furthermore, how does the length and information associated with longer result summaries affect the searcher's ability to discern the relevant from the non-relevant? We address these questions from the perspective of both:

- a **user study** examining this phenomenon, presented in Section 7.2; and
- a **simulated analysis**, closely examining how varying snippet lengths affect searcher performance and stopping behaviours, discussed in Section 7.3.

The outline for both of these studies follows the general methodology, as discussed in Chapter 6. Before discussing the studies and their results, we begin with an overview of prior work that has examined the length of SERPs and result summary snippets.

7.1 Background

Researchers have examined various aspects of SERPs, and how the designs of such aspects influence the behaviour of searchers. In this section, we provide a summary of the various aspects that have been examined over time. Specifically, we consider:

- the size of SERPs;
- how much text should be presented within each result summary;
- the layout and presentation of SERPs; and
- how snippet text for result summaries is generated.

Of the four areas of SERP research that we examine in this section, we consider the latter to be the main focus of this work. Each area is summarised below.

7.1.1 Results per Page

Today, a multitude of devices are capable of accessing the WWW – all utilising a wide range of different screen resolutions and aspect ratios. Therefore, the question of how many result summaries should be displayed per page becomes hugely important, yet increasingly difficult to answer. Examining behavioural effects of mobile devices when interacting with SERPs has attracted much recent research (e.g. Kim et al. (2012, 2014, 2016)), and with each device capable of displaying a different number of results *above-the-fold*[1], research has shown that the number of results presented on a SERP can influence the behaviour of searchers (Joachims et al., 2005; Kim et al., 2014). Understanding this change in behaviour can help guide and inform individuals charged with the design of user interfaces in contemporary retrieval systems.

However, Linden (2006) stated in a Google industry report that searchers desired more than 10 results per page, despite the fact that increasing the number of results displayed yielded a 20% drop in traffic. It was hypothesised that this was due to the extra time required to dispatch the longer SERPs. However, this drop could have been attributed to other reasons. Oulasvirta et al. (2009) discussed the *paradox of choice* (Schwartz, 2005) in the context

[1] Refer to Section 9.2.1 on page 311 for a detailed explanation on displaying results *above-the-fold* – also called the *viewport size*.

of search, where more options (results) – particularly if highly relevant – will lead to poorer decisions, degrading searcher satisfaction. In terms of searcher satisfaction, it can be argued that modern retrieval systems may be a victim of their own success, leaving searchers with *choice overload*. Oulasvirta et al. (2009) found that presenting searchers with a six item result list (rather than a list of 24) was associated with higher degrees of searcher satisfaction, confidence with choices and perceived carefulness.

Kelly and Azzopardi (2015) broadly agreed with the findings of Oulasvirta et al. (2009). Here, the authors conducted a between-subjects study with three conditions, where subjects were assigned to one of three interfaces – a baseline interface, showing 10 results per page (the traditional *ten blue links*), and two interfaces displaying 3 and 6 results per page respectively. Their findings showed that individuals using the 3 and 6 results per page interfaces spent more time (significantly so) examining top ranked results. They were also more likely to click on documents ranked higher than those using the 10 results per page interface. Findings from this study also suggested that subjects using the interfaces showing fewer results per page found it comparatively easier to find relevant content than those using the 10 results per page interface. Of course, examining results to shallow depths also means that searchers would have stopped examining content comparatively early, too. Displaying 10 results per page is still considered as the *de-facto* standard (Hearst, 2009), with this *de-facto* value our primary interest in examining result summary lengths in more detail.

7.1.2 Snippet Lengths: Longer or Shorter?

Snippet lengths have been examined in a variety of ways. A user study by Paek et al. (2004) compared a searcher's preference and usability against three different interfaces for displaying result summaries. With question answering tasks, the interfaces:

- displayed a *normal* SERP, consisting of a two line snippet for reach result summary, complete with a clickable hyperlink to the corresponding document;

- an *instant* interface, where an expanded snippet was displayed upon clicking it; and
- a *dynamic* interface, where hovering the cursor would trigger the expanded snippet.

The instant view interface was shown to allow searchers to complete the given tasks in less time than the normal baseline, with half of the participants preferring this approach.

Seminal work by Cutrell and Guan (2007) explored the effect of different snippet lengths, exploring *short* (1 line), *medium* (2-3 lines, the expected standard) and *long* (6-7 lines) snippets. They found longer snippets significantly improved performance for informational tasks (e.g. `find the address for Glasgow International Airport`[2]). Their subjects performed better for informational queries as snippet length increased. This work was extended by Kaisser et al. (2008). They conducted two experiments that estimated the preferred snippet length according to answer type (e.g. finding a person, time, or place), and comparing the results of the preferred snippet lengths to searchers' preferences, in order to see if this could be predicted. Their preferred snippet length was shown to depend upon the type of answer expected, with greater searcher satisfaction shown for the snippet length predicted by their technique. The findings also indicated that longer snippets may be more useful if the relevance of the snippet to the query was considered.

More recent work has begun to examine what snippet sizes are appropriate for mobile devices, with the multitude of screen resolutions available. Given smaller screen sizes when compared to desktop or laptop computers, this is particularly important – snippet text considered acceptable on a computer screen may involve considerable scrolling/swiping on a smaller screen. Kim et al. (2017) found that subjects using longer snippets on mobile devices exhibited longer search times and didn't lead to improvements in correctly identifying relevant content under informational tasks.[3] Longer reading times and frequent scrolling/swiping (with more viewport movement) were exhibited. Therefore, longer snippets did not

[2] Formerly *Abbotsinch Airport* and used as an airfield during World War II, *Glasgow International Airport* is located eight miles west of Glasgow city centre.

[3] The tasks considered by Kim et al. (2017) were similar to those defined by Cutrell and Guan (2007), where a single relevant document was sought.

appear to be very useful on a small screen. An *instant* or *dynamic* approach (as per Paek et al. (2004)) may have practical applications if searching were to be conducted on a mobile.

7.1.3 SERP Layout and Presentation

Early works regarding the presentation of result summaries examined different approaches to automatically categorising result summaries for searchers, similar to the categorisation approach employed by early retrieval systems (as shown in Figure 2.1 on page 24). Chen and Dumais (2000) developed an experimental system that automatically categorised result summaries on-the-fly as they were generated. For a query, associated categories were then listed as verticals, with associated document titles provided underneath each category header. Traditional result summaries were then made available when hovering over a document title (as illustrated below with a sample title and summary popup). Subjects of a user study found the interface easier to use than the traditional *ten blue links* approach – they were 50% faster at finding information displayed in categories. This work was then extended by Dumais et al. (2001), where they explored the use of hover text to present additional details about search results based upon user interaction. Searching was also found to be slower with hover text, perhaps due to the fact that searchers were required to make decisions as to when to explicitly seek additional information.

Chinese expert says South China tiger may be extinct in wild

...be too few to save the species from extinction, a Chinese wildlife expert said Thursday. The last... in May 1996, a report by the World Wildlife Fund for Nature said fewer than 50...
Associated Press Worldwide News Service

Popup Summary
Mouse Hover Event

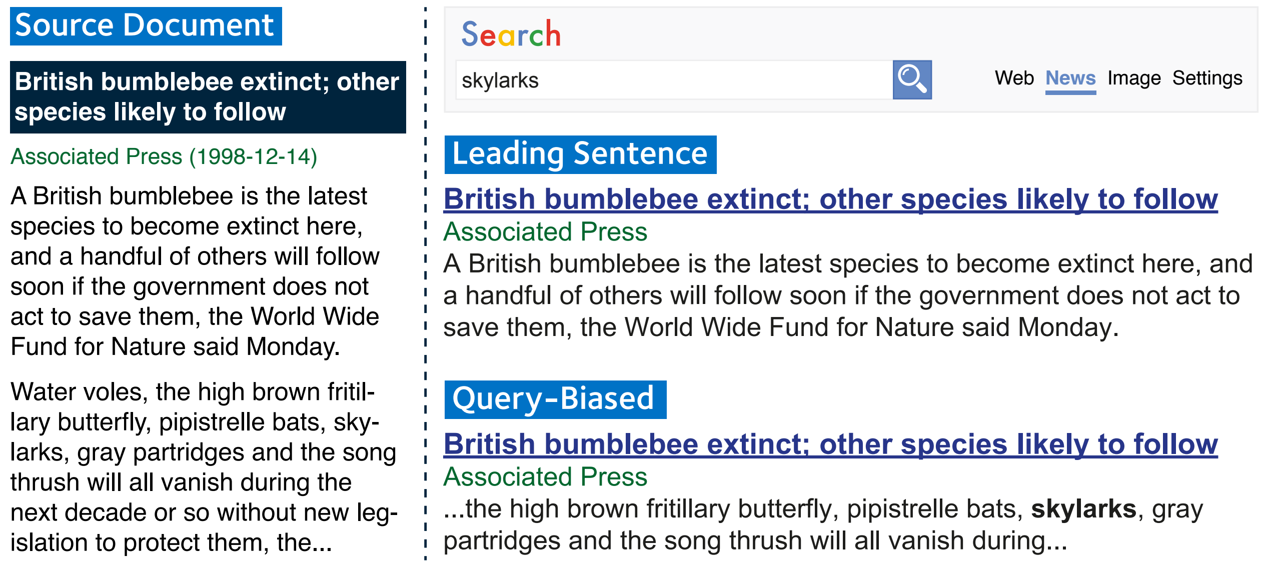
Alternatives to the traditional, linear list of result summaries have also been trialled – like grid-based layouts (Resnick et al., 2001; Kammerer and Gerjets, 2010; Chierichetti et al., 2011). From these examples, Kammerer and Gerjets (2010) examined differences in searcher behaviour when interacting with a standard list interface, compared against a tabular interface (title, snippet and source stacked horizontally in three columns for each result), and a grid-based layout (result summaries placed in three columns). Those using the grid layout spent more time examining result summaries, and demonstrated promise in overcoming issues such as *positional bias* (Craswell et al., 2008), as observed by Joachims et al. (2005).

Marcos et al. (2015) also performed an eye-tracking analysis, examining the effect of searcher behaviours while interacting with SERPs – and whether the *richness* of result summaries provided on a SERP (i.e. result summaries enriched with metadata from corresponding pages) impacted upon the user's search experience. Enriched summaries were found to help capture a searcher's attention.

Including both textual and visual information within results could have a positive effect on the assessment of relevance and the formulation of queries (Joho and Jose, 2006). Enriched summaries were also examined by Ali et al. (2009) in the context of navigational tasks. Striking a good balance between textual and visual cues (i.e. *proximal cues,* as discussed in Section 3.3.1.1) has been shown to improve a searcher's ability to complete tasks, and reduce search completion time.

7.1.4 Generating Snippet Text

Searchers may gain insight to the relevance of documents by examining the associated result summaries (He et al., 2012). Consequently, research has been undertaken that examined different kinds of snippets, and the optimal length of a snippet. Work initially focused upon how these summaries should be generated (Pedersen et al., 1991; Landauer et al., 1993; Tombros and Sanderson, 1998; White et al., 2003; Leal-Bando et al., 2015). These early works proposed the idea of summarising documents with respect to the query (query-biased summaries), or keywords-in-context – as opposed to simply extracting the representative or lead sentences from the document (Kupiec et al., 1995). Examples of both approaches are illustrated in Figure 7.1. Indeed, Tombros and Sanderson (1998) showed that subjects of their study were likely to identify relevant documents more accurately when using query-biased



British bumblebee extinct; other species likely to follow

Associated Press (1998-12-14)

A British bumblebee is the latest species to become extinct here, and a handful of others will follow soon if the government does not act to save them, the World Wide Fund for Nature said Monday.

Water voles, the high brown fritillary butterfly, pipistrelle bats, skylarks, gray partridges and the song thrush will all vanish during the next decade or so without new legislation to protect them, the...

British bumblebee extinct; other species likely to follow

Associated Press

A British bumblebee is the latest species to become extinct here, and a handful of others will follow soon if the government does not act to save them, the World Wide Fund for Nature said Monday.

British bumblebee extinct; other species likely to follow

Associated Press

...the high brown fritillary butterfly, pipistrelle bats, **skylarks**, gray partridges and the song thrush will all vanish during...

Figure 7.1 A visual example of two different types of summary, along with a portion of an example document from the TREC AQUAINT collection. Given the query `skylarks`, the Search result summaries for both leading sentence and query-biased summaries are shown. Note the highlighting of the term **skylarks** in the query-biased summary.

summaries, compared to summaries that were simply generated from the first few sentences of a given document. Query-biased summaries have also been more recently shown to be preferred on mobile devices (Spirin et al., 2016).

When constructing snippets using query-biased summaries, Rose et al. (2007) found that a user's perceptions of the result's quality were influenced by the snippets. If snippets contained truncated sentences or many fragmented sentences (denoted as *text choppiness*), searchers perceived the quality of the results more negatively, regardless of length. Kanungo and Orr (2009) found that poor readability also impacted upon how the resultant snippets were perceived. They maintain that readability is a crucial presentation attribute that needs to be considered when generating a query-biased summary. Clarke et al. (2007) analysed thousands of pairs of snippets where result *A* appeared before result *B*, but result *B* received more clicks than result *A*. As an example, they found results with snippets which were very short (or missing entirely) had fewer query terms, were not as readable and attracted fewer clicks. This led to the formulation of several heuristics relating to document surrogate features, designed to emphasise the relationship between the associated page and generated snippet. Heuristics included:

- ensuring that all query terms appeared in the generated snippet (where possible);
- withholding the repetition of query terms in the snippet if they were present in the page's title; and
- displaying shortened, easily readable URLs.

Recent work has examined the generation of snippets from more complex angles – from manipulating underlying indexes (Turpin et al., 2007; Bast and Celikik, 2014), to language modelling (Li and Chen, 2010; He et al., 2012), as well as using a searcher's recorded history to improve the generation of snippets (Ageev et al., 2013; Savenkov et al., 2011). The previous generation approaches also may not consider what parts of a document searchers actually find useful. Ageev et al. (2013) incorporated into a new model of post-click searcher behaviour data, such as mouse cursor movements and scrolling over documents, producing *behaviour-based snippets.* Results showed a marked improvement over a strong text-based snippet generation baseline. Temporal aspects have also been considered – Svore et al. (2012) conducted a user study that showed searchers preferred snippet text with *trending* content in snippets when searching for trending queries, but not so for general queries.

7.2 Varying Snippet Lengths

As can be seen from the background to this chapter, the presentation of result summaries has been demonstrated to strongly influence the ability of a searcher to judge relevance (He et al., 2012). Relevant documents may be overlooked due to uninformative or unattractive summaries – but conversely, non-relevant documents may be examined due to a misleading summary. However, longer summaries also increase the cost of the examination, so there is likely a tradeoff between informativeness/accuracy and length/cost. The current, widely accepted standard for result summaries are *two query-biased snippets/lines* (Hearst, 2009). However, does this concept hold in an ad-hoc context? In this context, do searchers – when presented with longer result summaries – gain an improved discrimination between relevant and non-relevant result summaries?

To address these questions, we now discuss a user study that investigated the effects of search behaviour (in particular, stopping behaviours) and search performance when we varied the lengths of result summaries, and thus the information content therein. The user study reported is crowdsourced ($n = 53$) and follows a within-subjects design. Under ad-hoc topic retrieval, subjects used four different search interfaces, each with a different size of result summary. Findings from this study allowed us to address two main research questions, which we enumerate below.

- **SNIPPET-RQ1** How does the length of a result summary affect searcher stopping behaviours, performance and user experience?
- **SNIPPET-RQ2** Does the length of each result summary affect a searcher's decision-making ability and their likelihood of identifying relevant documents?

If longer result summaries affect searcher judgements regarding their relevance, it follows that this will also influence their stopping behaviour, where searchers would stop at comparatively shallower depths (as per IFT and SET). This is because we assume that a greater volume of text will require a longer period of examination (a greater examination cost). Given sets of short and long result summaries, a searcher, over an identical time period, will therefore examine fewer long result summaries than when examining the shorter summaries. Refer to Section 3.4.3 on page 102 for more detail on the cost-interaction hypothesis (Azzopardi, 2011).

This explanation provides a justification for **SNIPPET-RQ2**. Corroborating evidence to support the research question would be likely, as with longer result summaries comes a greater volume of information gain, as we demonstrate in Section 7.2.1.1. We hypothesised that longer and more informative result summaries would enable participants to make better quality decisions, due to a higher information content within a greater volume of text. In the remainder of this section, we:

- discuss study-specific details to complement the general methodology outlined in Section 6.2;
- provide results and analysis from the study, providing insight into the two study-specific research questions outlined above; and
- discuss the implications of the study.

We then take the interaction data from this study forward to Section 7.3, using it as a means of grounding an extensive set of simulations of interaction. These are used to examine in greater depth how snippet length affects searcher stopping behaviours using our operationalised result summary level stopping strategies.

7.2.1 Methodology

In this section, we outline the user study's methodology. This section provides study-specific, supplementary details that complement the general user study methodology. For each subsection discussed below, we refer back to the relevant section in the general methodology to assist in understanding how each of the different components interact.

Below, we discuss the different search interfaces that we trialled, along with how we generated result summary snippets of varying length. We then provide a brief discussion of the 53 subjects who took part in the user study, explain the search task, and discuss the post-task surveys that subjects completed.

7.2.1.1 Search System and Interfaces

In conjunction with the common retrieval system, corpus and topics discussed earlier in Section 6.1.3, we trialled four different search interfaces as part of the within-subjects study design. The four interfaces presented snippets, as part of result summaries, of varying lengths. This allowed us to explore the influence of snippet length and snippet informativeness on search behaviours, performance and user experience.

Snippet Fragments

Example below

...fragment one...fragment two...fragment three...a summary can be comprised of many fragments.

To decide the length and informativeness of the result summaries, we performed a preliminary analysis to determine the average length (in words) and informativeness (as calculated by *Kullback–Leibler Divergence* (Kullback and Leibler, 1951) (Kullback and Leibler, 1951) to measure *information gain*, or *relative entropy*) of result summaries with the title, and a varying number of *snippet fragments*[4] (from 0 to 10). The closer the entropy value is towards zero, the more information that is gained. Figure 7.2 plots the number of words, the *log* of the information gain, and the information gain attained per word.[5][6] It is clear from the plots shown in Figure 7.2 that a higher level of information gain was present in longer snippets. However, as the length increased with each snippet fragment added, the informativeness per word flatlined, with each fragment added offering diminishing returns. Consequently, we selected the four different interface conditions in the region where informativeness has the highest change, i.e. from zero to four. The conditions selected[7] for this study were therefore:

[4] Figure 2.9 on page 50 illustrates snippet fragments in the wider context of a SERP.

[5] These values were obtained by submitting over 300 queries from a previous user study, conducted by Azzopardi et al. (2013). These were conducted on similar topics, the same retrieval system and the same corpus as used in the study reported in this chapter.

[6] Why did we take the *log* of the information gain? This was primarily done to show a visual increase in Figure 7.2 and extrapolate differences. Negative values stem from the fact that information gain values are small, and the *log* of a small value (< 1.0) corresponds to a negative value.

[7] Figure 7.3 provides a complete, rendered example of the different result summaries in each condition, using the same document.

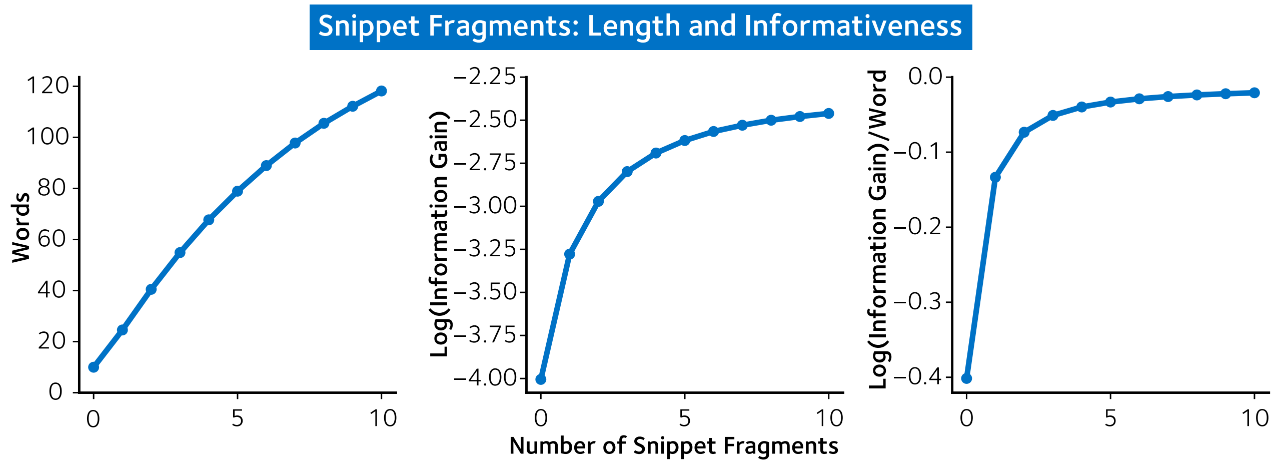


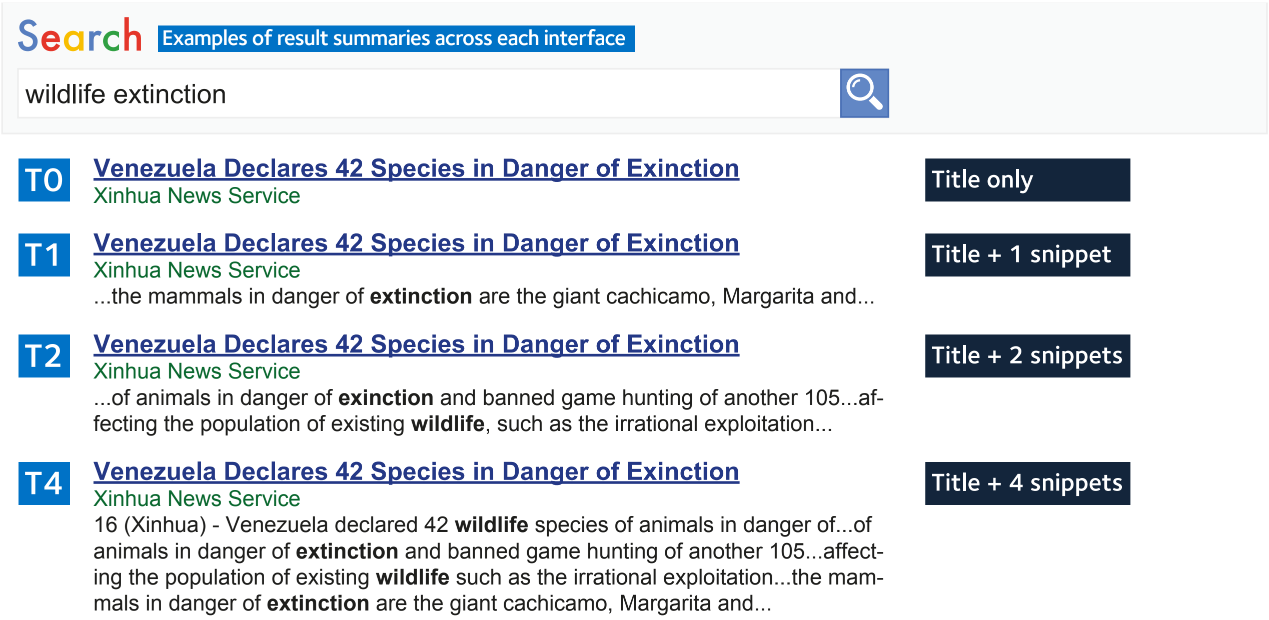
Figure 7.2 Plots showing the length (in words), informativeness (represented by the log of the information gain, or *IG*), and the log of the information gain *(IG)* per word for the title, plus 0 to 10 snippet fragments. The closer the value is to zero, the more information that is gained.

- T0, where only the title for each result summary was presented;
- T1, where for each result summary, a title and *one* query-biased snippet fragment were presented;
- T2, where a title and *two* query-biased snippet fragments were presented; and
- T4, where a title and *four* query-biased snippet fragments were presented.

From here, we carefully selected the order in which subjects were presented with each interface. For each of the four main topics discussed in Section 6.1.3, one of the four interfaces from T0, T1, T2 and T4 were assigned to a topic using a Latin-square rotation. For the practice topic, we used T2 – a title and two snippet fragments – the interface that represented the *de facto* standard for presenting result summaries (Hearst, 2009).

7.2.1.2 Snippet Generation

For interfaces T1 to T4, each result summary required one or more snippet fragments from the corresponding document. As illustrated in the complete, rendered example in Fig-



Search Examples of result summaries across each interface

wildlife extinction

T0 Venezuela Declares 42 Species in Danger of Exinction
Xinhua News Service
Title only

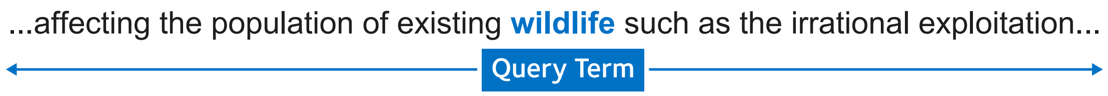
T1 Venezuela Declares 42 Species in Danger of Exinction
Xinhua News Service
...the mammals in danger of **extinction** are the giant cachicamo, Margarita and...
Title + 1 snippet

T2 Venezuela Declares 42 Species in Danger of Exinction
Xinhua News Service
...of animals in danger of **exinction** and banned game hunting of another 105...affecting the population of existing **wildlife**, such as the irrational exploitation...
Title + 2 snippets

T4 Venezuela Declares 42 Species in Danger of Exinction
Xinhua News Service
16 (Xinhua) - Venezuela declared 42 **wildlife** species of animals in danger of...of animals in danger of **extinction** and banned game hunting of another 105...affecting the population of existing **wildlife** such as the irrational exploitation...the mammals in danger of **extinction** are the giant cachicamo, Margarita and...
Title + 4 snippets

Figure 7.3 Examples of the result summaries generated by each of the interfaces, T0, T1, T2 and T4. The same document is used. Demonstrated by Search, each of the result summaries consists of: a **title** (in blue, underlined); none, one, or more **snippet fragments** (in black, with fragments separated by ellipsis); and a **newswire source** (under the title, in green).

ure 7.3, each of the fragments generated was query-biased in nature (Tombros and Sanderson, 1998). Fragments were generated by splitting a given document into sentences (delimited by a period), and scoring each of the sentences according to BM25 (where $\beta = 0.75$). Fragments were then extracted from the ordered series of sentences by identifying query terms within said sentences. Fragments were created by creating a window of 40 characters from either side of the identified query term, as illustrated in the example figure below.



...affecting the population of existing **wildlife** such as the irrational exploitation...

The ordered set of fragments were then joined together, in order of relevance and separated by ellipses (with one only for T1, two for T2, and four for T4). These were combined together with the document's title and source newswire to form the complete result summary.

7.2.1.3 Search Task

As discussed previously in Section 6.2.1, subjects were grounded by instructing them to imagine that they were newspaper reporters. As such, they were required to gather (save) documents to write stories about each of the four topics for which they were asked to search. For this study, subjects were assigned ten minutes to each of the four search tasks. They and were specifically instructed to find and identify as many relevant documents as they could during this allotted time. As shown in Section 6.2.2, subjects interacted with the standard search interface. Documents were *saved* by subjects when they were considered to be relevant to the given TREC topic.

7.2.1.4 Crowdsourced Subjects

A total of 60 subjects took part in the MTurk platform. However, seven subjects were omitted due to quality control constraints imposed on the study, as outlined in Section 6.2.4 on page 147. This left 53 subjects who satisfied the conditions of the experiment. In all, of the 53 subjects who satisfied the criteria, 28 were male, with the remaining 25 female. The average age of the subjects was 33.8 years ($min = 22$; $max = 48$; $stdev = 7.0$). A total of 19 subjects reported possessing a bachelor's degree or higher, with all expressing a high degree of search literacy, and reportedly conducted at least five searches for information online per week.

With a total of 53 subjects considered in the results of this study, each searching over four topics, this meant a total of 212 search sessions being logged and available for analysis. Finally, we report results over a reduced time period of six minutes (360 seconds). This decision was taken as not all of the 53 subjects spent the full ten minutes searching, with these subjects skewing results. By reducing the time period that we considered, we ensured that we could extract meaningful data from the interaction logs, and guaranteed that subjects were interacting with the experimental system up until the cutoff point.

7.2.1.5 Post-Task Surveys

With the pre-task survey the same as that outlined in the general methodology (refer to Section 6.3), surveys for this study differed post-task and post-experiment. Here, we discuss the questions posed in each of the four post-task surveys.

A seven-point Likert scale was used for post-task surveys, similar to the pre-task surveys. Upon completion of each search task, the following statements were completed by each subject, selecting from *strongly disagree* to *strongly agree.*

- **Clarity** The result summaries presented were clear and concise.
- **Confidence** The result summaries presented increased my confidence in my decisions.
- **Informativeness** The result summaries presented were informative to me.
- **Relevance** The result summaries presented allowed me to judge the relevance of the associated document.
- **Readability** The result summaries presented were readable.
- **Size** The result summaries presented were of an appropriate size and length.

These questions allowed us to obtain quantitative data alluding to how each subject perceived the search interface with which they interacted, and allowed us to ascertain the subjects' evaluations of the differing result summary lengths.

7.2.1.6 Post-Experiment Survey

At the end of the experiment, subjects also undertook a post-experiment survey. Five questions were posed, this time asking subjects to select which one of the four different interfaces best reflected their opinion of the question asked.

- **Most Informative?** Of the four interfaces, what one yielded the most informative result summaries?
- **Least Helpful?** Of the four interfaces, which one provided the most unhelpful result summaries?
- **Easiest?** Which of the four interfaces provided result summaries that were easiest to understand?
- **Least Useful?** Of the four interfaces, which one provided the least useful result summaries?
- **Most Preferred?** Of the four interfaces, what interface did you prefer using the most?

These questions allowed us to determine which interface delivered the optimal outcome for searching when considering different criteria.

7.2.2 Results and Analysis

From the four aspects we highlighted in Section 6.3, we report results from the study across four main sections, including analysis of: *behavioural* measures (interactions); *time-based* measures; *performance* measures; and *user experience* (surveys).

Each measure was analysed over the four different search interfaces. To perform our analysis, *Analysis of Variance (ANOVA)* tests were used using the interfaces as factors; main effects were examined with $\alpha = 0.05$. The Bonferroni correction was used for post-hoc analysis to determine what interfaces offered significant differences.

To check whether the interfaces were sufficiently different with respect to snippet length and information gain, we performed an analysis of the result summaries that were presented to the subjects. Table 7.1 summarises, over each interface, the number of words and

Table 7.1 Characters, words and the log of the *Information Gain (IG)* across each of the four interfaces trialled. Significant differences were revealed, with follow-up tests showing that each interface was significantly different to others. An greater **IG** value denotes a higher level of IG.

	T0	T1	T2	T4
Words	6.58±0.01	25.21±0.06	44.29±0.10	77.06±0.13
Characters	37.37±0.05	103.29±0.17	168.36±0.23	284.78±0.31
IG	-6.35±0.01	-3.59±0.00	-3.00±0.00	-2.67±0.00
IG/Word	-1.17±0.00	-0.18±0.00	-0.08±0.00	-0.04±0.00

characters that result summaries contained on average. As expected, the table shows an increasing trend of words and characters as the number of snippet fragments were increased. Information gain (or relative entropy), as previously discussed, was calculated using KL-divergence (Kullback and Leibler, 1951).[8] A two-tailed Student's t-test (where $\alpha = 0.05$) showed that the differences between snippet length ($F(3,208 = 1.2x10^5, p < 0.001)$) and information gain ($F(3,208) = 2.6x10^5, p < 0.001)$) were significant. Follow up tests revealed that this was the case over all four interfaces, indicating that our conditions were indeed different over these dimensions. These findings provide some justification for our choices of the number of snippet fragments used with each interface. A diminishing increase in information gain after four snippet fragments suggested that there would not have been much point generating result summaries that were any longer.

7.2.2.1 Interaction Measures

Across the four experimental interfaces trialled, Table 7.2 presents the mean (± standard deviations) of:

- the number of queries issued (**#Queries**);

[8] For consistency with the pilot study reported in Figure 7.2, we again took the *log* of the information gain and reported them in Table 7.1.

Table 7.2 Various measures reported over each of the four experimental interfaces, **T0**, **T1**, **T2** and **T4**. Included are interaction and time-based measures (behavioural), as well as performance-based measures. No significant differences were observed, bar for the time per result summary, as highlighted. Refer to Section 7.2.2.2 for details.

		T0	T1	T2	T4
Interactions	**#Queries**	3.72±0.34	3.19±0.35	3.30±0.35	3.28±0.31
	#SERPs/Query	2.87±0.29	2.69±0.23	2.43±0.13	2.40±0.20
	#Docs./Query	4.23±0.55	4.83±0.54	5.14±0.66	4.76±0.62
	Depth/Query	15.44±1.81	17.00±2.21	14.37±1.39	13.53±1.95
Performance	**P@10**	0.25±0.02	0.23±0.02	0.27±0.02	0.25±0.03
	#Saved	6.68±0.66	7.00±0.63	6.49±0.58	7.60±0.79
	#TREC Saved (iP)	2.58±0.34	2.28±0.25	2.47±0.28	2.66±0.32
	#TREC Non.	1.85±0.32	2.08±0.29	1.98±0.24	1.68±0.32
Times	**Per Query**	8.29±0.57	7.99±0.57	9.42±0.79	8.12±0.48
	Per Document	17.32±2.12	22.82±6.03	17.19±1.86	18.99±2.13
	Per Summary	1.63±0.13	2.21±0.21	2.35±0.23	2.60±0.27

- the number of SERPs viewed per query (**#SERPs/Query**), considering pagination;
- the number of documents viewed per query (**#Docs./Query**); and
- the mean click depth – or stopping depth (**Depth/Query**).

These are all presented within the **Interactions** grouping. Across the four experimental interfaces of **T0**, **T1**, **T2** and **T4**, there were no significant differences reported between any of these measures. The number of queries issued follows a slight down-

Table 7.3 A summary of the various interaction probabilities over each of the four experimental interfaces examined. Note the increasing trends for each probability from T0 → T4. Section 6.4.2.3 on page 167 provides an explanation of the various probabilities listed here. No significant differences were observed across any of the probabilities and interfaces.

		T0	T1	T2	T4
Click	P(C)	0.20±0.02	0.25±0.02	0.26±0.03	0.28±0.03
	P(C|R)	0.28±0.03	0.34±0.03	0.35±0.03	0.40±0.04
	P(C|N)	0.18±0.02	0.23±0.02	0.25±0.03	0.24±0.03
Save	P(S)	0.61±0.04	0.68±0.04	0.65±0.03	0.71±0.03
	P(S|R)	0.66±0.06	0.69±0.05	0.67±0.05	0.66±0.05
	P(S|N)	0.55±0.04	0.65±0.04	0.58±0.04	0.67±0.04

ward trend as the length of the result summaries (dictated by the interface conditions) increased (3.72 ± 0.34 for T0, to 3.28 ± 0.31 for T4). This was also true for the number of SERPs examined. However, the depth to which subjects went to per query follows a downward trend. As the length of each result summary increased, subjects were likely to go to shallower depths per query when examining result summaries (15.44 ± 1.81 for T0, to 13.53 ± 1.95 for T4). Although not significantly different, we observe that result summary length does appear to influence searcher stopping behaviours. When using interface T4 for example, subjects viewed fewer SERPs per query, and correspondingly examined to shallower depths. Taken together, these two measures demonstrate that a relationship may exist between these factors.

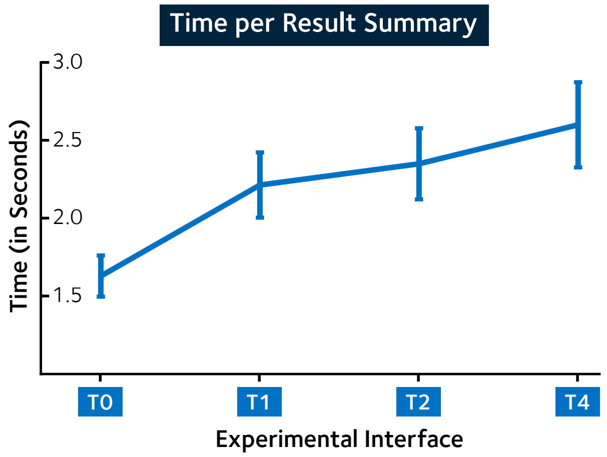
Interaction probabilities all showed an increasing trend as result summary length increased over the four experimental interfaces, as shown in Table 7.3. Explanations for what each of the different probabilities represents can be found in Section 6.4.2.3 on page 167. Although no significant differences were observed over the four interfaces and the different probabilities examined, values reported across all probabilities generally showed an increasing trend as result summary lengths increased. An increase was observed for both the probability of clicking result summaries on the SERP ($P(C)$) and saving the associated documents ($P(S)$) as relevant were observed. When these probabilities are examined in more detail by separating the result summaries clicked and documents saved by their TREC relevance, we see increasing trends for clicking and saving – both for TREC relevant ($P(C|R)$ and $P(S|R)$ for clicking and saving, respectively), and TREC non-relevant documents ($P(C|N)$ and $P(M|N)$). This interesting finding shows that an increase in result summary length does not necessarily improve a subject's accuracy, but simply the likelihood that they will consider documents to be relevant, making them more *click happy*.

7.2.2.2 Time-Based Measures

Table 7.2 also presents three time-based measures (within the **Times** grouping) that were observed across the four experimental interfaces. We show:

- the mean time spent by subjects issuing queries (**Per Query**);
- the mean time spent by subjects examining individual documents (**Per Document**); and
- the mean time spent examining individual result summaries (**Per Summary**).

No significant differences were found between the time spent per query, and the time spent examining individual documents. However, a difference did exist for the time spent per result summary, as can be seen from the table. A clear upward trend in the time spent examining result summaries can be seen in Figure 7.4 as their lengths increased, from 1.63 ± 0.13 for T0 to 2.6 ± 0.27 for T4, and this difference was statistically significant



Interface	Time (in Seconds)
T0	1.63±0.13
T1	2.21±0.21
T2	2.35±0.23
T4	2.60±0.27

Increasing Time

The longer the presented result summary, the greater the time (on average) required to examine them

Figure 7.4 Plot and table illustrating the mean time spent examining result summaries across each of the four experimental interfaces trialled. Note the increasing mean examination time as the result summary length increases, from T0 → T4. Error bars denote the standard deviation.

$(F(3, 208) = 3.6, p = 0.014)$. The follow-up Bonferroni test showed that significant differences did exist between interfaces T0 and T4. This finding suggests that as result summary lengths increased, the amount of time spent examining individual result summaries also increased. This also complies with trends observed regarding examination depths. When the length of result summaries increased, subjects were likely to examine result summaries to shallower depths. This is an intuitive result; given the cost-interaction hypothesis (Azzopardi, 2011), one would expect a searcher to examine less content if the cost of such content (i.e. the greater length, and amount of information) were to increase.

7.2.2.3 Performance

Also included within Table 7.2 are our reported performance measures, this time shown under the **Performance** grouping. Again, these are reported over each of the four experimental interfaces trialled. We report the mean performance of:

- the issued queries, with the corresponding **P@10** score;
- the number of documents that were saved by subjects (**#Saved**); broken up into

- the interactive precision, or the number of saved documents that were TREC relevant (**#TREC Saved (iP)**); and
- the number of saved documents that were not TREC relevant (**#TREC Non.**).

Like the interaction measures examined previously, no significant differences were observed over the four experimental interfaces examined. The performance of queries issued by subjects was very similar across all interfaces ($P@10 \approx 0.25$), along with the number of documents identified by subjects as relevant (6.49 ± 0.58) for T2 to 7.6 ± 0.79 for T4, and the interactive precision (2.28 ± 0.25 for T1 to 2.66 ± 0.32 for T4). Considering the number of saved TREC non-relevant documents, subjects saved on average two of these documents. However, there were no significant differences between the four interfaces.

7.2.2.4 User Experience

In this section, we analyse the results of the post-task and post-experiment surveys that subjects completed. Examining the results from these surveys allowed us to capture the perceived experiences of the subjects when using the experimental system across all four interfaces trialled.

Post-Task Surveys Table 7.4 presents the mean set of results (and their standard deviations) from subjects across the four interfaces trialled. The survey questions are detailed in Section 7.2.1.5. Using a seven-point Likert scale for their responses (with 1 denoting a strong disagreement, and 7 denoting a strong agreement), significant differences were found in all question responses, as shown below.

- **Clarity** $F(3, 208) = 5.22, p = 0.001$
- **Confidence** $F(3, 208) = 5.2, p = 0.001$
- **Informativeness** $F(3, 208) = 5.22, p = 0.001$

Table 7.4 Summary table of the recorded observations for the post-task surveys, indicating the preferences of subjects over the six criteria and four experimental interfaces. Across all criteria, T0 was significantly different from the other three interfaces. Using the seven-point Likert scale, results are shown from 1 (strongly disagree) to 7 (strongly agree).

	T0	T1	T2	T4
Clarity	4.16±0.27	5.00±0.21	5.06±0.24	5.40±0.20
Confidence	3.71±0.26	4.66±0.26	4.75±0.24	5.06±0.25
Informativeness	4.20±0.30	5.38±0.24	5.27±0.24	5.62±0.20
Relevance	3.84±0.28	4.89±0.25	5.08±0.24	5.36±0.20
Readability	5.18±0.31	6.32±0.17	6.46±0.14	6.36±0.14
Size	4.00±0.31	4.94±0.25	5.21±0.22	5.36±0.19

- Relevance $F(3, 208) = 6.44, p < 0.001$
- Readability $F(3, 208) = 9.25, p < 0.001$
- Size $F(3, 208) = 7.28, p < 0.001$

However, follow-up Bonferroni tests show that a significant difference occurred only between interface T0 and T1, T2 and T4. A series of discernible trends can be observed throughout the responses, with subjects regarding longer result summaries as clear and concise, and possessing a higher degree of clarity (4.16 ± 0.27 for T0 to 5.4 ± 0.2 for T4). This improved clarity also provided subjects with greater confidence that longer result summaries helped them better determine the degree of relevance to a given topic. Interaction results presented above however differ from this (as shown in Table 7.3), where the overall probability of saving documents increased, regardless of the document/topic relevance judgement. Other notable trends observed from the results included an increase in how informative subjects perceived results to be – again, with longer result summaries

Table 7.5 Raw results of responses from the post-experiment exit survey completed by each subject. More information on the survey can be found in Section 7.2.1.6, with results discussed in Section 7.2.2.4. Questions recording the highest value(s) for each interface are highlighted.

	T0	T1	T2	T4
Most Informative	1	4	19	29
Least Helpful	46	4	1	2
Easiest	4	4	24	21
Least Useful	48	4	0	1
Most Preferred	3	5	19	26

proving to be more informative. Subjects also reported a general increase in satisfaction of the length of the presented result summaries. However, as mentioned, no significant difference existed between the three interfaces in which snippets were generated as part of the result summaries (i.e. T1, T2 and T4).

Post-Experiment Survey As detailed previously in Section 7.2.1.6, subjects completed a post-experiment exit survey. Responses from the subjects are presented in Table 7.5. From the results, subjects found result summaries of longer lengths (i.e. those generated by interfaces T2 and T4) to be the more informative, and those generated by T0 – without any snippet(s) – to be both the least helpful and useful. Longer result summaries were also consistently favoured by subjects who preferred them over the result summaries generated by interfaces T0 and T1. Subjects also found the result summaries of longer lengths to be better in helping them satisfy their given information need.

From the results, it is clear that a majority of subjects preferred longer result summaries to be presented on SERPs, generated by interfaces T2 and T4. This is illustrated in Table 7.5 by the highlighting of the key results. Note that interface T0 tended to be the most popular option for questions with a negative tendency.

7.2.3 Discussion and User Study Conclusions

This user study investigated the influence of result summary length on search behaviour and performance. Using KL-divergence (Kullback and Leibler, 1951) as a measure of information gain, we examined result summaries of different lengths. We selected a series of result summary lengths (comprised of snippet fragments) where there was a significant difference in information gain between them, which in turn yielded the configurations for our four experimental conditions, T0, T1, T2 and T4. A crowdsourced, within-subjects user study was performed comprising of 53 subjects, each of whom undertook four search tasks, using each of the four experimental interfaces. This work addressed two key research questions, which explored how: SNIPPET-RQ1 the length of a result summary affected search behaviour and user experience; and SNIPPET-RQ2 whether the length of result summaries affected the decision making ability and accuracy of the subjects.

Addressing SNIPPET-RQ1 first in terms of search behaviour, there were few significant differences, but we did observe the following trends. As result summary length increased, subjects issued fewer queries and examined fewer SERPs, *but importantly demonstrated a higher probability of clicking result summary links.* Our results also show that in terms of experience, subjects broadly preferred longer result summaries. Subjects reported that longer summaries were more clear, informative, and readable. In addition to this, the longer result summaries also gave subjects more confidence in their relevance decisions.

With respect to SNIPPET-RQ2, we again observed little difference in subjects' decision making abilities and accuracy between the four experimental interfaces. While subjects perceived longer result summaries to help them infer relevance more accurately, our empirical evidence suggests otherwise. In fact, it would appear that longer result summaries were more attractive, increasing the information scent of the SERP. This may account for the increase in clicks at higher ranks. However, the accuracy of our subjects did not improve with longer result summaries, nor did they find more relevant documents. Increased confidence in the result summaries (from interfaces T0 → T4) may have led to a more relaxed approach at saving content as relevant, as can be seen by increasing click and mark probabilities for both relevant and non-relevant content. It is also possible that the *paradox of choice* (Oulasvirta et al., 2009) could play a role in shaping a searcher's preferences. For example, in interface T4, subjects viewed fewer results/choices than when using other interfaces. This may have contributed to their feelings of greater satisfaction and increased confidence in their decisions.

These novel findings provide new insights into how searchers interact with result summaries in terms of their experiences and search behaviours. Previous work had only focused upon task completion times and accuracy of the first result while not considering their experiences (Cutrell and Guan, 2007; Kaisser et al., 2008). Our findings show that while containing a greater amount of information content, longer result summaries are not necessarily better in terms of decision making. However, subjects perceived this to be the case. We also show a positive relationship between the length and informativeness of result summaries and their attractiveness (clickthrough rates). These results show that the experiences and perceptions of searchers (and the actual performance of those searchers) is different, and when designing interfaces, this needs to be taken into account.

7.3 Simulated Analysis

With our user study now reported, we move onto our corresponding simulations of interaction. In particular, this section reports on how the twelve different result summary level stopping strategies performed (defined earlier in Chapter 5 on page 121):

- HL-RQ3a perform; and
- HL-RQ3b approximate actual searcher stopping behaviours.

For both research questions, these are addressed under the context of varying result summary lengths. In the remainder of this section, we provide methodology details specific to this study (Section 7.3.1), before providing the results of the simulations (Section 7.3.2).

7.3.1 Methodology

This methodology section outlines the details specific to this set of simulations. One can assume that any components not discussed here were instantiated as presented in the general simulation methodology, provided in Section 6.4 on page 157. As we wish to examine how stopping behaviours vary when searchers are exposed to interfaces with result summary snippets of different lengths, we utilised all four of the experimental interfaces defined earlier. These are discussed below (Section 7.3.1.1), before we outline the interaction costs and probabilities extracted for each interface (Section 6.4.2.1).

7.3.1.1 Experimental System and Interfaces

The experimental system used for these simulations was largely the same as outlined in Section 6.4 on page 157 – save for the incorporation of the snippet generation components as outlined previously in Section 7.2.1.2 within the **SimIIR** framework. By incorporating this component, this allowed us to run simulations whose simulated searchers were also subjected to interfaces T0, T1, T2, and T4. We could then mirror closely – given the experimental setup – the interfaces that the real-world searchers were subjected to.

7.3.1.2 Interaction Costs and Probabilities

We then took the interaction log data from the associated user study. Given the four experimental interfaces, we could then (following the methodology outlined in Sections 6.4.2.3 and 6.4.2.1) extract different interaction probabilities and costs to ground our simulations.

Table 7.6 Summary table of the different **interaction costs** (in seconds) and probabilities, with **P(C)** denoting the probability of a click, and **P(S)** denoting the probability of saving a document (considering it relevant). Refer to Sections 6.4.2.1 and 6.4.2.3 respectively for further information on how the costs and probabilities were extracted. All probabilities in this table is attained from interaction data from the user study reported in Section 7.2.

		T0	T1	T2	T4
P(C)	**P(C|R)**	0.28	0.34	0.35	0.40
	P(C|N)	0.18	0.23	0.25	0.24
P(S)	**P(S|R)**	0.66	0.69	0.67	0.66
	P(S|N)	0.55	0.65	0.58	0.67
Costs (in seconds)	**Query**	8.29	7.99	9.42	8.12
	SERP	3.22	3.56	3.93	3.45
	Result Summary	1.63	2.21	2.35	2.60
	Document	17.32	22.82	17.19	18.99
	Save	1.26	1.11	1.26	1.17
	Time Limit	360 seconds (refer to Section 7.2.1.4)			

The interaction probabilities and costs for each of the four interfaces are presented in Table 7.6. Included in the table under the **P(C)** and **P(S)** groupings are:

- the probabilities for clicking a result summary link, broken down over whether the associated document is TREC relevant (**P(C|R)**) or not (**P(C|N)**); and
- the probabilities for saving a document (denoting its relevance to the given TREC topic), again broken down over whether the document is TREC relevant (**P(S|R)**) or not (**P(S|N)**).

Interaction costs denote the time required by the simulated searchers to undertake different tasks. Interaction costs listed in Table 7.6 (under the **Costs** grouping) include:

- the time taken to issue a query (labelled **Query**);
- the time taken to perform an initial examination of the SERP (**SERP**);
- the time taken for a simulated searcher to examine an individual result summary (**Result Summary**);
- the time taken for a document examination (**Document**); and
- the time required to save a document (**Save**).

Details for what constitutes each individual interaction cost can be found, as previously stated, in Section 6.4.2.1 on page 163. Note that the SERP examination cost was included even though the SERP stopping decision point was disabled in these simulations; this ensured that a cost was still paid when performing a SERP examination, even if the outcome was always the same.

Regarding the total session time, simulated searchers were permitted a total of 360 seconds to perform each search session. This was the same total session time used in the user study analysis, as discussed in Section 7.2.1.4. Simulated searchers within this time period would save as many documents as possible that were judged to be relevant according to the prerolled action judgement files (refer to Section 6.4.2.3 on page 167).

7.3.2 Results

We now report the results of our simulations of interaction. We discuss our findings over two subsections, considering:

- the *performance* runs (Section 7.3.2.1), where we discuss the highest levels of performance attained by simulated searchers under different *what-if* scenarios; and

- the *comparison* runs (Section 7.3.2.2), where we provide results of the simulations that were directly compared to actual mean searcher stopping behaviours.

Both of these sections provide an answer for high-level research questions HL-RQ3a and HL-RQ3b respectively, under the context of varying result summary lengths.

Significance Testing In order to determine what result summary level stopping strategies were different from others, we employed significance testing. All tests in this section utilise the two-tailed Student's t-test, where $\alpha = 0.05$. We compared the best performing or approximating stopping strategies against the other eleven. Here, we are interested in *statistical non-significance* (i.e. $\alpha > 0.05$), meaning that the compared stopping strategies are *similar* to one another in terms of performance or approximations.

7.3.2.1 Performance

Before discussing the performance *(what-if)* results, we must first determine whether the implemented querying strategy QS13 delivered queries of expected performance. Recall that QS13 is an *interleaved querying strategy*, where queries from two other querying strategies QS1 (poor) and QS3 (good) were interleaved together.[9] In turn, this allowed us to determine how *robust* a given result summary level stopping strategy was. With a poor query, searchers would be best placed to abandon the associated SERP at shallow depths, as an example. As such, we first examined the average performance of all generated queries issued to the underlying retrieval system. An example of the interleaving approach that we employed is demonstrated in Figure 7.5. In this illustration, we see four actual queries that were issued for the *piracy* topic. Single term queries (i.e. `` `clashes' ``) were generated by QS1; three term queries (i.e. `` `piracy taking control' ``) were generated by QS3.

Table 7.7 reports on a number of different *P@k* measures for each of the three individual querying strategies. We consider *P@1*, *P@5*, *P@10* and *P@20*. Values (± standard devia-

[9] Refer to Section 6.4.2.2 on page 164 for additional information on how these querying strategies were implemented.

Table 7.7 Mean *P@k* values (± standard deviations) of all generated queries issued for performance runs. Precision values are reported at depths of **1**, **5**, **10** and **20** over QS1 (single term queries), QS3 (three term queries) and interleaved querying strategy QS13. Note the general increase in average query performance as we tend from QS1 → QS3.

	QS1	QS13	QS3
P@1	0.04 ± 0.20	0.19 ± 0.39	0.23 ± 0.43
P@5	0.03 ± 0.07	0.14 ± 0.21	0.18 ± 0.23
P@10	0.02 ± 0.07	0.12 ± 0.18	0.14 ± 0.19
P@20	0.03 ± 0.07	0.08 ± 0.13	0.10 ± 0.14

tions) for each of these measures were computed over each of the individual queries issued during the performance runs. A total of 101 unique queries were identified across the five topics trialled to produce these results. As per the querying strategy descriptions outlined in Section 6.4.2.2, we split queries into sets for either QS1 (for single term queries) or QS3 (for three term queries), and both sets for QS13. From the table, we can see from left (QS1) to right (QS3) an increasing trend in performance, demonstrating that the performance of the queries that were issued was in line with our expectations. Moving forward as we report the performance of individual stopping strategies, this provides confirmation that the querying strategies were working as intended. We believe that QS13 provides a good test environment to evaluate the robustness of the twelve individual result summary level stopping strategies.

We now turn our attention to the main results of the performance *(what-if)* simulations. We primarily consider HL-RQ3a in our reporting, which requires an examination of the performance for each result summary level stopping strategy. Before this, we consider the general trends that we observed from the results of the experiments and consider the difference between the four experimental interfaces.

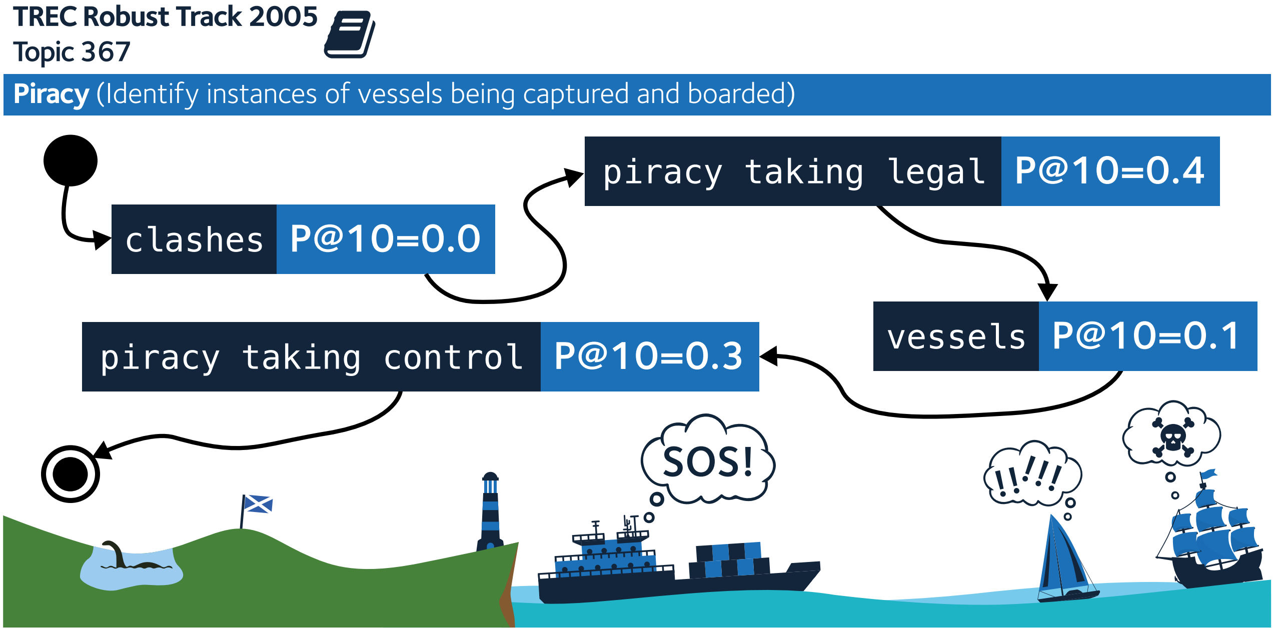

TREC Robust Track 2005
Topic 367
Piracy (Identify instances of vessels being captured and boarded)
clashes P@10=0.0
piracy taking legal P@10=0.4
vessels P@10=0.1
piracy taking control P@10=0.3
SOS!
!!!!!

Figure 7.5 Illustration highlighting several queries issued during the performance *(what-if)* experiments, as generated by **QS13**. The selected topic illustrated is TREC topic 367, *piracy*. Notice the interleaving between single term and three term queries, along with the varying levels of performance between single term and three term queries, represented here by the *P@10* score.

Figure 7.6 provides twelve individual plots, one per result summary level stopping strategy. The plots represent the mean levels of performance attained at varying depths per query, averaged over the five individual topics[10] and 50 individual trials. This is shown across interfaces **T0**, **T1**, **T2** and **T4**. Each point on a plotted line represents a stopping threshold parameter configuration for a given stopping strategy. The mean depth per query is represented along the x axis, with the performance attained (represented as CG) represented on the y axis. Although some stopping strategies caused simulated searchers to browse to depths greater than 25 on average, we cut all plots at this value for consistency, and to better illustrate how performance varies at shallower depths.

General trends across all twelve stopping strategies can be observed from the plots in Figure 7.6. We first note that as we alter the various stopping threshold parameter values that we trialled, we see that mean performance is attained from shallower to greater depths per

[10] This figure includes the practice topic, *privacy* – as real-world queries were not required for this set of experiments, we could use this topic.

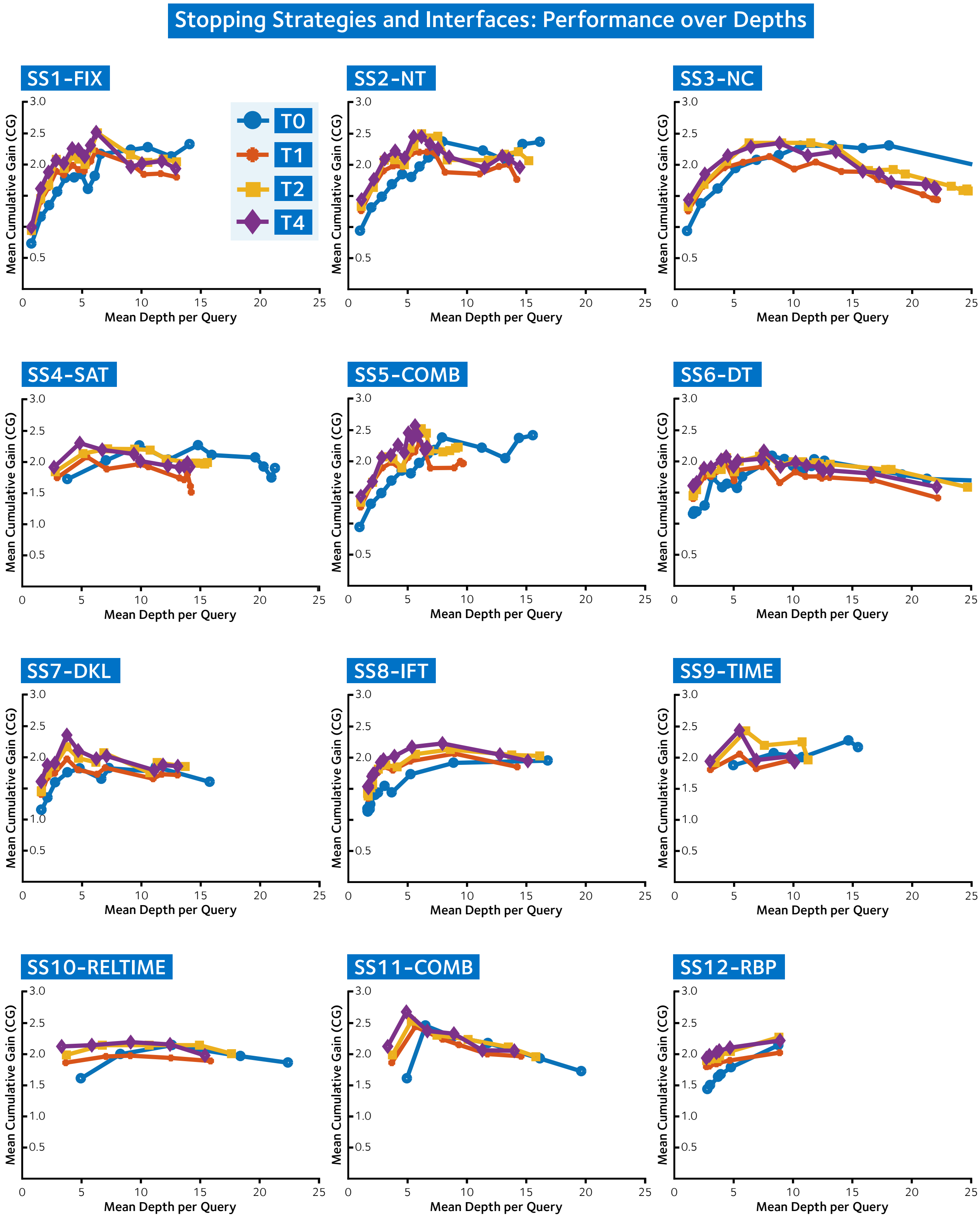


Figure 7.6 Plots showing the varying levels of performance, measured in CG, against the mean depth per query. Each result summary stopping strategy is shown on an individual plot, with each of the four experimental interfaces shown within each plot. The depth per query reported on each *x* axis is cut at 25 to allow for an easier comparison between different stopping strategies.

query. In turn, this generally results in a gain in overall performance, with the mean CG attained generally increasing as the depth per query increases. However, this is true only until a certain point, representing the maximum level of mean CG attained. After this point, which is illustrated in a more profound way with some stopping strategies than others, a simulated searcher would begin to browse to greater depths on average. At greater depths, the searcher would encounter fewer and fewer potentially relevant documents, and thus would begin to waste time examining the same SERP. This is represented in the plots as the downward trend of mean CG at greater depths per query, clearly visible after the highest level of CG is attained. This general trend can be clearly observed from our baseline, fixed-depth stopping strategy, approximately at SS1-FIX @6. An increase in mean CG up until a mean depth per query of ≈ 6 (for most interfaces) can be observed; after this point, we generally observed a gradual drop-off in performance.

Considering the four plots in Figure 7.6 from the perspective of the four experimental interfaces, little difference can be observed across all twelve stopping strategies, and across the varying depths per query. The same general trends can be observed across T0, T1, T2 and T4, with plotted lines representing each interface being largely invariant to each other. However, some interesting observations can be made. Even though the mean levels of CG are all very similar, we do find that interfaces yielding snippet text of greater length (interfaces T2 and T4) generally outperform the interfaces with minimal and no snippets (interfaces T1 and T0, respectively). This is an unsurprising result – performance of subjects in interface T0 was generally worse in the reported user study, and this poorer performance had a subsequent impact upon the simulations of interaction which were grounded by interaction data from the said user study.[11] However, as the mean depth per query increases, we find in all plots reported in Figure 7.6 that the mean level of CG begins to close up over each of the four interfaces. This can be attributed to the fact that at greater depths, the likelihood of encountering relevant material decreases and will likely converge. Generally however, results are consistent across the four experimental interfaces.

[11] Refer to Table 7.6 for the different interaction costs and probabilities used to ground the simulations.

Table 7.8 Results from the simulated *what-if* performance runs, showing the highest levels of CG attained for each result summary level stopping strategy trialled. **x_n** denotes the parameter threshold(s), with **DQ** denoting the depth per query at which the greatest **CG** value was attained at. For each interface, the stopping strategy which attained the highest level of CG is highlighted. Light blue highlighting denotes *no significant difference* from the best performing strategy, with **no highlighting** denoting a significant difference at $\alpha=0.05$. For combination thresholds, x_2,x_4 are presented for SS5-COMB, with x_{10},x_4 for SS11-COMB.

		T0			T1			T2			T4		
		x_n	DQ	CG	x_n	DQ	CG	x_n	DQ	CG	x_n	DQ	CG
FIX	SS1	24	14.09	2.31	10	6.19	2.20	10	6.33	2.50	10	6.23	2.50
FRUS	SS2	10	7.94	2.36	8	6.61	2.20	7	6.16	2.49	6	5.49	2.44
	SS3	8	13.22	2.31	5	7.91	2.13	4	6.19	2.35	5	8.76	2.35
SAT	SS4	5	14.79	2.26	2	5.40	2.08	3	7.15	2.21	2	4.80	2.30
COM	SS5	24,8	15.58	2.41	8,4	6.05	2.30	8,4	6.17	2.52	9,3	5.65	2.56
DIFF	SS6	0.55	8.23	2.08	0.35	4.35	1.96	0.55	7.48	2.12	0.55	7.53	2.16
	SS7	3.5	11.24	1.84	6.0	3.73	1.97	6.0	3.71	2.16	6.0	3.72	2.35
IFT	SS8	0.002	16.83	1.95	0.004	8.96	2.06	0.006	8.59	2.13	0.006	7.93	2.23
TIME	SS9	120	14.61	2.27	60	5.43	2.05	60	5.95	2.42	60	5.41	2.43
	SS10	30	12.60	2.14	30	9.09	1.98	20	6.73	2.14	30	9.12	2.19
COM	SS11	10,8	6.53	2.45	10,10	5.66	2.44	10,10	5.33	2.52	10,10	4.91	2.67
RBP	SS12	0.99	8.78	2.15	0.99	8.87	2.03	0.99	8.81	2.27	0.99	8.90	2.22

With the twelve plots in Figure 7.6 presenting a broad overview of the variation in performance as the mean depth per query increases, we now turn our attention to the peaks in each plot for the four individual experimental interfaces – or the *highest levels of CG that were attained.* Table 7.8 reports these values, across each of the twelve stopping strategies (rows), and over the four individual experimental interfaces (columns, grouped by T0, T1, T2 or T4). For each of the twelve stopping strategies, we report: the highest level of CG attained (**CG**); the mean depth per query at which this value was reached (**DQ**); and the stopping threshold parameter value(s) used to reach this value ($\mathbf{x_n}$). Highlighted are the stopping strategies that yielded the highest overall level of CG. Across all four interfaces, combination stopping strategy SS11-COMB attained this, with mean CG values of 2.45, 2.44, 2.52 and 2.67 reported for interfaces T0, T1, T2 and T4 respectively. These values were all reached at similar depths per query ($5 - 6.5$), and with a similar range of threshold values, with $x_4 \approx 8 - 10$, and $x_{10} = 10$ seconds.[12] The relatively low levels of mean depths per query (6.53, 5.66, 5.33 and 4.91 for interfaces T0, T1, T2 and T4 respectively) at which the best CG was attained also demonstrates that SS11-COMB was particularly robust at detecting a SERP with good results, and vice versa. As such, the low depths per query indicate that the simulated searchers were confidently able to abandon poor quality SERPs without affecting their overall performance. Of course, this result is unsurprising considering how we instantiated SS11-COMB. Using *P*@1 to determine the patch yield type, our approach drew upon TREC QRELs. This meant that the strategy would make the correct decision (as per the theory) for every SERP examined.

Upon closer examination of Table 7.8, we also find that the other combination strategy SS5-COMB (relating to a combination of both frustration and satisfaction stopping strategies) consistently placed second in performance rankings across the four experimental in-

[12] As a reminder, SS11-COMB considers the type of patch presented by a given SERP, before employing either the satisfaction-based stopping strategy SS4-SAT for high yields early on. Alternatively, the give-up time-based stopping strategy SS10-RELTIME was selected if the SERP did not appear to yield promising results early on.

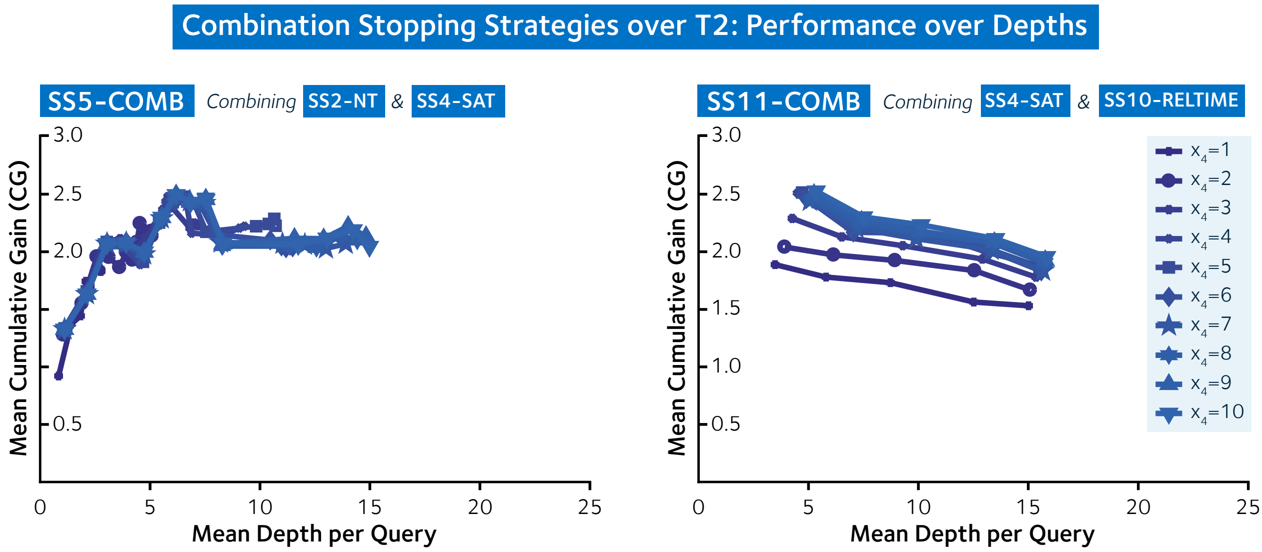


Figure 7.7 Plots illustrating performance over varying depths per query. Reported are performance values over combination strategies SS5-COMB (left) and SS11-COMB (right). With each line representing a value from x_4 (refer to legend), each point on the lines represents performance and depth for a threshold value from x_2 on the left, and x_{10} on the right. Little difference in performance is observed between variations of parameter combinations.

terfaces trialled, very close behind the mean CG values attained by SS11-COMB (with CG values ± standard deviations for both combination strategies reported in Table 7.9). Again, this demonstrates that a combination strategy appears to be very effective at eliciting good levels of performance, and suggests that a degree of flexibility in selecting their stopping criterion/criteria is advantageous to searchers.

It should be noted that the two plots reported in Figure 7.6 on page 222 for both SS5-COMB and SS11-COMB are shown over the best performing x_4 value for each stopping strategy, as reported in Table 7.8. With two sets of parameters for these combination strategies, reporting in these plots would have been difficult – Figure 7.7 instead reports the varying levels of CG across different mean depths per query, for each of the different x_4 values trialled. These are shown over interface T2 only; similar plots were observed for the other three experimental interfaces, and as such are not reported here. We can see from these plots that similar trends can be observed across the range of mean depths per query over

Table 7.9 The highest levels of CG attained by the two combination result summary level stopping strategies, across the four different experimental interfaces. Reported in the table are the standard deviations, demonstrating a high variance between trials.

	T0	T1	T2	T4
SS5-COMB	2.41±2.47	2.30±2.44	2.52±2.67	2.56±2.27
SS11-COMB	2.45±2.59	2.44±2.41	2.52±2.52	2.67±2.52

both stopping strategies. The change in performance for SS11-COMB is more profound – as x_4 increases, so too does the mean level of CG attained.

With both SS5-COMB and SS11-COMB performing very well in terms of the highest levels of CG attained, a cursory examination of Table 7.8 will also confirm that several other strategies also perform to a high standard. With these generally high levels of performance being reported, we decided to perform statistical significance testing to determine if the performance of any stopping strategies were significantly different (or not) from that of SS11-COMB. As discussed previously in this section, we performed two-tailed Student's t-tests over the CG values, comparing SS11-COMB against the other eleven stopping strategies. Results of the statistical testing showed that a majority of stopping strategies were indeed *not significant* ($p > 0.05$), denoting that the performance was similar to SS11-COMB. In Table 7.8, highlighted cells denote that the represented stopping strategy was similar in terms of performance to the best performing. Stopping strategies without any cell highlighting offered statistically significant differences, meaning that CG values were worse than the best reported over SS11-COMB. Of the eleven remaining stopping strategies, we generally observe – across the four interfaces – that the following showed similar levels of performance:

- SS1-FIX, the fixed-depth stopping strategy;
- SS2-NT and SS3-NC, the frustration-based strategies;

- SS4-SAT, the satiation-based stopping strategy;
- SS5-COMB, the frustration and satiation combination strategy;
- SS8-IFT, the IFT-based stopping strategy; and
- SS9-TIME and SS10-RELTIME, the time-based stopping strategies.

This accounts for a majority of the remaining eleven stopping strategies trialled. The remaining three:

- SS6-DT and SS7-DKL, the difference-based stopping strategies; and
- SS12-RBP, the RBP-based stopping strategy

generally offered significantly different levels of CG (across the four experimental interfaces) from those of SS11-COMB. We now briefly examine the remaining stopping strategies, before moving to the reporting of the real-world simulated comparisons.

We first consider SS1-FIX, SS2-NT and SS3-NC together. Examining the figures for the first three stopping strategies in Figure 7.6, we observe very similar plots following the aforementioned trends in performance over mean depths per query. Comparing the plots for SS1-FIX and SS2-NT in particular, striking similarities can be observed. SS3-NC yields similar plots, but spread over greater mean depths per query.[13] Examining Table 7.8, similar levels of CG can also be attained over three stopping strategies, all at similar mean depths per query (2.50 at a DQ of 6.33, 2.49 at a DQ of 6.16 and 2.35 at a DQ of 6.19 for SS1-FIX, SS2-NT and SS3-NC respectively, over interface T2). This result is interesting as intuitively, one would expected both SS2-NT and SS3-NC to offer greater performance over SS1-FIX. This is because the frustration-based stopping strategies are *adaptive* in nature, curtailing the examination of result summaries early when results are mostly non-relevant. This is in contrast to the fixed-depth strategy (and baseline) of SS1-FIX.

[13] This is due to the fact that the stopping criterion for SS3-NC considers a series of contiguous, non-relevant items to be found. This can mean that a searcher subscribing to this stopping strategy will typically examine content to greater depths before meeting this criterion.

Moving to the satiation-based stopping strategy SS4-SAT, we find that the associated plot in Figure 7.6 looks somewhat invariant compared to other strategies. A relatively consistent level of CG can be attained at a range of mean depths per query. It is somewhat surprising how this stopping strategy performs so well, given that it may have been better suited to a session-based stopping decision point (e.g. as applied in the study reported in Chapter 8) than being applied at the result summary level. Nevertheless, SS4-SAT yields good levels of CG (2.10 – 2.30). A low stopping threshold (x_4) value ranging from 2 to 5 provides these levels of CG across the four experimental interfaces, suggesting that to acquire good levels of CG, finding 2 to 5 potentially relevant documents is a good approach to follow.

Turning our attention to the IFT-based stopping strategy SS8-IFT, we notice the lower values of CG that are attained in Table 7.8. These values are generally reached at greater mean depths per query. If we examine the plot for SS8-IFT in Figure 7.6, we notice a drop in mean accumulated CG after a $D/Q \approx 7$. However, the low levels of CG in comparison to other stopping strategies (e.g. 2.13 vs. 2.52 for SS11-COMB over interface T2) suggests that this approach does not work particularly well. This perhaps can be attributed to how the *rate of gain* was calculated – a difficult value to estimate. We leave the issue of calculating this rate of gain parameter to our discussion, presented in Section 10.2.2 on page 340.

We next consider the time-based stopping strategies, SS9-TIME and SS10-RELTIME. The first stopping strategy here can be considered analogous to a fixed-depth approach, considering the time from the point at which a SERP is presented – and is therefore agnostic of relevance. The second strategy can be considered adaptive in the sense that it considers the time from which the last relevant document was saved. To this end, it is a somewhat interesting result that SS9-TIME consistently yields a higher level of CG across all four experimental interfaces (2.27, 2.05, 2.42 and 2.43 for T0, T1, T2 and T4 respectively) than when compared to SS10-RELTIME (2.14, 1.98, 2.14 and 2.19 for T0, T1, T2 and T4 respectively – with the CG attained for T4 significantly different from best-performing strategy, SS11-COMB). Intuitively, one would expect higher performance to be attained by the adaptive approach, where searchers would stop earlier when examining a poor set of results with few (if any) documents saved. The performance across SS10-RELTIME is invariant across mean depths per query, as illustrated in Figure 7.6. A smaller number of points reflects the smaller number of threshold values that we trialled. These are summarised as x_9 and x_{10} in Table 6.3 on page 178.

Difference-based stopping strategies SS6-DT and SS7-DKL are considered next. As reported earlier, these stopping strategies offered significant differences in performance across interfaces T1 and T4 when compared to stopping strategy SS11-COMB. Indeed, performance is generally poor in comparison to other stopping strategies that were trialled in this study. As can be seen from the performance plots in Figure 7.6 however, the greater the depth a searcher was to go on average, the better performance would be. We discuss these findings in more detail in Section 10.2.2 on page 340 – with an emphasis on the poor performance that these strategies yielded.

Our final stopping strategy is the RBP-based approach, SS12-RBP. From the plot in Figure 7.6, it is clear that RBP provides lower levels of CG when compared to other stopping strategies. An intuitive result is that as the patience parameter was increased, simulated searchers would traverse to greater depths on average – the highest levels of CG as reported in Table 7.8 are attained at greater depths on average than SS11-COMB.

While SS11-COMB consistently offers the highest level of CG across the four experimental interfaces, it is clear from our results that this is not significantly so. Several other stopping strategies offer maximum CG values that are very close to that of SS11-COMB. As such, this combination stopping strategy does not offer a significant improvement in performance over our fixed-depth baseline, SS1-FIX. It does, however, offer a marginally greater level of CG at comparatively lower depths per query on average, demonstrating that it is a more robust and efficient means for determining when to stop. This means that following such a strategy may be prudent for a searcher to follow – at least under the search context examined in this work.

7.3.2.2 Real-World Comparisons

From our *what-if* performance simulations that examined what would happen if a particular stopping strategy were to be rigidly followed, we now examine how closely each of the aforementioned stopping strategies compares to actual searcher behaviours. As such, this provides an answer to **HL-RQ3b** under the context of varying result summary lengths. These simulations *replayed* all of the queries issued by real-world searchers, allowing us to compare real-world and simulated click depths.

Figure 7.8 again presents twelve plots, one for each result summary level stopping strategy trialled. Each of the plots illustrates the mean depth per query, again on the x axis. This is plotted against the *Mean Squared Error (MSE)*[14] of the real-world vs. simulated click depths. Each point on the plotted lines represents a given stopping threshold parameter configuration, and its position represents how close the click depth approximation was on average to real-world searcher click depths. The closer the MSE tends towards zero, the closer the simulated searcher's approximation to actual stopping behaviours. These are shown over each of the four experimental interfaces. For reference, we also include on each plot four vertical dashed lines, representing the mean click depths reached by the real-world searchers. Again, a separate line is presented for each experimental interface. For most of the plots, notice how the lowest point for the simulated results tends towards the dashed lines. This indicates that the simulations offered a good approximation of real-world searcher click depths on average. As an example of how to interpret these plots, interface **SS2-NT** over **T2** reaches its lowest MSE value of 77.76 at a mean depth per query very close to the real-world mean (14.67 for **T2** vs. 14.39 for real-world).

From the plots in Figure 7.8, we can observe a number of notable trends. Stopping strategies that offered good levels of performance (as reported in Section 7.3.2.1) generally yielded smoother MSE curves, a trait indicative of providing good approximations of actual searcher

[14] Refer to Section 6.4.3.2 on page 182 for further information on how we computed the *Mean Squared Error (MSE)*.

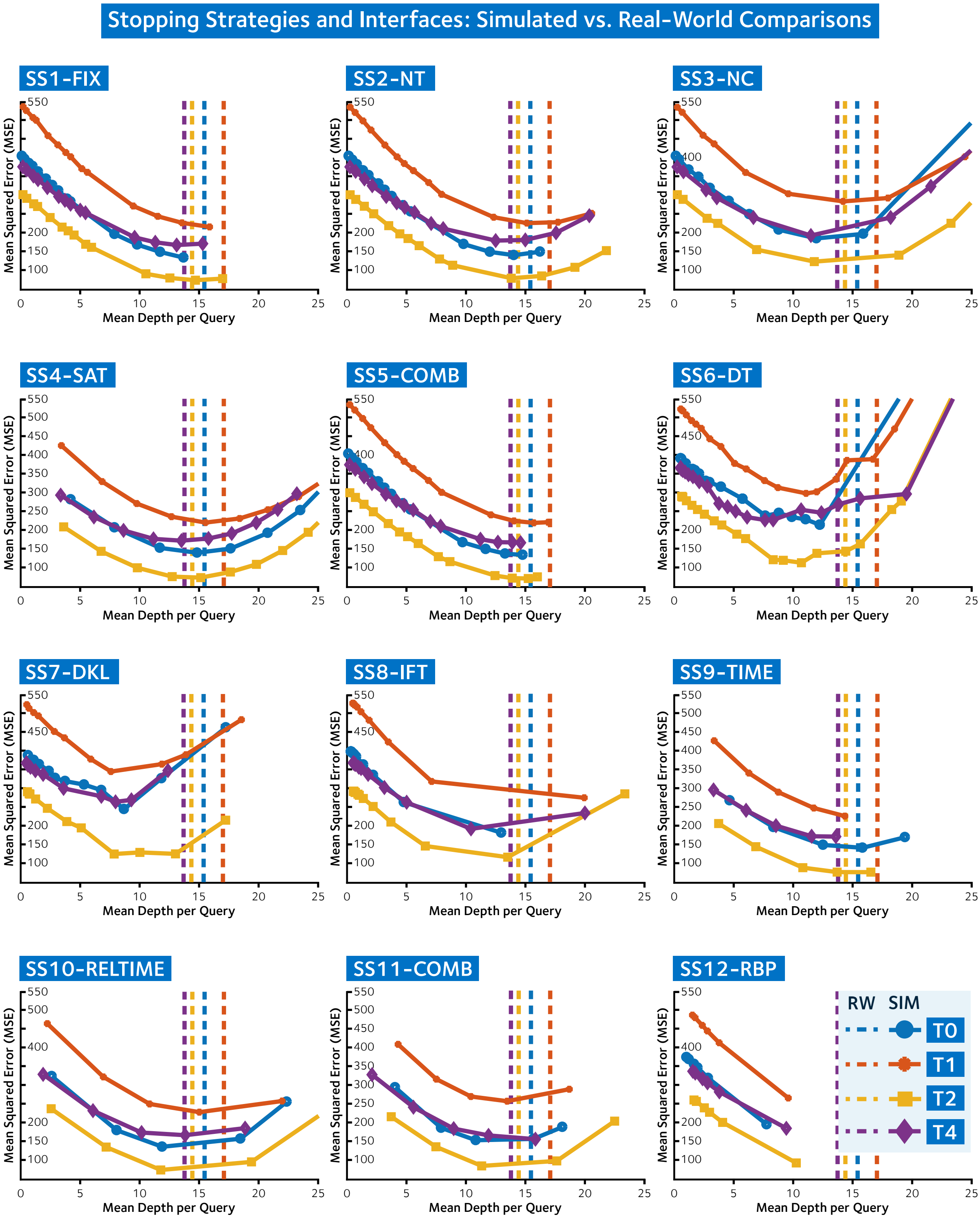


Figure 7.8 Plots reporting the comparison runs, reporting the MSE vs. the mean depth per query. Trials over each of the four experimental interfaces are shown. Also included in the plots are a series of dashed lines denoting the mean depth per query reached by the real-world subjects of the user study. Depths per query (and mean CG values) are also reported in Table 7.11.

Table 7.10 Results from the simulated comparison runs, showing the *lowest* MSE value reached over each result summary level stopping strategy trialled (grouped by their type). $\mathbf{x_n}$ denotes the parameter threshold(s) that the lowest **MSE** was reached with. Results are presented across the four experimental interfaces, including an average over the four to examine if a particular strategy emerges as a better approximation. For each interface, the stopping strategy that attained the lowest MSE is highlighted. For the combination stopping strategies, two parameters are presented, with x_2,x_4 presented for SS5-COMB and x_{10},x_4 presented for SS11-COMB. Significance testing yielded no significant differences between strategies at α=*0.05*.

		T0		T1		T2		T4		Average	
		x_n	MSE	x_n	MSE	x_n	MSE	x_n	MSE	x_n	MSE
FIX	SS1	24	133.90	24	215.48	21	74.04	21	167.38	24	149.66
FRUS	SS2	21	139.15	18	224.61	15	77.76	15	178.05	18	159.02
	SS3	9	183.45	7	282.34	6	121.60	6	191.21	6	224.69
SAT	SS4	4	138.75	5	219.43	5	72.36	5	171.06	5	153.21
COM	SS5	24,6	135.31	21,8	216.96	21,6	72.81	24,6	167.69	21,6	160.63
DIFF	SS6	0.90	214.53	0.70	298.30	0.70	113.59	0.65	227.33	0.70	219.51
	SS7	4.0	244.09	4.5	344.04	4.5	124.92	4.0	263.53	4.0	250.13
IFT	SS8	0.002	180.77	0.002	273.69	0.004	115.98	0.004	191.13	0.004	221.44
TIME	SS9	120	140.16	150	224.59	150	75.36	150	170.02	120	158.16
	SS10	30	136.63	40	229.02	30	74.81	40	167.54	30	159.20
COM	SS11	30,9	152.60	40,10	256.30	30,8	84.39	50,2	155.33	40,9	168.75
RBP	SS12	0.99	195.56	0.99	265.90	0.99	93.02	0.99	185.22	0.99	184.93

behaviours. For example, plots for SS6-DT are more variable in nature; this stopping strategy was also reported in Section 7.3.2.1 as one of the worst performing on average. We also notice a variation in how predictions differ across the four experimental interfaces, with this trend observed over all twelve stopping strategy plots. Interface T2 consistently served better approximations than its counterpart interfaces, with T1 always offering the worst. Interfaces T0 and T4 appear between the two extremes, often interleaving with one another as the mean depth per query increases.

These trends can also be observed in Table 7.10. In this table, we report for each stopping strategy and interface the point on the corresponding plots in Figure 7.8 where the lowest MSE value is attained (**MSE**), and the stopping threshold value(s) ($\mathbf{x_n}$) that were used to attain it. For interface T2 , notice how the MSE values are lower than those of the other three experimental interfaces. The table also highlights the stopping strategy combination that yielded the lowest MSE for each interface, with somewhat surprising results. The baseline, fixed-depth stopping strategy SS1-FIX offered the best approximations for interfaces T0 , T1 and T4 , with satiation stopping strategy SS4-SAT yielding the lowest MSE for interface T2 . These results are somewhat surprising: it would make sense for a searcher to employ a more adaptive approach in determining when they should stop. Furthermore, the satiation stopping strategy would likely make more sense at a session level.

With these interesting results in mind, we also ran a series of statistical significance tests over the reported MSE values. This was to determine if any significant difference existed between approximations for each reported stopping strategy. Our findings showed that when compared to the best stopping strategy for each interface, *no significant differences* were observed. These findings highlight that even though SS1-FIX and SS4-SAT offered the lowest MSE overall, the other eleven stopping strategies *all* offered good approximations of stopping depths, some better than others. In other words, it was hard to deduce from the results a stopping strategy offering a clearly superior means of approximating searchers' mean stopping depths.

Table 7.11 Additional results from the searcher comparisons runs, with this table reporting mean depth per query (**DQ**) and **CG** values, along with the mean interactive precision value (**iP**). All these values are reported at the configuration yielding the lowest MSE (refer to Table 7.10), indicating the best approximation to real-world stopping behaviours. Also included are the mean real-world (**RW**) values over each interface for a direct comparison. Note that result summary level stopping strategies offering the lowest MSE are highlighted — cell colouring here does not denote the outcome of any significance testing.

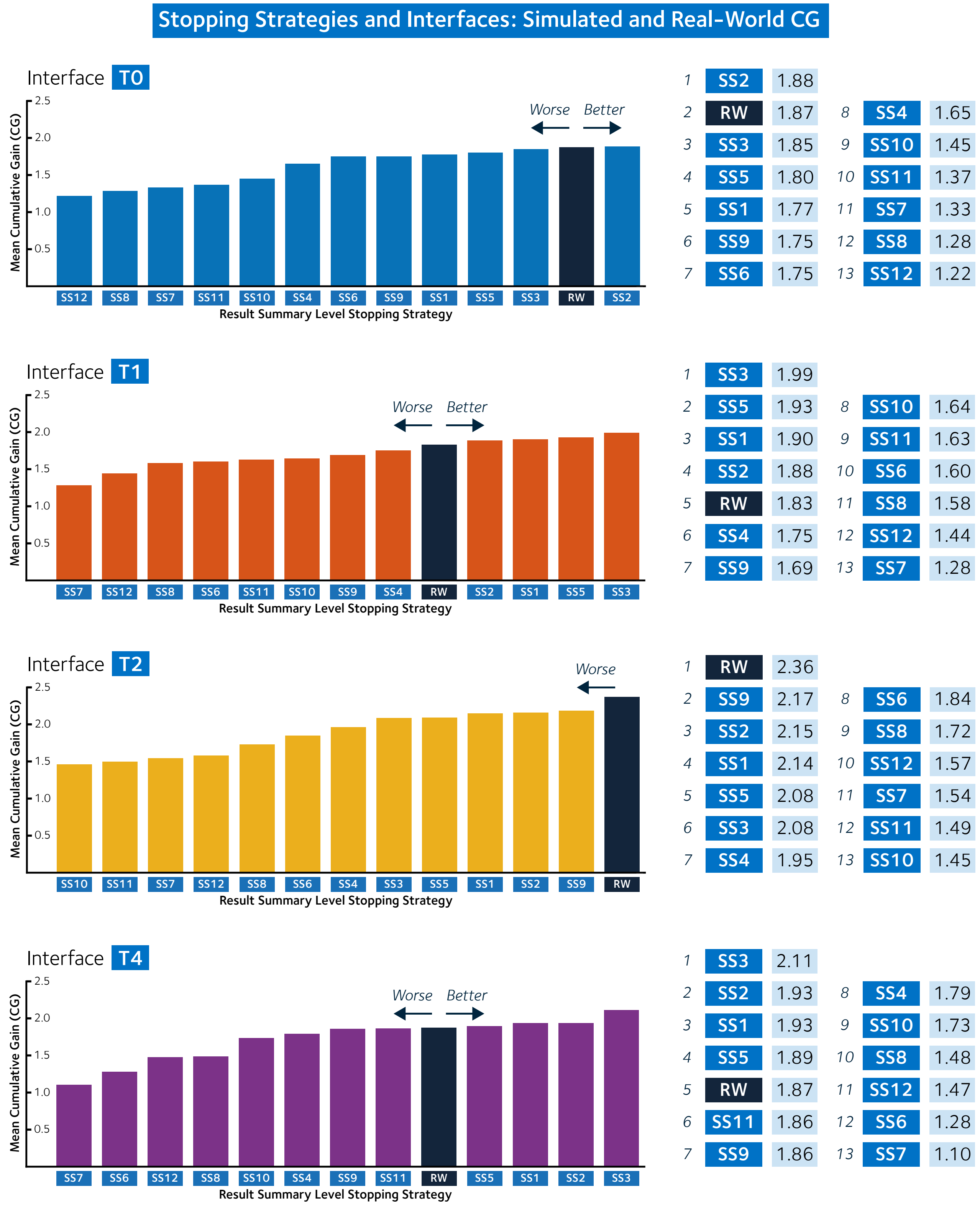
		T0			T1			T2			T4		
		DQ	CG	iP	DQ	CG	iP	DQ	CG	iP	DQ	CG	iP
	RW	15.42	1.87	2.58	17.04	1.83	2.28	14.39	2.36	2.47	13.74	1.87	2.66
FIX	SS1	**13.68**	**1.77**	**1.17**	**15.86**	**1.90**	**1.28**	14.67	2.14	1.41	**13.10**	**1.93**	**1.27**
FRUS	SS2	14.04	1.88	1.25	15.13	1.88	1.25	13.81	2.15	1.42	12.50	1.93	1.27
	SS3	11.99	1.85	1.22	14.23	1.99	1.32	11.80	2.08	1.37	11.52	2.11	1.38
SAT	SS4	14.80	1.65	1.10	15.51	1.75	1.17	**15.10**	**1.95**	**1.26**	13.48	1.79	1.18
COM	SS5	14.76	1.80	1.19	16.45	1.93	1.29	15.22	2.08	1.34	14.58	1.89	1.23
DIFF	SS6	12.25	1.75	1.15	11.08	1.60	1.05	10.69	1.84	1.21	7.62	1.28	0.85
	SS7	8.74	1.33	0.89	7.63	1.28	0.85	7.90	1.54	1.01	8.02	1.10	0.74
IFT	SS8	12.96	1.28	0.86	19.92	1.58	1.06	13.50	1.72	1.12	10.41	1.48	0.97
TIME	SS9	15.79	1.75	1.16	14.30	1.69	1.14	16.49	2.17	1.41	13.54	1.86	1.22
	SS10	11.86	1.45	0.97	14.98	1.64	1.09	11.74	1.45	0.98	13.79	1.73	1.14
COM	SS11	10.78	1.37	0.92	13.42	1.63	1.07	11.27	1.49	1.01	15.80	1.86	1.24
RBP	SS12	7.78	1.22	0.82	9.59	1.44	0.97	10.26	1.57	1.05	9.42	1.47	0.97

Given the findings outlined above, we also decided to examine if a particular stopping strategy emerged as providing a good approximation of stopping behaviours when considering *all four interfaces on average.* Results of this analysis are shown in the **Average** grouping in Table 7.10. Statistical tests comparing the best-approximating strategy (again SS1-FIX @24) against the remaining eleven stopping strategies once again yielded no significant differences, highlighting that all stopping strategies offered similar approximations. As such, we did not explore the concept of averaging over the four experimental interfaces any further.

Moving back to our per interface examination, Table 7.11 reports additional information relating to the best approximations offered by each stopping strategy. We report for each stopping strategy (across each experimental condition) the: mean depth per query (**DQ**); **CG**; and the mean number of saved TREC relevant documents, or interactive precision (**iP**). These values are attained at the stopping threshold parameter(s) that yielded the lowest MSE, as reported in Table 7.10. Also included in Table 7.11 are the mean real-world (**RW**) values attained by the subjects of the user study. We also once again highlight the stopping strategy for each interface that yielded the lowest MSE, as reported in Table 7.10.

Trends from Table 7.11 are largely to be expected: stopping strategies that yielded the closest approximations to actual mean stopping behaviour parallel the mean depth per query to the real-world (**RW**) counterparts. In contrast, we find that for stopping strategies such as SS6-DT and SS12-RBP, the mean depth per query is lower than the mean values across each of the four interfaces. As such, these strategies largely *underestimate* the stopping depth of the real-world searchers. We also find that for interface T2, mean depths per query all appeared to offer closer representations to the real-world mean. This may be an artefact of the probabilities that were used to ground the simulations of interaction.

Stopping strategies offering better approximations also reported a higher mean number of TREC relevant documents that were identified (saved). For example, SS4-SAT @5 over interface T2 reported a mean of 1.26 saved TREC relevant documents. However,



Stopping Strategies and Interfaces: Simulated and Real-World CG

1	SS2	1.88			
2	RW	1.87	8	SS4	1.65
3	SS3	1.85	9	SS10	1.45
4	SS5	1.80	10	SS11	1.37
5	SS1	1.77	11	SS7	1.33
6	SS9	1.75	12	SS8	1.28
7	SS6	1.75	13	SS12	1.22

1	SS3	1.99			
2	SS5	1.93	8	SS10	1.64
3	SS1	1.90	9	SS11	1.63
4	SS2	1.88	10	SS6	1.60
5	RW	1.83	11	SS8	1.58
6	SS4	1.75	12	SS12	1.44
7	SS9	1.69	13	SS7	1.28

1	RW	2.36			
2	SS9	2.17	8	SS6	1.84
3	SS2	2.15	9	SS8	1.72
4	SS1	2.14	10	SS12	1.57
5	SS5	2.08	11	SS7	1.54
6	SS3	2.08	12	SS11	1.49
7	SS4	1.95	13	SS10	1.45

1	SS3	2.11			
2	SS2	1.93	8	SS4	1.79
3	SS1	1.93	9	SS10	1.73
4	SS5	1.89	10	SS8	1.48
5	RW	1.87	11	SS12	1.47
6	SS11	1.86	12	SS6	1.28
7	SS9	1.86	13	SS7	1.10

Figure 7.9 Bar charts, one per experimental interface, demonstrating the mean level of CG attained by each result summary level stopping strategy. Ordered by CG, these values are reached using the threshold configurations yielding the best approximations to actual searcher behaviour, as shown in Table 7.10. Also included are the mean real-world searcher CG values for each interface.

this is lower than the real-world mean of 2.47. Interestingly, we find that stopping strategies SS1-FIX, SS2-NT and SS3-NC all consistently reported the highest levels of mean saved TREC documents across the four interfaces, albeit still lower than the real-world means. For example, SS2-NT @21 reports a mean of 1.25 saved documents over interface T0, compared to 0.89 over SS7-DKL @4.0. This is in contrast to the real-world mean of 1.87 over the same interface.

Considering the levels of CG attained, we also find interesting results. Values reported in Table 7.11 indicate the level of CG that searchers would have accumulated on average if they rigidly followed a given stopping strategy (using the stopping threshold value(s) that yielded the best approximations of actual stopping behaviours). In other words, if rigidly following a given stopping strategy, *would searchers have been able to accumulate higher levels of CG (on average) compared to what they actually achieved?* To better represent these results, we generated a series of bar charts as shown in Figure 7.9. Each bar chart represents an individual experimental interface, with each bar representing the mean level of CG attained over each of the twelve stopping strategies, plus the additional mean real-world CG.

Results show that on average, real-world searchers typically appear on the high end of the spectrum across all four experimental interfaces. This suggests that given the twelve stopping strategies that were trialled, not many would have offered improvements in overall levels of CG. This is especially true for interface T2, where the real-world CG mean topped all twelve stopping strategies. For interfaces T0, T1 and T4 where simulations do offer better levels of CG, we find the same stopping strategies appearing above the real-world mean: SS1-FIX, SS2-NT, SS3-NC and SS5-COMB. These results are interesting, as they suggest that a simple stopping strategy is an effective means for attaining high levels of CG, even that of the fixed-depth baseline. Combination stopping strategy SS11-COMB consistently ranked lower across all four interfaces, even though this yielded the highest levels of CG in the *what-if* performance runs reported in Section 7.3.2.1. This suggests that if employed by our real-world subjects, SS11-COMB (on average) would have potentially allowed the subjects to enjoy higher levels of gain (finding relevant documents).

7.4 Chapter Summary

In this chapter, we have examined how result summary snippet lengths affect a searcher's behaviour, performance and user experience via a crowdsourced user study (Section 7.2). From this user study, we then subsequently used interaction data from said user study to ground an extensive set of simulations of interaction (Section 7.3). These simulations were trialled to determine how each of the twelve stopping strategies proposed in Chapter 5 performed and approximated the mean stopping behaviours of searchers. In turn, these findings provide answers to our two high-level research questions HL-RQ3a and HL-RQ3b when varying result summary snippet lengths.

The main finding from the user study showed that as snippet lengths increased across the four experimental interfaces (i.e. T0 → T4), subjects reported that they became more confident with the decisions they were making with respect to identifying relevant material. This can be cited due to the fact that more text in the result summary yielded a greater insight into the corresponding document – at the cost of greater examination time. However, a disconnect existed between how subjects *believed* they performed, and what was actually attained through empirical evidence. Here, we found that as snippet lengths increased, subjects became more *click happy*, marking more documents as relevant, even though accuracy did not improve. This is particularly clear when examining the interaction probabilities that we extracted from interaction data, as reported in Table 7.3 on page 208.

These interaction probabilities (and costs) were then used as a basis for grounding an extensive set of simulations of interaction. Split across performance (addressing HL-RQ3a) and comparison runs (addressing HL-RQ3b), we examined each individual stopping strategy in terms of both overall performance and how well they approximated actual searcher stopping behaviours. This was considered across the four aforementioned experimental interfaces. Findings for HL-RQ3a show that all twelve stopping strategies offered reasonable levels of CG – although we found that combination stopping strategy SS11-COMB consistently provided the highest levels of CG across all four experimental interfaces. This was however largely reached without achieving statistical significance from the remaining eleven strategies. Likewise, for HL-RQ3b, we found that SS1-FIX appeared to offer the lowest MSE (and thus best approximations) when tuned to actual searcher behaviours – a surprising result.[15] This was largely consistent across interfaces. We also showed that if followed rigidly, several stopping strategies offered improved levels of CG when compared to the real-world mean. No significant differences were obtained between the stopping strategies. The lack of significant differences may be due to the interface variations not possessing a large enough effect, or an insufficient number of subjects to detect it (whose interaction data would have been used for subsequent grounding). This may also be attributed to the way in which we operationalised the stopping strategies examined, something that is discussed later in Section 10.2.2 on page 340. However, from the results, it is clear that a notable trend exists. With fewer SERP pages examined and subjects examining content to shallower depths when result summary lengths increase, this suggests that the length of result summaries do indeed affect stopping behaviours.

Findings from this chapter will be discussed further in Chapter 10. Along with the findings of the two remaining empirical contribution chapters, we will consider all of our findings from the stopping strategy simulations in detail, determining what conclusions can be drawn from this work. Our next chapter considers how altering search tasks and goals affects stopping behaviours – and whether this is reflected by what stopping strategies perform and approximate well.

[15] We leave further discussion of this finding to Section 10.2.2 on page 340.

Chapter 8

Result Diversification and Stopping Behaviour

As we've discussed previously, snippet text will typically provide a query-biased summary of a document. The provided text is used by a searcher to help him or her in satisfying their underlying information need. Often, this information need may be very *diverse,* with searchers learning about a diverse topic by issuing multiple queries to explore the topic space (Kelly et al., 2015).

What species?
Topic 347
Wildlife Extinction
What measures are being taken to prevent extinction?
What countries?

These topics or information needs are considered as *aspectual* in nature, where an underlying goal is to find out about the different facets, dimensions or *aspects*[1] of the topic. An example of different aspects within is illustrated above, showing the *wildlife extinction* topic used in this thesis. An aspect here includes for example endangered species of animal.

[1] We consider aspects in this chapter, defined as *"roughly one of many possible answers to a question which the topic in effect poses"* (Over, 1998).

While aspectual retrieval has been heavily studied in the past (most prominently during the period of the *TREC Interactive Tracks* (Over, 2001)), there has been renewed interest in this type of search task (Collins-Thompson et al., 2017). Under this context, retrieval systems are tasked with helping searchers learn more about a topic. With this goal, it makes sense to return results that are more diverse in nature, presenting the searcher with a broader view of the topic. This *should* assist searchers in learning more, and would likely lead to an improved search and learning experience (Syed and Collins-Thompson, 2017).

In this chapter, we consider how task types and the diversification of search results affect stopping behaviours. Complementing ad-hoc retrieval tasks, we introduce aspectual retrieval tasks – and compare searcher behaviours between the two. We also consider the retrieval systems that are used, allowing for comparisons in searcher behaviours when exposed to a baseline retrieval system (using BM25), and a retrieval system that diversifies the results presented to searchers (employing the *XQuAD* framework (Santos et al., 2010)).

The intuition behind the aforementioned variations in tasks and retrieval systems suggest that we will observe a difference in stopping behaviours. Under an aspectual retrieval task on a standard retrieval system, searchers will likely issue more queries as they attempt to explore the topic space – all the while stopping at relatively shallow depths. This requires a higher degree of effort on the part of the searcher. When switching to a retrieval system that diversifies results, we would expect a searcher to subsequently issue fewer queries and browse to greater depths. With this intuition demonstrating that stopping behaviours are likely to be influenced, this chapter reports on:

- a user study, exploring how diversifying results (or not) affects the performance and stopping behaviours of searchers when undertaking different search tasks, from one of ad-hoc or aspectual retrieval (Section 8.2); and
- a simulated analysis, examining how the various stopping strategies proposed in Chapter 5 perform under these conditions (Section 8.3).

In particular, we consider IFT (Pirolli and Card, 1999) to theoretically ground a number of different hypotheses relating to stopping behaviours. We begin this chapter with a discussion of prior work in the area (focusing upon aspectual retrieval), before moving to the introduction of our IFT-based experimental hypotheses.

8.1 Background, Motivation and Hypotheses

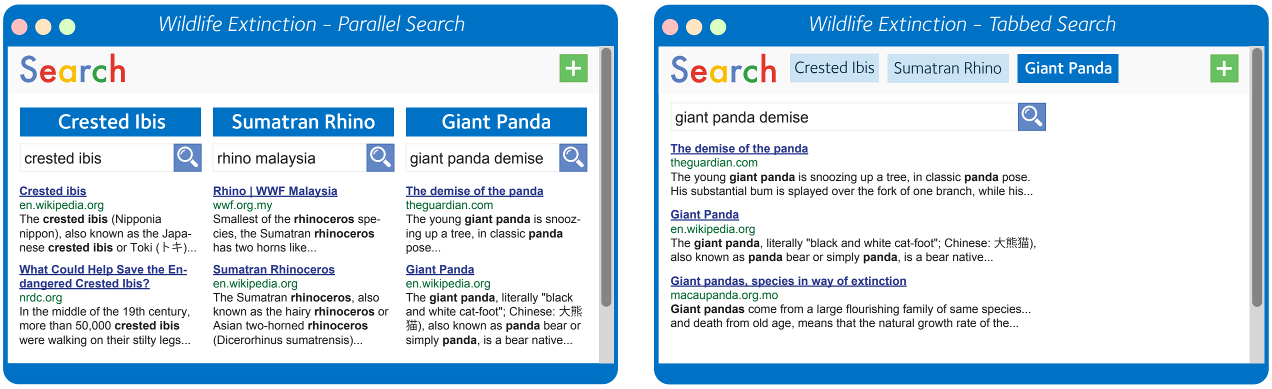
As discussed previously, a searcher will likely pose a varying number of queries (examining SERPs), and examine a number of documents (if any) before issuing a new query, or stopping their search altogether – *session level stopping* (refer to Section 4.2 on page 111). The reasons for stopping at the session level are numerous, and can occur because searchers:

- have found enough information (Prabha et al., 2007; Dostert and Kelly, 2009; Hassan et al., 2013);
- have run out of time (Zach, 2005);
- become dissatisfied with what they found (Kiseleva et al., 2015); or
- simply give up their search (Diriye et al., 2012).

Studies have shown that different factors influence search behaviours. This is demonstrated in Chapter 7, for instance, which showed how varying the length of result summaries influences behaviour. However, of particular relevance to this chapter is how different *search tasks* influence a searcher's behaviours (Kelly et al., 2015).

8.1.1 Aspectual Retrieval

An interesting search task that has not received much attention in contemporary research is *aspectual retrieval*. Aspectual retrieval is a type of search task that concerns the identification



crested ibis

Crested ibis
en.wikipedia.org
The **crested ibis** (Nipponia nippon), also known as the Japanese **crested ibis** or Toki (トキ)...

What Could Help Save the Endangered Crested Ibis?
nrdc.org
In the middle of the 19th century, more than 50,000 **crested ibis** were walking on their stilty legs...

rhino malaysia

Rhino | WWF Malaysia
wwf.org.my
Smallest of the **rhinoceros** species, the Sumatran **rhinoceros** has two horns like...

Sumatran Rhinoceros
en.wikipedia.org
The Sumatran **rhinoceros**, also known as the hairy **rhinoceros** or Asian two-horned **rhinoceros** (Dicerorhinus sumatrensis)...

giant panda demise

The demise of the panda
theguardian.com
The young **giant panda** is snoozing up a tree, in classic **panda** pose...

Giant Panda
en.wikipedia.org
The **giant panda**, literally "black and white cat-foot"; Chinese: 大熊猫), also known as **panda** bear or simply **panda**, is a bear native...

giant panda demise

The demise of the panda
theguardian.com
The young **giant panda** is snoozing up a tree, in classic **panda** pose. His substantial bum is splayed over the fork of one branch, while his...

Giant Panda
en.wikipedia.org
The **giant panda**, literally "black and white cat-foot"; Chinese: 大熊猫), also known as **panda** bear or simply **panda**, is a bear native...

Giant pandas, species in way of extinction
macaupanda.org.mo
Giant pandas come from a large flourishing family of same species... and death from old age, means that the natural growth rate of the...

Figure 8.1 Mockups of Search interfaces that consider the different *AspectBrowser* interfaces, as examined by Villa et al. (2009). Considering the *wildlife extinction* topic, the **left** illustration denotes a parallel interface, with the **right** demonstrating a tabbed interface. Refer to Villa et al. (2009) for further information on the interfaces, and how they performed.

of different *aspects* of a given topic. This task type differs from traditional ad-hoc retrieval in the sense that ad-hoc retrieval is concerned only with what constitutes a *relevant* document to a given topic, rather than identifying relevant documents and whether they are *different* to what has been previously observed.

A relevant and different document will contain unseen aspects associated with the topic in question. With a graphical example provided at the beginning of this chapter, we now provide a further example to aid understanding. Consider the topic *wildlife extinction* from the TREC 2005 Robust Track (Voorhees, 2006). In an ad-hoc search task, if the searcher manages to find several documents concerning `Pandas in China', these would all be considered relevant. However, for an aspectual retrieval task where *different* aspects must be found, the first document concerning `Pandas in China' is considered to be relevant, and other aspects (in this case, the species of endangered animal) would need to be found, such as `Sumatran Rhinos in Malaysia', `Crested Ibis in Japan', etc.

Aspectual retrieval found significant traction in *TREC Interactive Tracks* (Over, 2001) from 1997-2002. The overarching goal of these tracks was to investigate searching during an interactive search task by examining the processes involved, as well as the outcome (Over, 2001). Interaction was considered from the inaugural *TREC-1* in 1993 (Harman, 1993), where one group investigated interactive searching under the so-called *interactive query mode* while undertaking an ad-hoc task. From *TREC-6* (1997) to *TREC 2002*, a substantial volume of research was directed towards the development of systems and search interfaces that:

- assisted searchers in exploring and retrieval various aspects of a topic, such as cluster-based and faceted interfaces that explicitly showed different aspects (McDonald et al., 1998; Villa et al., 2009) (refer to Figure 8.1 for a visual example);
- provided tiles and stacks to organise documents (Hearst, 1995, 1997; Harper and Kelly, 2006; Iwata et al., 2012); and
- provided mechanisms to provide query suggestions that led to subjects following different search paths (Kato et al., 2012; Umemoto et al., 2016).

However, a disappointing conclusion from this initiative was that little difference was observed between such systems and the standard control systems (i.e. the traditional *ten blue links*, as previously discussed in this thesis) – both in terms of behaviour and performance (Voorhees and Harman, 2005).

As work shifted from aspectual retrieval to other areas, studies related to determining the intent of a searcher's query began to take hold, where the goal here was to diversify the results retrieved with respect to the original query (Rose and Levinson, 2004). Thus, this addresses the problem of *ambiguity* for short, impoverished queries. This led to a series of diversification algorithms (and intent-aware evaluation measures), changing focus from the interface to the underlying algorithms and their evaluation measures. However, while there have been numerous studies investigating the effectiveness of diversification algorithms for the problem of query intents (e.g. one query, several possible interpretations), little work has looked at studying how such algorithms apply in the context of aspectual retrieval (e.g. one topic, many aspects). This is mainly due to the fact that most of these algorithms were developed *after* the TREC Interactive Track concluded in 2002.

Recently, a growing interest in new, more complex and exploratory search tasks has taken hold. This is true within the context of *"searching as learning"* (Collins-Thompson et al., 2017). Syed and Collins-Thompson (2017) hypothesised that diversifying results presented to searchers would improve their learning efficiency. This would then be observed by a change in vocabulary expressed in their queries. This is coupled with a hypothesis of stopping behaviours, with diversifying results leading to searchers issuing more queries, and examining content to comparatively shallow depths. These hypotheses provide motivation for examining the effects of diversification when considering the task of aspectual retrieval, where a searcher needs to learn about different aspects of a topic. To ground our work, we now consider how search behaviours are likely to be changed by generating a series of hypotheses based upon IFT.

8.1.2 Tasks, Systems and Information Foraging Theory

To motivate our hypotheses for this chapter, we draw upon IFT (Pirolli and Card, 1999) and the patch model, in particular, to ground our research, and provide insights into how search behaviours may change. To recap, the patch model, as detailed in Section 3.3.1 on page 90, provides a mechanism for predicting how long foragers (searchers) will stay in a patch before moving onwards to the next. Using the established approach discussed previously – where moving between patches is akin to issuing a new query, while staying within a patch is considered as examining a SERP and any associated documents – we can then make a series of predictions as to how searchers will behave – and most importantly for this work, stop – under different experimental conditions.

These predictions are graphically illustrated in the four plots shown in Figure 8.2 – over a diversified D and non-diversified ND system, with ad-hoc AD and aspectual AS retrieval tasks.[2] Gain curves for each of the four conditions are shown. In Figure 8.2 **(a)**

[2] The system and task are combined together to produce a complete condition, such as ND AD representing a non-diversified system ND with an ad-hoc retrieval task AD.

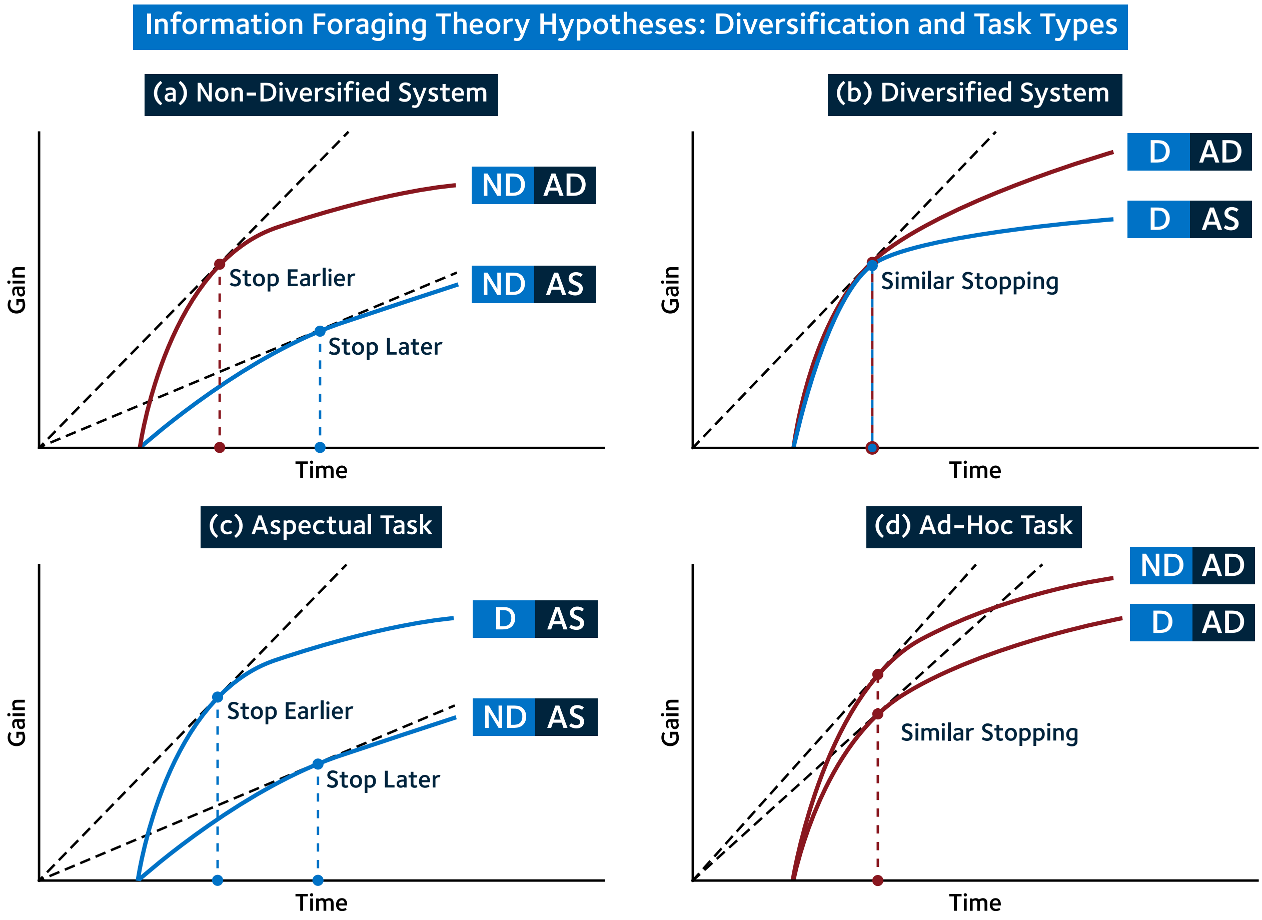


Figure 8.2 Plots of the hypotheses motivated by IFT, with each plot showing how stopping behaviour is likely to be affected when using a system that **(a)** diversifies results and **(b)** doesn't, and over **(c)** aspectual and **(d)** ad-hoc tasks. Section 8.2.1 enumerates the four different experimental conditions shown here, such as ND AD for instance.

where a non-diversified system is being used, the gain curve for the ad-hoc retrieval task is higher. This is because any relevant document would contribute to the searcher's gain. Conversely, the gain curve is lower for the aspectual retrieval task. This is because similar relevant documents that are encountered would not contribute to the overall level of gain experienced by the searcher.

From IFT, the optimal stopping point would be different between the two tasks. As we discussed in Section 3.3.1 on page 90, we can graphically find this point by drawing a line from the origin to the tangent of the gain curve. Red and blue dots indicate the optimal stopping points for ad-hoc and aspectual retrieval respectively. IFT suggests that with a non-diversified system, searchers will examine more documents per query for aspectual retrieval tasks than when compared to ad-hoc tasks.

Figure 8.2 **(b)** illustrates gain curves where a diversified system would be used, with gain curves for ad-hoc and aspectual retrieval being similar in nature. This is because the diversified system should bring relevant but different documents closer to the top of the rankings earlier. In the case of ad-hoc retrieval, these relevant (even if different) documents would still contribute to the overall level of gain. For aspectual retrieval, relevant and different documents will also contribute to the overall level of gain experienced by the searchers – up to the point where the documents become similar to the previously examined material. Therefore, IFT appears to suggest that similar stopping behaviours would be observed when searchers use a diversified search system.

Figure 8.2 **(c)** shows the predicted stopping behaviour for the aspectual retrieval task, where we have plotted the aspectual gain curves from system plots **(a)** and **(b).** Interestingly, IFT suggests that searchers will stop sooner when using the diversified system. As such, if searching for the same length of time, searchers would thus issue more queries. Finally, Figure 8.2 **(d)** shows the predicted stopping behaviour for the ad-hoc retrieval task, where again we plot the curves from the respective systems in plots **(a)** and **(b).** Note that here, the gain curve for the diversified system may be a little lower as some non-relevant but different material may bubble up the rankings. However, we expect little difference overall between the two systems, and so we hypothesise that the two levels of gain (and searcher behaviours) will approximately be the same. Consequently, IFT suggests that there will be little observable difference in terms of stopping behaviours between the two systems with ad-hoc retrieval tasks.

Therefore, we found IFT to counter our intuitions as to how searchers would behave. When using a standard, non-diversified retrieval system, our intuition suggests that since the aspectual retrieval task is rather exploratory, searchers are then more likely to issue more queries as they learn about the topic, and try to explore efforts made by different countries to protect different species. Kelly et al. (2015) for example showed that more complex search tasks required a greater number of queries. If a searcher submits a query that retrieves relevant material such as `protecting Pandas in China`, then one would expect them to only select one or two examples, rather than many more. In the case of ad-hoc topic retrieval, we intuitively expected that searchers would issue fewer queries and examine more documents. This is because they don't need to find multiple aspects. However, when using a diversified system that attempts to promote different aspects of a given topic, we would intuitively expect that the stopping behaviours of searchers using it would change. Under an aspectual retrieval task, searchers would issue fewer queries (when compared to ad-hoc tasks) and examine a greater number of documents per query.

8.1.2.1 Hypotheses

From the plots and descriptions provided above, we can formulate a number of different hypotheses relating to the expected searcher behaviours in different contexts.

Under aspectual retrieval search tasks, using a diversified system D will lead to:

- H1 fewer documents examined per query (stopping earlier); and
- H2a more queries issued; or
- H2b a decrease in the task completion time.

With ad-hoc retrieval AD tasks, diversification will lead to:

- H3 no difference in the number of documents examined (invariant stopping behaviour); and
- H4 no difference in the number of queries issued.

The contradiction between IFT and our intuitions provide an ulterior hypothesis. In addition, given the findings demonstrated by Syed and Collins-Thompson (2017), we also hypothesise that diversification will lead to a greater awareness of the topic, regardless of the task put forward, because more aspects will be encountered and found.

8.2 Diversifying Search Results

Following on from the motivation and IFT-based hypotheses outlined above, this section discusses the user study that examined the aforementioned hypotheses. As per our general user study methodology discussed previously in Section 6.2 (page 141), we conducted a within-subjects experiment. Specific details relating to this study are detailed in Section 8.2.1 below.

The primary research question for this user study is as follows.

- **DIVERSITY-RQ** How does diversification affect the search performance and stopping behaviours of searchers under ad-hoc and aspectual retrieval tasks?

This research question is addressed in tandem with the hypotheses put forward above in Section 8.1.2.1. Below, we now discuss the specific details for this user study, before discussing the results, with an emphasis on stopping behaviours, and whether or not the empirical evidence supports our hypotheses.

8.2.1 Methodology

The same basic retrieval system, document corpus and topics were used as reported in Section 6.2 (page 141).

The within-subjects study considers two key factors: the *system* and the *task.* For the system factor, our baseline control system was based upon BM25 (i.e. no diversification), and a diversified system. The details of our diversification approach are discussed in Section 8.2.1.5. For the task factor, we used the standard ad-hoc retrieval task and compared this against the aspectual retrieval task. This resulted in a 2 × 2 factorial design. Each subject who took part in the study completed four different search tasks. Each of those tasks utilised a different experimental condition, as we enumerate below. Conditions were assigned using a Latin square rotation to minimise any ordering effects. The conditions listed below are also used in Section 8.1.2 when explaining the plots supporting our hypotheses. Note that for all conditions we list below, two snippet fragments are used when generating result summaries, as per T2 in Chapter 7.

The first two conditions consider a non-diversified retrieval system ND. Our baseline, this uses BM25 as the retrieval model.

- ND AS A non-diversified system, with an aspectual retrieval task.
- ND AD A non-diversified system, with an ad-hoc retrieval task.

Our second set of conditions consider a diversified system D, using BM25 with an additional re-ranking, diversifying component. We discuss this later in Section 8.2.1.5.

- D AS A diversified system, with an aspectual retrieval task.
- D AD A diversified system, with an ad-hoc retrieval task.

With non-diversifying and diversifying systems, we developed different sets of branding for each system, each with their own distinct colour scheme, name and logo. This was to assist searchers in differentiating between the two. First, in terms of branding, we created two fictional retrieval system names:

- Hula Search, representing the non-diversified system ND ; and
- YoYo Search, representing the diversified system D .

These names were chosen as they were not associated with any major retrieval system (to the best of our knowledge), nor did they imply that one of the systems performed better than the other – both systems presented results in an identical way. Colour schemes were chosen to provide the greatest difference in visual appearance to those with colour blindness.[3] This was to ensure that subjects could later indicate which one of the two systems they preferred. Note that only the colour schemes and logos varied – the same basic interface layout as previously discussed in Section 6.2.2 (page 144) was employed. Figure 8.3 demonstrates the two different colour schemes and logos for the systems.

For the practice task, it should be noted that the standard, blue colour scheme as shown in Figure 6.2 on page 145 was used. This is the same colour scheme as used in the user study reported in Chapter 7. A standard `News Search System Study' title was also used in place of any logos. This decision was taken to remove any impact that incorporating an individual system's colour scheme in the practice task would have on searcher behaviour or perceptions. All subjects used the ND AS system and task for the practice task.

8.2.1.1 Search Tasks

As we discussed in Section 6.2.1, subjects were grounded by instructing them to imagine that they were newspaper reporters. As such, they were required to gather documents to write stories about the four topics for which they had been asked to search. Given each topic, each subject was then instructed to search while considering different search goals.

- For ad-hoc retrieval tasks, subjects were simply instructed to find documents that were *relevant* to the topic provided.

[3] Two of the more common variants of colour blindness – *protanopia* and *deuteranopia* – were both considered.

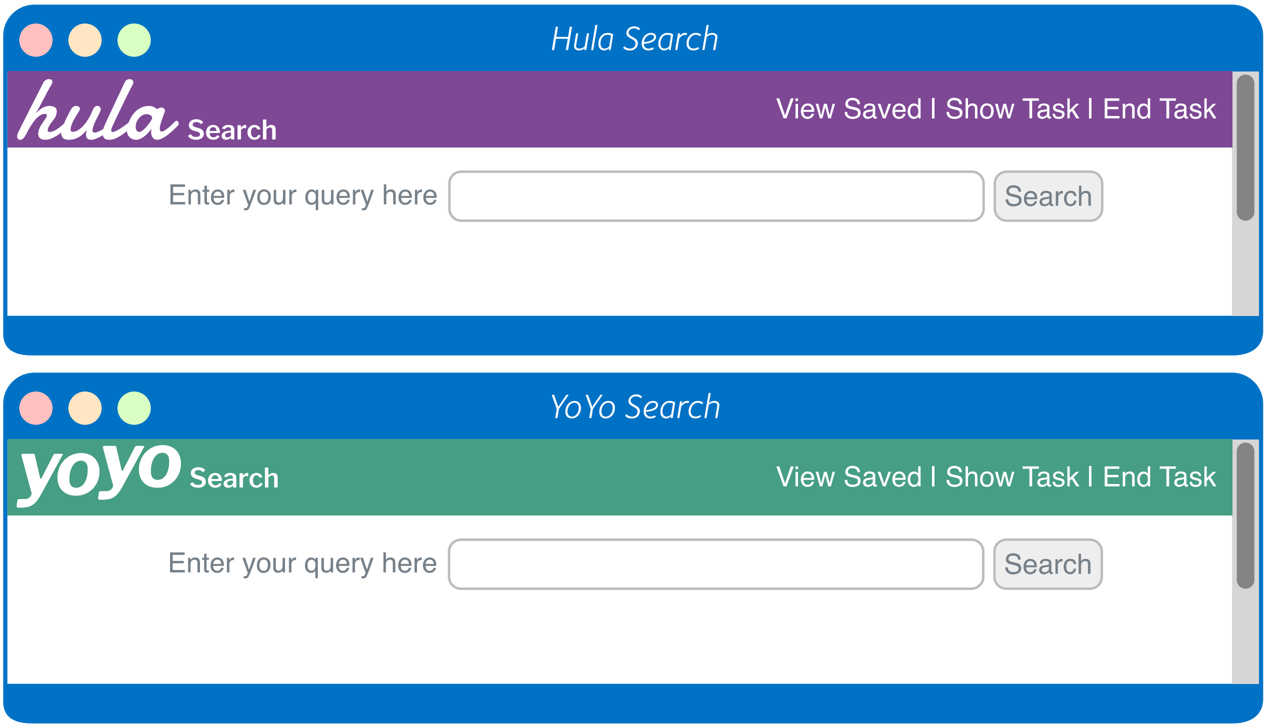


Figure 8.3 Mockups of the two interfaces used to differentiate between the two experimental systems of this user study. **Hula Search** and **YoYo Search** represented the non-diversified and diversified systems respectively. Refer to Section 8.2.1 for more information.

- For **aspectual retrieval** tasks, subjects were instructed not only to find documents that were relevant but also discussed *different aspects* of the provided topic.

For example, take the *Airport Security* topic (refer to Section 6.1.3 on page 138). Under an ad-hoc retrieval task, subjects were required to learn about the efforts taken by international airports to better screen passengers and their carry-on luggage. For aspectual retrieval tasks, subjects were also asked to find relevant documents that are different, mentioning *new airports*. Thus, subjects were explicitly instructed to find a number of examples from different airports, as opposed to a discussion of the same airport over several documents.

Task Goal Rather than imposing a session time limit (as used in Chapter 7), subjects were requested to find and save at least four novel documents which they judged to be either relevant (ad-hoc) or relevant and different (aspectual) for their given topic. (Refer to the following section for details on the reasons behind selecting this value.) As such, subjects had the liberty to end the search task when they chose to do so by selecting the End Task option at the top right of the search interface – refer to Figure 8.3 for examples of this. This is in direct contrast to the user study reported in Chapter 7, where the End Task option was not present – the ten minute limit dictated when their search session ended.

8.2.1.2 Crowdsourced Subjects and Controls

Subjects undertaking the user study were informed that from a small-scale pilot study, it would take approximately 7-10 minutes of their time to find at least four useful documents per task. Combining everything together, this meant that the entire experiment would take approximately 40-50 minutes. Since we did not impose any time constraints on how long subjects searched, we instead established an accuracy-based control. We informed subjects that their accuracy in identifying useful material would be examined, and that they were required to find four useful documents with at least 50% accuracy (based upon TREC relevance judgements as the gold standard). Using data from the prior user study reported in Chapter 7, the accuracy of those subjects was between 25% and 40% on average, depending upon the topic. While we stipulated a higher accuracy, this was to motivate subjects to work in a diligent manner.

In all, a total of 64 subjects performed the experiments that complied with the MTurk recruiting constraints imposed, as we outlined in Section 6.2.4 on page 147. However, a total of 13 were omitted from this population because they either:

- failed to complete all the search tasks (a total of five subjects were removed);
- failed to mark at least four documents (two subjects); or
- spent less than two minutes per task, and failed to retrieve any relevant documents (six subjects).

Of the 51 subjects who successfully completed the experiment, 26 females and 25 males participated. The average age of the subjects was 38.66 years ($min = 20$; $max = 71$; $stdev = 11.43$). In addition to these basic demographics, a total of 22 subjects reported possessing a bachelor's degree or higher, with the remaining 29 possessing an associate's degree or lower. All subjects bar one expressed a preference to Google as their everyday retrieval system of choice. All subjects indicated that they conducted many searches for information via a retrieval system per week.

8.2.1.3 Extracting Aspects

For each topic, we used the corresponding TREC QRELs derived from the 2005 Robust Track (Voorhees, 2006). However, to assess how many aspects were retrieved by subjects, we needed to commission additional labels as existing labels were not available for all the selected topics. First, for each topic, we examined the topic descriptions to identify what dimensions could be considered aspects of the topic. We noted that for each topic, there were at least two ways this could be achieved: *entity-* or *narrative-based*. For example, a useful document within the *Curbing Population Growth* topic could either state the country in which measures were taken (entity-based) or a description of the actual measure used to reduce population growth (narrative-based).

For this study, it was decided that we should focus on entity-based aspects. This decision was taken as *different narratives* were subject to greater interpretation than *different entities* – it is easier to identify from a document that China, for example, is the country being discussed, rather than the measures the country took – and their effects. For each TREC relevant document across the five topics considered, the author and his supervisor manually extracted the different aspects for each, with higher agreement (95% vs. 67%) between them across entity-based aspects. Both the entity- and narrative-based approaches for each of the five topics are shown in Table 8.1.

Table 8.1 A list of the different entity- and narrative-based approaches trialled during the aspect extraction process. As discussed in Section 8.2.1.3, the entity-based approach was carried forward for this study with a higher agreement rate between assessors.

	Entity	Narrative
Airport Security	Airports	Security measures taken
Wildlife Extinction	Species	Protection and conservation measures
Piracy	Vessels boarded	Acts of piracy
Tropical Storms	Storms	Lives lost, destruction caused
Curb. Pop. Growth	Countries	Population control methods

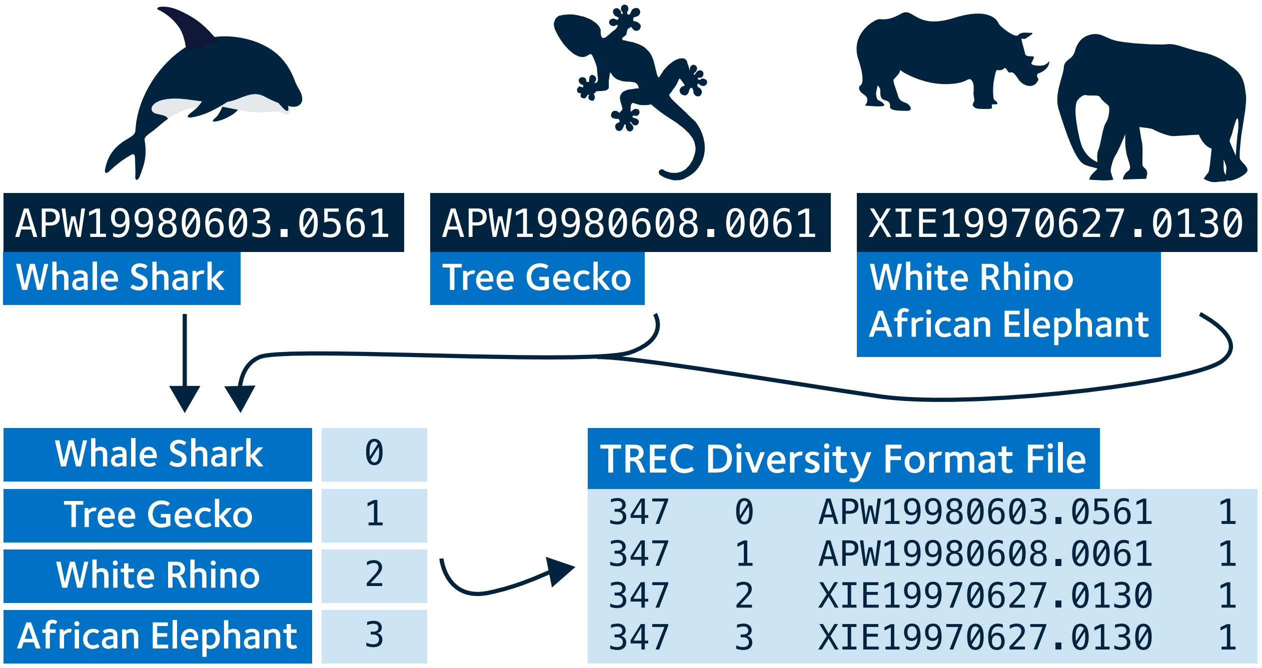
To complement Table 8.1, we also list below a number of different example entity-based aspects that were extracted for each of the five topics. The number provided with the topic title denotes the number of individual aspects that were extracted for a specific topic.

- **Airport Security** **14 unique aspects** Considering different *airports* in which additional security measures were taken, examples include *John F. Kennedy International Airport, Boston Logan International Airport,* or *Leonardo da Vinci International Airport.*

- **Wildlife Extinction** **168 unique aspects** Considering different *species of endangered animals* under protection by states around the world, such as the *golden monkey, Javan Rhino,* or *Manchurian tiger.*

- **Piracy** **18 unique aspects** Considers different *vessels* that were either boarded or hijacked, such as the *Petro Ranger, Achille Lauro* or *Global Mars.*

- **Tropical Storms** **43 unique aspects** Considers different *tropical storms* where individuals were killed, and/or there was major damage, such as *Hurricane Mitch, Typhoon Linda* or *Tropical Storm Frances.*



Whale Shark	0
Tree Gecko	1
White Rhino	2
African Elephant	3

TREC Diversity Format File

```
347   0   APW19980603.0561   1
347   1   APW19980608.0061   1
347   2   XIE19970627.0130   1
347   3   XIE19970627.0130   1
```

Figure 8.4 Illustration of the process used to create a TREC diversity format file. The identified aspects are assigned a unique identifier per topic. With the diversity format file, we could then use tools such as `ndeval` to compute diversity-based measures.

- Curbing Population Growth 26 unique aspects Considers different *countries* where population control methods were employed, such as *China, India* or *Zimbabwe.*

Each of these unique aspects was assigned an identifying number, and stored in the TREC diversity format. By storing the aspects in this format, we could then use existing evaluation tools, such as `ndeval`[4]. This tool was used to compute a number of measures related to aspectual retrieval. This process is illustrated in Figure 8.4.

8.2.1.4 Additional Performance Measures

In conjunction with the standard performance measures that we discussed in Section 6.3.3 on page 155, we also include for this chapter two measures allowing for the examination

[4] The `ndeval` source code can be acquired from the TREC website at `https://trec.nist.gov/data/web/10/ndeval.c`. LA *2018-06-24*

of searcher and system performance regarding the entity-based aspects. While traditional measures consider what documents are relevant, these additional measures allow us to determine *why* said documents are relevant (i.e. what aspects each document covers).

The first measure we consider is Aspectual Recall (AR). Defined by Over (1998), AR was introduced as part of the TREC-6 campaign. It was defined as:

> *"...the fraction of the submitted documents which contain one or more aspects."*
>
> Over (2001)

3
NEW
1
NEW
0
NEW
AR@3
=1.33

Given a ranking, aspectual recall can be therefore computed by summing the number of *unseen aspects* regarding a given topic up to some depth k, and dividing by the rank. This is in contrast to more simplistic relevance measures that consider only the TREC relevance judgement score for a document and topic combination. An example is provided above: given three documents, with three, one and zero new aspects to a topic, the aspectual recall at rank 3 is therefore $(3 + 1 + 0)/3 = 1.33$.

The second measure that considers the diversity of the results returned is αDCG. A *Cumulative Gain (CG)*-based approach, we discussed CG basics in Section 2.4.1.4. An extension of *Discounted Cumulative Gain (DCG)* (Järvelin and Kekäläinen, 2002), α*DCG* employs a position-based searcher model (Clarke et al., 2008). The measure takes into account the position at which a document is ranked, along with the aspect(s) mentioned within the documents. α*DCG* ranks by rewarding newly-found aspects, and penalising redundant aspects geometrically, discounting all rewards with a discounting rank function. As the name of the measure might imply, α is a tuneable parameter, controlling the severity of redundancy penalisation. As used in prior TREC experimentation, we used $\alpha = 0.5$ for all reporting of α*DCG* in this chapter.

8.2.1.5 Diversifying Search Results

As discussed earlier, our system factor considered both a baseline BM25 retrieval system and a diversified approach, again using BM25 as an initial ranking baseline. The algorithm that we employed, based upon the *XQuAD* framework by Santos et al. (2010), re-scores and subsequently re-ranks documents based upon the number of unseen entities that appear within the document. The algorithm is presented as pseudo-code in Algorithm 8.5 below. Essentially, documents are re-ranked according to the number of new entities that are contained within them, with w determining the weighting of the aspectual scoring component.

In order to select a reasonable approximation for the algorithm's weighting, we performed a pilot study running the diversification algorithm over the set of 715 queries that were issued by subjects of the user study reported in Chapter 7. Results of the pilot study are presented in Table 8.2. As can be seen from the table, we explored a range of cutoff (**k**) and weighting (**w**) values, with $10-50$ trialled for k and $0.1-1.0$ trialled for w. We selected $k = 30, w = 0.7$ as this combination provided the best results ($AR@10 = 6.61$, $\alpha DCG = 0.075$, $P@10 = 0.36$) in terms of performance and efficiency. A higher k for example only slightly increased performance but took longer to compute. Indeed, $k = 30$ was deemed to be a sensible choice as subjects from the prior user study didn't go lower than a depth of 24 on average over interface T0.

For the diversity re-ranking to work in this scenario, the algorithm must be aware of the ground truths which were collated, as described in Section 8.2.1.3 above. The reasons for following this approach were:

- not having to invest a significant amount of effort into tuning a different diversification algorithm to return acceptable results; and
- that it would guarantee that TREC relevant documents, containing different entities, would bubble up to the top of the rankings, increasing the effect (and hopefully the subject's observation) that the results were indeed ranked differently.

Input Parameters

- Original ranking, `existingResults`
- Diversification depth, set in this study to `k=30`
- `w(=0.7)`, weighting for diversification scoring component

Output

- Manipulated array of results, diversified to depth `k` (see above)

Helper Functions

- `getEntities(x,y,z)` Given an array of results (documents), returns an array of entities present in the results array from range `y` to `z`
- `getLength(x)` Returns the length of array `x`
- `getUnseenEntities(x,y)` Returns entities in document `x` that have not yet been observed in ranked document array `y`
- `sortByScore(x)` Sorts document array `x` by `score` in descending order
- `<array>.pop()` Removes the top entry from an array, returning the popped value

```
SET entities TO []
SET newRankings TO []
SET i TO 1

# Take the top result from the baseline results, popping results
SET newRankings[0] TO existingResults.pop()

WHILE i <= k DO
    # Obtain all entities from the first to ith result
    SET entities TO getEntities(existingResults, 0, i-1)
    SET j TO 0

    # Now rescore all remaining results, considering weighting w
    WHILE j <= getLength(existingResults) DO
        SET newEntityCount TO
                getUnseenEntities(document, existingResults)
        SET existingResults[j].score TO score + (w·newEntityCount)
        SET j TO j + 1
    END WHILE

    # Reorder existingResults; move top result to new array
    sortByScore(existingResults)
    SET newRankings[i] TO existingResults.pop()
    SET i TO i + 1
END WHILE
```

Figure/Algorithm 8.5 Pseudo-code of the diversification algorithm used in this study, based on the XQuAD framework by Santos et al. (2010). As described in Section 8.2.1.5, the algorithm guarantees that TREC relevant documents containing different aspects from each other will bubble up the baseline (BM25) rankings. Pseudo-code is provided in *HAGGIS* (Cutts et al., 2014).

Table 8.2 Table illustrating the effects of varying the diversification weighting parameter, **w**, and diversification cutoff **k** when using the diversification algorithm as discussed in Section 8.2.1.5. Values in the table represent the aspectual recall in the top 10 documents *(AR@10)* after re-ranking, on average, over the 715 queries issued by subjects of the user study reported in Chapter 7. At *w=0.0*, diversification is not applied — this configuration therefore enjoys the same performance as our baseline, non-diversified system **ND**, utilising BM25 *(b=0.75)*. Cells highlighted denote the selected configurations for systems **ND** and **D**.

	Cutoff Range (k)				
Weighting Parameter (w)	**10**	**20**	**30**	**40**	**50**
0.0 (ND)	3.64				
0.1	3.64	4.94	5.51	5.95	6.37
0.3	6.58	6.58	6.64	6.59	6.59
0.5	6.58	6.58	6.58	5.58	6.58
0.7 (D)	6.56	6.56	6.61	6.51	6.60
0.9	6.52	6.52	6.61	6.57	6.63
1.0	6.63	6.63	6.59	6.61	6.56

Without such a ground truth based approach, ensuring that TREC relevant documents would bubble up would have been difficult to achieve. Given the effects of document pooling as part of how the TREC QRELs were created (refer to Section 2.3.1.1), it is highly likely that many other documents exist within the corpus that could be considered to be useful to a given topic, but were not assessed.

One major pitfall of this approach is that the diversification algorithm employed would have provided results that were *too good*, thus not presenting much of a challenge to subjects. To mitigate this issue, we also included documents that were considered non-relevant by TREC assessors (i.e. a TREC assessment of `0`) when performing re-ranking. Rather than always bubbling up relevant documents discussing new aspects, the inclusion of these documents would also mean the bubbling up of *non-relevant* documents that ultimately mention one or more aspects.[5] From this, an additional 2, 663 documents that were not relevant were included within the diversity re-ranking ground truths. For the five topics we considered, the number of non-relevant documents from the TREC QRELs over each topic are reported in Table 6.1 on page 140.

Rather than manually assess each non-relevant document, we took the list of entities that were discovered for TREC relevant documents, and performed an exact keyword search for each of the entities within each of the 2, 663 documents. Any matches would have the corresponding entity attached to the document.

8.2.1.6 Post-Task Surveys

With the pre-task survey the same as that outlined in the general methodology (refer to Section 6.3.4 on page 155), questions for this study differed only over post-task and post-experiment surveys. Here, we discuss the questions posed in each of the four post-task surveys.

On the completion of each of the four search tasks, subjects were asked to answer questions that were split into two broad categories, examining:

- their perceived behaviours when interacting; and
- how they felt the retrieval system they used had performed.

Answers were compulsory; we provided a seven-point Likert scale for responses, providing the ability to give a neutral response, as well as strong disagreement *(1)* or agreement *(7)* with the questions that were asked.

[5] For example, a non-relevant document for the *wildlife extinction* topic may have discussed a species of animal, but would not have discussed approaches used to correct its endangered status.

Considering the subject's behaviours, we asked their opinions on the following areas.

- **Success** How successful they thought they were at completing the given search task.
- **Subject Speed** How quickly subjects felt that they completed the search task.
- **Queries** Whether the subjects issued different queries to explore the topic.
- **Documents** If they only examined a small number of documents per query.
- **Checks** Whether they checked each document carefully before saving.
- **Enough** Whether the subjects saved more documents than was required (remembering that subjects were instructed to save at least four per task).

In addition to the behavioural component of the survey, the system-sided component of the survey asked an additional six questions, again using a seven-point Likert scale. The questions posed to the subjects are enumerated below.

- **System Speed** How well subjects thought the system helped them complete the given search task quickly.
- **Difficulty** Whether they felt the system made things difficult to find useful information.
- **Ease** If the system made it easy for subjects to complete the given search task.
- **Happiness** Whether the subjects were happy or not with how the system performed.
- **Cumbersome** Whether they felt the system was cumbersome to use or not.
- **Confidence** How confident the subjects were in the decisions that they had taken.

8.2.1.7 Post-Experiment Survey

In addition to the post-task surveys, we also asked subjects to answer a post-experiment survey upon completion of all four search tasks. Here, we wanted to ascertain which of the two retrieval systems (Hula Search, representing the baseline non-diversified system ND, and YoYo Search Search, representing the diversified system D) offered subjects a better experience, and which one of the two they preferred overall.

Seven questions were posed, with answers again provided on a Likert scale. However, this time we provided six possible choices, from 1 (definitely Hula Search) to 3 (slightly Hula Search), from 4 (slightly YoYo Search) to 6 (definitely YoYo Search). We opted not to include a neutral option to force subjects into deciding between one of the two systems.

- Informative Which one of the two retrieval systems returned the most informative results?
- Unhelpful What one of the two retrieval systems was more unhelpful?
- Easiest Of the two retrieval systems, what one was easier to use?
- Least Useful Which retrieval system was less useful?

The final three questions then asked subjects about what one of the two systems they felt yielded the most relevant and diverse content.

- Most Relevant Which of the two retrieval systems yielded more relevant information?
- Most Diverse Which of the two retrieval systems offered the more diverse set of results?
- Most Preferable Which retrieval system did you prefer overall?

We discuss the results from this survey in Section 8.2.2.4.

8.2.2 Results

We now move onto an analysis of the user study results, addressing the overarching study research question DIVERSITY-RQ, and the five hypotheses posed in Section 8.1.2.1. In this analysis, we examine both the behaviour and performance of subjects across the four different experimental conditions, D AS, ND AS, D AD and ND AD. Both task (considering AD vs. AS) and system (considering ND and D) effects were also examined. To evaluate these data, ANOVAs were conducted using the experimental conditions, tasks and systems each as factors; main effects were examined with $\alpha = 0.05$. Bonferroni tests were then used for post-hoc analysis. To reiterate, αDCG was computed at $\alpha = 0.5$.

To begin our analysis, we first examined whether the performance demonstrated by subjects over the two retrieval systems was in fact different – as indicated it would be by our pilot study (refer to Section 8.2.1.5). We took the queries subjects issued to each of the two systems and measured the performance according to αDCG, AR and precision. Results are presented in Table 8.3. Statistical testing confirms that the two systems were significantly different in terms of diversity (i.e. αDCG@10: $F(1, 1272 = 28.74, p < 0.001)$ and AR@10: $F(1, 1272 = 55.43, p < 0.001)$). However, P@10 was not significantly different between the two retrieval systems. This suggests that the re-ranking promoted relevant and diverse documents, mostly from the top ten results (on average).

Aside from showing query performance, Table 8.3 also reports the number of terms issued per query over retrieval systems ND and D. Of the 1273 total queries issued, those issued to ND were shorter on average, with 3.59 terms compared to 3.80 terms for system D. However, the vocabulary used by subjects issuing queries to ND was more diverse than D – queries issued to ND contained a total of 345 unique terms compared to 292. This provides our first finding of note from the interaction data. When using retrieval system ND that did not diversify search results, subjects issued a greater number of queries – but with slightly shorter and more varied queries (in terms of the vocabulary used) – in order to accomplish their tasks.

Table 8.3 Query statistics and performance measures across experimental systems ND (baseline, non-diversified) and D (diversified). Note the significant differences between the diversity-centric measures, α**DCG** (where α=0.5) and aspectual recall (**AR**), demonstrating that the diversification algorithm did indeed provide subjects with a more diverse set of results with which to examine. Highlighted cells denote a significant difference between systems.

		ND	D
	Queries Issued	718	555
	Terms per Query	3.59	3.80
	Unique Terms	345	292
Precision	P@5	0.25±0.01	0.29±0.01
	P@10	0.22±0.01	0.24±0.01
αDCG	αDCG@5	0.02±0.00	0.04±0.00
	αDCG@10	0.03±0.00	0.04±0.00
AR	AR@5	1.40±0.11	3.39±0.21
	AR@10	2.11±0.14	4.07±0.24

8.2.2.1 Interaction Measures

Firstly, we examine the different interactions between searchers and the retrieval systems. Tables 8.4 and 8.5 both present the mean (and standard deviations) of:

- the number of queries issued (**#Queries**);
- the number of SERPs that were examined by subjects per query (**#SERPs/Query**);
- the number of documents examined (clicked) per query (**#Docs./Query**); and
- the click depth (or stopping depth) per query (**Depth/Query**).

These are reported in the **Interactions** category in both tables. Table 8.4 reports over the four different system and task combinations trialled, while Table 8.5 reports over each individual system and task. ANOVAs revealed no effects across conditions, systems or tasks. However, there are trends that are worth discussing. Firstly, we notice that when subjects used system D to complete the aspectual retrieval task, they examined fewer documents per query than when completing the same task on system ND (12.85 ± 1.49 vs. 15.73 ± 1.45) – which is in line with H1. We also observed that subjects issued slightly more queries on D compared to ND under the aspectual retrieval task (5.92 ± 0.88 vs. 5.25 ± 0.80). This is in line with H2a – these results, however, were again not statistically significant.

Now we turn our attention to the ad-hoc retrieval tasks. Our hypotheses claimed that there would be no differences in terms of the number of documents examined (H3) or in the number of queries issued (H4) – which was the case. However, we note that subjects using D examined more results than when using ND (16.19 ± 2.14 vs. 13.94 ± 1.93), and they issued slightly fewer queries (4.96 ± 0.74 vs. 5.20 ± 0.69). We can see the trade-offs between queries and the number of results inspected per query, where more queries tend to lead to fewer results being examined, and vice versa. This result suggests that subjects, when searching using diversified system D, under an ad-hoc task, may have had to examine to greater depths to find more relevant material due to the system's performance. Alternatively, this trend could be explained by suggesting that the system encouraged subjects to go deeper, something that we intuitively expected when subjects were searching for aspectual, diversified information. Either way, no conclusive evidence to support our hypotheses exists with statistically significant differences – merely trends.

8.2.2.2 Performance Measures

Tables 8.4 and 8.5 also report a number of different performance measures, reported within the **Performance** grouping. Included in these tables are:

- the number of saved documents (**#Saved**); also broken down into:

Table 8.4 Behavioural (including interaction and time-based) and performance measures, across each of the experimental conditions D AS, ND AS, D AD and ND AD. Cells that are highlighted denote statistically significant differences between conditions.

		D-AS	ND-AS	D-AD	ND-AD
Interactions	**#Queries**	5.92±0.88	5.25±0.80	4.96±0.74	5.20±0.69
	#SERPs/Query	1.78±0.14	2.42±0.24	2.28±0.31	2.28±0.20
	#Docs./Query	3.02±0.39	3.65±0.46	3.48±0.51	3.23±0.37
	Depth/Query	12.85±1.49	15.73±1.45	16.19±2.14	13.94±1.93
Performance	**#Saved**	5.80±0.26	5.96±0.25	5.92±0.25	5.78±0.20
	#TREC Saved (iP)	2.63±0.22	2.18±0.23	2.51±0.23	2.22±0.22
	#TREC Non.	1.75±0.22	1.96±0.23	1.37±0.22	1.82±0.23
	#Ent. Found	**7.22±0.94**	**4.31±0.60**	5.82±0.77	**4.37±0.59**
	#Docs. New Ent.	**3.20±0.21**	**2.35±0.20**	2.63±0.23	**2.02±0.18**
Times	**Total Session**	443.65±45.05	430.50±38.39	432.18±49.87	447.55±47.82
	Per Query	8.80±0.89	9.99±1.21	9.69±0.79	8.69±0.57
	Per Document	15.97±1.96	13.03±1.01	13.66±1.02	15.09±2.20
	Per Summary	1.59±0.09	1.75±0.15	1.71±0.11	1.71±0.13

  - the number that were TREC relevant, or interactive precision (**#TREC Saved (iP)**); and
  - the number that were TREC non-relevant (**#TREC Non.**);

- the number of new entities that were found (within saved documents, with new entities being in the context of a search session) (**#Ent. Found**); and

Table 8.5 Behavioural (including interaction and time-based) and performance measures, across the two experimental systems **ND** and **D**, as well as the two tasks, **AD** and **AS**. Cells that are highlighted denote statistically significant differences between conditions.

		ND	D	AD	AS
Interactions	#Queries	5.23±0.53	5.44±0.58	5.08±0.51	5.59±0.59
	#SERPs/Query	2.35±0.16	2.03±0.17	2.28±0.18	2.10±0.14
	#Docs/Query	3.44±0.29	3.25±0.32	3.36±0.31	3.34±0.30
	Depth/Query	14.84±1.58	14.52±1.31	15.07±1.44	14.29±1.47
Performance	#Saved	5.87±0.16	5.86±0.18	5.85±0.16	5.88±0.18
	#TREC Saved (iP)	2.20±0.16	2.57±0.16	2.36±0.16	2.40±0.16
	#TREC Non.	1.89±0.16	1.56±0.16	1.60±0.16	1.85±0.16
	#Ent. Found	4.34±0.42	6.52±0.61	5.10±0.49	5.76±0.57
	#Docs. New Ent.	2.19±0.13	2.91±0.16	2.32±0.15	2.77±0.15
Times	Total Session	439.02±30.52	437.91±33.44	439.86±34.38	437.08±29.45
	Per Query	9.34±0.67	9.25±0.59	9.19±0.49	9.39±0.75
	Per Document	14.06±1.21	14.81±1.10	14.37±1.21	14.50±1.11
	Per Summary	1.73±0.10	1.65±0.07	1.71±0.08	1.67±0.09

- the number of documents containing at least one new entity (**#Docs. New Ent.**).

In terms of the number of documents saved, there were no significant differences between conditions, systems or tasks. On average, subjects saved around six documents on average, which was two more than the minimum goal of four. This suggests that subjects wanted to make sure that they found a few extra, potentially useful documents.

Table 8.6 Interaction probabilities, as observed over the four experimental conditions. Cells that are highlighted denote statistically significant differences between conditions. Refer to Section 6.4.2.3 on page 167 for an explanation of the different probabilities listed here.

		D-AS	ND-AS	D-AD	ND-AD
Click	P(C)	0.16±0.01	0.21±0.02	0.16±0.01	0.20±0.01
	P(C|R)	0.27±0.03	0.30±0.04	0.25±0.03	0.31±0.04
	P(C|N)	0.13±0.02	0.18±0.02	0.13±0.01	0.17±0.02
Save	P(S)	0.67±0.03	0.66±0.03	0.70±0.03	0.71±0.04
	P(S|R)	0.78±0.04	0.63±0.05	0.74±0.04	0.67±0.05
	P(S|N)	0.59±0.04	0.61±0.04	0.65±0.04	0.65±0.04

When we turn our attention to the entity-related measures, we note that subjects found more documents that contained new entities, and found more new entities overall when using the diversified system D. This was statistically significant (6.52 ± 0.61 compared to 4.34 ± 0.42 for systems D and ND respectively, where $F(1, 203 = 8.70, p < 0.05)$). When examining each condition, the Bonferroni follow-up test showed significant differences between conditions D AS and conditions D AD and ND AD, where $F(3, 203 = 3.49, p < 0.05)$. We also noticed that subjects found more documents with new entities, and thus more entities generally, for task D AD than when using system ND (documents with new entities: 2.63 ± 0.23 vs. 2.02 ± 0.18, new entities: 5.82 ± 0.77 vs. 4.37 ± 0.59). Though this was not significantly different, it does suggest that when subjects used system D, they did learn more about the different aspects of the given topic (or at least encountered more aspects) than when using system ND that did not diversify results.

Table 8.6 reports interaction probabilities associated with searcher interactions, the details of which are discussed in Section 6.4.2.3 on page 167. From the table, we can see that there was a significant difference between conditions for the probability of clicking on a result summary link, and the probability of clicking on TREC non-relevant items. Comparing systems indicated that subjects clicked more when using the non-diversified system, and clicked on more non-relevant documents. However, we did not observe any task effects and thus do not report these measures here. This suggests that the non-diversified system ND led to subjects examining more documents, but often more non-relevant documents. This is reflected by the fact that across all the performance measures, subjects when using system ND, performed worse.

8.2.2.3 Time-Based Measures

Tables 8.4 and 8.5 also report a third grouping of results, showing a series of times recorded for various interactions. These are all reported within the **Times** grouping. Across both tables (conditions, systems and tasks), we report:

- the mean total session time (denoted as from the first query focus to ending the task, **Total Session**);
- the mean time spent entering queries (**Per Query**);
- the mean per document examination time (**Per Document**); and
- the mean time spent examining an individual result summary (**Per Summary**).

All values in the two tables for time-based measures are reported in seconds. Surprisingly, no significant differences were found between any of the comparisons over the total session times, the per query times, the per document times, and the individual result summary examination times. However, results do show a relatively constant mean session time over each of the four experimental conditions, as shown in Table 8.4. At ≈ 438.5 seconds, this is around seven minutes on average – in line with the time taken to find four documents in the previous user study reported in Chapter 7.

Considering H2b, no evidence was found to support that task completion times were lower under the diversifying retrieval system with an aspectual retrieval task D AS. From Table 8.5, we can see that subjects in actuality spent slightly longer on the task, with 443 seconds reported for D vs. 430 seconds for ND – essentially, the difference of examining approximately one document on average.

8.2.2.4 User Experience Measures

There were no significant differences between conditions, tasks, or systems for any of the post-task surveys. For the post-experiment survey, subjects were roughly evenly split between their preference for system D or ND – again with no significant differences. This finding suggests that despite the substantial (and significant) difference in aspectual recall and other system performance measures between the systems, subjects seemed largely unaware of the influence of the two systems. However, their observed behaviours do suggest that the system (and task) did affect their performance, as Table 8.4 demonstrates.

Post-Task Surveys Table 8.7 provides the results of the post-task surveys. Questions were provided in Section 8.2.1.6. To recap, a seven-point Likert scale was used for all responses, ranging from 1 (strongly disagree) to 7 (strongly agree). Turning our attention first to the **Behavioural** survey results, we observed no significant differences across conditions, systems or tasks. However, across all conditions, systems and tasks, subjects broadly agreed with the statements that they were presented with, suggesting that they felt successful in completing the search tasks, and were able to complete them quickly. All were in agreement that they carefully checked their documents for usefulness (i.e. relevance and/or new entities, depending upon the task) before saving, but were in broad disagreement that they had examined a *few* documents per query, indicating that across all conditions, subjects felt as though they had examined more than they felt they needed to (or more than the requested minimum). A positive sentiment to this question was recorded across 68.2% of all logged search sessions, with 23% and 8.33% for negative and neutral sentiments respectively.

Table 8.7 Results from the post-task surveys, consisting of both the behavioural- and system-based questions. Results shown are averages recorded for each of the four experimental conditions when considering the seven-point Likert scale. Significant differences are highlighted.

		D-AS	ND-AS	D-AD	ND-AD
Behavioural	**Success**	5.90	5.53	5.98	5.98
	Subject Speed	4.24	4.33	4.61	4.45
	Queries	5.75	5.35	5.24	5.47
	Documents	2.78	3.00	2.67	2.69
	Checks	6.08	6.10	6.14	6.02
	Enough	5.00	5.06	4.84	5.43
System	**System Speed**	4.55	4.16	4.84	4.42
	Difficulty	3.78	4.20	3.31	3.38
	Ease	4.53	4.00	4.47	4.32
	Happiness	4.45	4.18	4.73	4.46
	Cumbersome	**3.31**	**3.50**	3.18	3.00
	Confidence	5.25	5.04	5.63	5.36

Regarding the **System**-sided survey questions presented in Table 8.7, subjects considered that both systems offered a reasonably quick and straightforward approach to finding results, with a generally positive outcome for both. The systems generally did not appear to be considered cumbersome to use, and subjects did not find the system made it overly difficult to complete the tasks – a significant difference existed between the two tasks for this question, where $F(1, 201 = 5.51, p < 0.05)$. Overall, subjects felt happy with how the system performed and had some confidence in their decisions.

Table 8.8 Raw results of the post-experiment survey. Values denote subjects who selected an answer (columns) for each question (rows). The lower the value, the stronger the preference to YoYo Search; the higher the value, the stronger the preference to Hula Search.

	Preference to YoYo (D)			Preference to Hula Search (ND)		
	1	**2**	**3**	**4**	**5**	**6**
Most Informative	5	0	20	20	0	6
Most Unhelpful	7	0	21	17	0	6
Easiest	7	0	18	20	0	6
Least Useful	6	0	20	18	0	7
Most Relevant	10	0	16	18	0	7
Most Diverse	8	0	18	19	0	6
Most Preferable	9	0	16	18	0	8

Post-Experiment Survey Upon finishing the experiment, subjects completed the post-experiment survey as detailed in Section 8.2.1.7. Here, we asked subjects which system they preferred (from either D or ND) over a number of different questions. Results from the survey, as shown in Table 8.8, provide a mixed picture – neither system was favoured by the subjects, with all questions recording a near 50-50 split. This is an interesting finding, as results – especially from Table 8.3 on page 266 – showed that there was a significant difference between the two systems. Subjects simply had difficulty attempting to determine which system was more attractive to use.

8.2.2.5 Gain over Time

Back in Section 8.1.2, we motivated this study – and indeed the wider work reported in this chapter – using IFT, where we constructed a number of gain curves reflecting our beliefs about how the search performance experienced by subjects would look on each of the four combinations of system and task. This was done in order to generate our hypotheses outlined in Section 8.1.2.1. In this final section of the user study results, we examined how subjects performed over time for each of the experimental conditions trialled, allowing us to infer the gain curves. We then compare each of the curves generated with our initial expectations, shown in Figure 8.2 on page 247.

To create empirical gain curves, we plotted *Cumulative Gain (CG)* against time, where gain was defined to be either:

- the number of saved relevant documents under ad-hoc retrieval tasks; or
- the number of saved, relevant and different documents when undertaking an aspectual retrieval task.

These definitions are what constituted as a useful document for both of the tasks, defined previously in Section 8.2.1.1. As both of these definitions can be expressed in the same units, they can be also plotted on the same axes.

In parallel with expectation plots shown in Figure 8.2 on page 247, Figure 8.6 plots the corresponding *empirical gain curves* for:

(a) the non-diversified system, ND, over both search tasks AD and AS;

(b) the diversified system, D, over both search tasks AD and AS;

(c) the aspectual search task AS for both retrieval systems; and

(d) the ad-hoc task AD for both retrieval systems.

Compared to our expectations in Figure 8.2 on page 247, on visual inspection, we see that our predictions were roughly in line with the average levels of CG experienced by the subjects. With Figure 8.6 **(a)** for example, we hypothesised that using retrieval system ND,

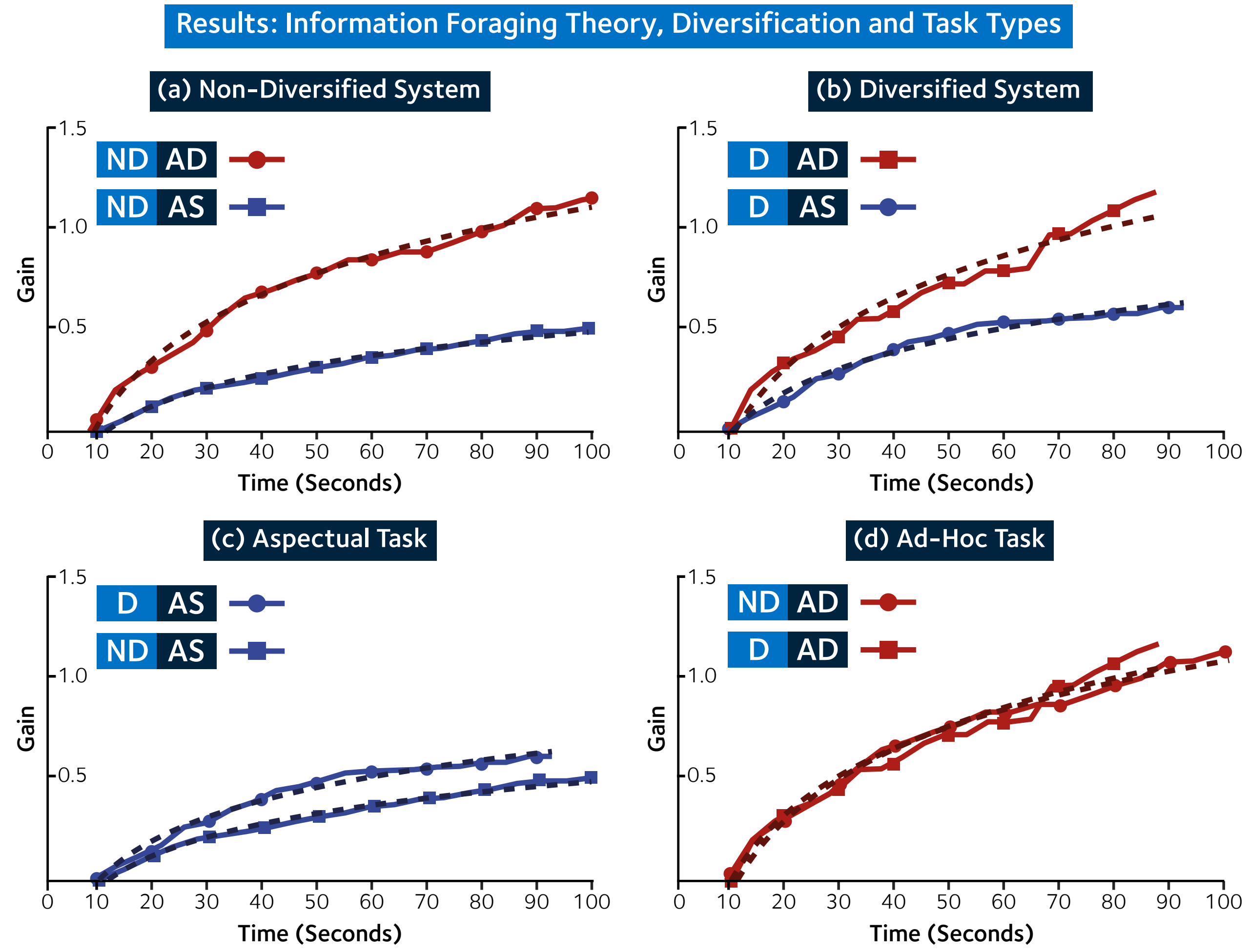


Figure 8.6 Plots illustrating the *Cumulative Gain (CG)* attained by subjects of the user study (on average), over the first 100 seconds of a search session. Shown are plots with the four different combinations of experimental condition trialled. Dashed lines represent fitted curves.

subjects would have experienced greater levels of gain. The empirical gain curves demonstrate that this actually occurred. A critical difference however though is for plot **(b)**. Here, it is clear that subjects went through a very different experience when searching, and this motivated a revision of our expectations.

To do so, we first fit a logarithmic function to each of the gain curves given session time, such that: $gain = b \cdot log(time) - a$, as used by Athukorala et al. (2014). Table 8.9 presents the parameters and correlation coefficients for fit ($\mathbf{r^2}$) for each of the four experimental conditions. We could then calculate how many documents subjects examined by drawing the tangent line to the estimated gain functions from the origin. This resulted in the predicted

Table 8.9 Fitting parameters for the gain curves illustrated in Figure 8.6, over each of the four experimental conditions trialled. Also included are the estimations from the model of the predicted number of documents that subjects would examine, and the actual number from the empirical data.

	Model Fitting Parameters			Predictions	
	a	b	r^2	Pred. D.	Actual D.
ND-AD	-1.08	0.48	0.989	3.68	3.23
ND-AS	-0.57	0.23	0.987	4.92	3.65
D-AD	-1.22	0.52	0.959	4.98	3.48
D-AS	-0.68	0.29	0.985	4.36	3.02

number of documents examined (**Pred. D.**), which we see are in line with the actual number of documents examined (**Actual D.**). With respect to plot **(b)**, we see that for diversified system D, the theory, given their performance, suggests that subjects should examine more documents per query under the aspectual task AS than when undertaking the ad-hoc task AD (i.e. 4.98 vs. 3.68 for AS and AD respectively, as shown in Table 8.9). We observed that subjects examined 3.48 and 3.02 documents per query (shown in Table 8.4 and repeated in Table 8.9) – which follows a similar trend but not to the same magnitude. Thus, this revises our expectations regarding how people would search differently between these tasks.

With respect to H1, we see that the theory, given their performance, suggests that subjects – when undertaking the aspectual retrieval task – would examine fewer documents per query when using the diversified system D than when using the non-diversified system, ND (4.36 vs. 4.92). Again, we see that subjects examined 3.02 and 3.65 documents per query respectively, again following the same trend – but not to the same magnitude. This post-hoc analysis provides justification for some of our initial hypotheses regarding how search behaviour would change under the different experimental conditions. However, it has also led to us revising our expectations based upon the empirical data.

8.2.3 Discussion

This user study has investigated the effects of diversifying search results when searchers undertook complex search tasks, requiring one to learn about different aspects of a topic. To test the series of hypotheses, derived from IFT outlined in Section 8.1.2.1, we conducted a within-subjects user study, using:

- a non-diversified system ND; versus
- a diversified system D.

These were tested over two different search tasks, where the task was set to either:

- ad-hoc topic retrieval AD; or
- aspectual retrieval AS.

This led to four experimental conditions. Our findings lend evidence to support the IFT hypotheses broadly. However, we only observed statistically significant differences across a subset of behavioural and temporal measures. This was despite the fact that there were significant differences in performance between systems ND and D. Diversified system D was able to, on average, return a ranked list of results with a greater number of documents containing new, unseen entities. This finding is in line with past work which found that interface-based interventions seemingly had little influence on search performance and search behaviours. Clearly, bigger differences need to be present – or larger sample sizes are required – to determine if the difference between systems over all examined indicators are significant. Despite these results, there were a number of clear trends.

When performing the aspectual task AS on the diversified system D (in contrast to the non-diversified system ND): subjects examined fewer documents per query (3 vs. 3.7 documents/query), issued slightly fewer queries (5.92 vs. 5.25 queries), and didn't go to as great a depth when examining SERPs (depths of 12.85 vs. 15.73). Taken together, this resulted in a lower probability of clicking ($P(C)$ = 0.16 vs 0.21, which was significantly different) and interestingly a lower probability of clicking on non-relevant document ($P(C|N)$ = 0.13 vs. 0.18, which was also significantly different). While subjects spent a similar amount of time searching on both systems, subjects on the diversified system spent slightly more time examining each document (15.97 seconds vs. 13.03 seconds) – suggesting that more effort was directed to assessing rather than searching. However, subjects found significantly more entities (7.22 vs. 4.31 entities) and found more documents that contained new/different entities (3.20 vs 2.35). Both of these findings were statistically significant. This shows that the diversification algorithm led to a greater awareness of the topics, and provided subjects with greater coverage of the topic. In turn, this also suggests that subjects were able to learn more about the topic, and were exposed to less bias.

When performing the ad-hoc task AD over the diversified system D (in contrast to the non-diversified system ND): subjects examined more documents per query (3.48 vs. 3.23 documents/query), issued slightly more queries (4.96 vs. 5.20 queries), and examined content to greater depths presented on SERPs (depths of 16.19 vs. 13.94). Again, this meant that the probability of clicking was lower on the diversified system D (0.16 vs. 0.20); this was significantly so. Subjects spent similar amounts of time searching on both systems. However, unlike on the aspectual tasks AS, subjects spent less time examining potentially relevant documents on system ND (13.66 vs. 15.09 seconds). This suggests that less effort was directed at assessing, rather than searching. This could be possibly due to the performance of D being higher than ND ($P@5 = 0.29$ vs. 0.25, which was significantly different). Alternatively, it could be because the results returned were easier to identify as relevant, as the probability of marking a document given it was relevant was higher (0.74 vs. 0.67). This suggests that subjects may be more confident when using the diversified system. Although not explicitly requested in the task description, subjects encountered more novel entities when using D (5.82 vs. 4.37). Subjects also found more documents with new entities using D (2.63 vs. 2.02). Taken together, this suggests that subjects again implicitly learn more about the topic because the diversified system D surfaced content that presented a more varied view on the topic.

With regards to the application of IFT, we showed that the generated hypotheses were largely sound. However, the empirical data prompted us to revise the hypotheses. Initially, we hypothesised that the performance and behaviour on both tasks would be similar when using the diversified system D (see Figure 8.2 **(b)**). However, post-hoc analysis revealed that the performance (and subsequent behaviour) was different (see Figure 8.6 **(b)**). Here, subjects obtained higher levels gain for the ad-hoc task AD. Thus, under such conditions, IFT would stipulate that they would examine more documents per query (3.48 vs 3.02 documents/query) and issue fewer queries (4.96 vs. 5.92 queries) when undertaking the ad-hoc retrieval task AD vs. the aspectual retrieval task AS (as opposed to there being no difference). Encouragingly, our application of IFT (before and after the experiment) led to new insights into how behaviours are affected under different conditions. This shows that IFT is a useful tool in developing, motivating and analysing search performance and behaviours. Furthermore, counter to our intuition about how we *believed* people would behave in these conditions, the theory provided *more informed and accurate hypotheses* which tended to hold in practice.

In past work, many interface-based solutions were studied, where a few significant differences in behaviour were found when compared to a standard interface. Disappointingly, we also found that an algorithmic solution has little impact or influence either, though there were trends which indicated that diversifying search results does indeed lead to better performance, greater awareness of the topic (even when not specifically instructed, i.e. *find relevant only*), and fewer examinations of non-relevant items. Thus, this allows the suggestion that diversification should be employed more widely (in particular, in the context of news search) where bias is an issue, and diversification algorithms can present a broader overview of the aspects within a topic. From this discussion, we now move to the next section, outlining our corresponding simulations of interaction.

8.3 Simulated Analysis

From the user study, we now move to a simulated analysis of searcher stopping behaviour and performance. In this section, we detail how stopping behaviour and performance vary when simulated searchers utilise different systems (i.e. D vs. ND) and search tasks (i.e. AD vs. AS). These are again, like in Section 7.3, addressed in the context of the two high-level research questions. Considering each of the twelve result summary level stopping strategies enumerated in Chapter 5, how does each strategy:

- HL-RQ3a perform; and
- HL-RQ3b approximate actual searcher stopping behaviour?

In the remainder of this section, we discuss the specific details of our methodology (Section 8.3.1), discussing in particular how we instantiated the task goals for these experiments. We then move onto an examination of the results from our simulations (Section 8.3.2).

8.3.1 Methodology

This methodology section provides the details specific to how we instantiated this set of simulations. One can assume that any components that are not discussed in this section were instantiated as shown in the general simulation methodology. The general methodology is presented in Section 6.4 on page 157.

Two of the key differences between the simulations reported in Chapter 7 and the simulations discussed here are how we operationalise the underlying retrieval system to support diversity, and the simulated searcher *task goals*.

8.3.1.1 Experimental System and Conditions

The **SimIIR** framework was adapted to incorporate additional components, allowing the simulations to employ the use of the diversified system D, as well as the standard BM25-based non-diversified system, ND. This involved the development of a further retrieval system component (refer to Section 6.4.1 on page 159 – in particular, the **Retrieval System** block in Figure 6.5 on page 160) that catered for system D. Therefore, this new component allowed us to run simulations over the four experimental conditions trialled as discussed in Section 8.2.1. Given the same queries, results returned from systems D and ND in the simulations were identical to results returned to the real-world subjects of the corresponding user study.

The second major component of the **SimIIR** framework that we considered was the result summary and document decision makers. These components determine whether a result summary is considered attractive enough to click and whether a document is relevant to the given TREC topic. For aspectual search tasks AS, the focus was not to simply save relevant documents but to save relevant documents containing at least one new entity associated with the topic. However, we decided to keep the decision makers the same for both search tasks, meaning that only the relevance to the topic was considered. We do however report results with aspectual measures, such as AR, in Section 8.3.2.

8.3.1.2 Interaction Costs and Probabilities

Interaction log data from the associated user study was taken and filtered by the four experimental conditions. Following the methodology outlined in Sections 6.4.2.3 and 6.4.2.1 on pages 167 and 163 respectively, we then extracted the different interaction probabilities and costs to ground our simulations.

Table 8.10 presents the interaction probabilities and costs across the four experimental in-

Table 8.10 Summary table of the different interaction costs (in seconds) and probabilities, with **P(C)** denoting the probability of a click, and **P(S)** denoting the probability of saving a document (considering it relevant). Also included are probabilities broken down over TREC relevant (**P(C|R)** and **P(S|R)**) and non-relevant (**P(C|N)** and **P(S|N)**). Values are reported across the four experimental conditions trialled. Refer to Sections 6.4.2.1 and 6.4.2.3 respectively for further information on how the costs and probabilities were derived. All data in this table are attained from interaction data extracted from the user study reported in Section 8.2.

		D-AS	ND-AS	D-AD	ND-AD
P(C)	**P(C|R)**	0.27	0.30	0.25	0.31
	P(C|N)	0.13	0.18	0.13	0.17
P(S)	**P(S|R)**	0.78	0.63	0.74	0.67
	P(S|N)	0.59	0.61	0.65	0.65
Costs (in seconds)	**Query**	8.80	9.99	9.69	8.69
	SERP	5.92	6.29	6.36	5.79
	Result Summary	1.59	1.75	1.71	1.71
	Document	15.97	13.03	13.66	15.09
	Save	1.73	1.78	1.58	1.68
	Task Goal	Find 6 relevant (refer to Section 8.3.1.2)			
	Session Timeout	500 seconds (refer to Section 8.3.1.2)			

terfaces trialled. Included within the table are the interaction probabilities for clicking on result summaries (under the **P(C)** grouping) and saving documents (under the **P(S)** grouping). Also included are the five main interaction costs, presented under the **Costs** grouping.

Task Goal Also included in Table 8.10 is the task goal. In the user study, subjects were instructed to find and save a minimum of four useful documents, as reported back in Section 8.2.1.1. Across all four experimental interfaces, subjects saved on average 5.87 documents, perhaps to hedge their bets in the eventuality that certain saved documents turned out to not be useful after all. These values can be found in the **#Saved** row in Table 8.4 on page 268. As such, simulated searchers were given a goal of finding *six* useful documents to mirror the average saved by their real-world counterparts.

Session Timeout A session timeout was also provided such that if simulated searchers failed to find the minimum of six useful documents, a time limit would prevent the simulation from 'getting stuck', where the query lists would be exhausted. Again referring to Table 8.4 on page 268, the mean total session time (reported on row **Total Session**) was reported to be 438.47 seconds across all four experimental conditions. With a large variance in total session time reported across search sessions, we set a time limit of 500 seconds (equating to the upper side of the variance) to grant sufficient time for searchers to find the six required documents, yet restricting runaway (and unrealistic) behaviours.

Simulations of interaction when considering aspectual retrieval tasks have to the best of our knowledge not been performed before. The majority of prior work purely considers ad-hoc retrieval, given that this is a relatively straightforward search task to model. Therefore, we suggest that the assumptions made for these simulations of interaction provide a reasonable approximation for how real-world searchers actually performed and behaved, and leaves scope for development of these simulations in future work. We leave the discussion on these assumptions and potential issues that may arise from them to Section 10.3.1 on page 349.

8.3.2 Results

We now report the results of our simulations of interaction, under different search tasks and goals. As presented in Chapter 7, we discuss our findings from these experiments over two subsections, considering:

- the *performance* runs (Section 8.3.2.1), where we discuss the highest levels of performance attained by simulated searchers under different *what-if* scenarios; and
- the *comparison* runs (Section 8.3.2.2), providing results of the simulations that were directly compared to the actual mean.

Both of these sections provide an answer for high-level research questions HL-RQ3a and HL-RQ3b respectively, this time under the context of varying search tasks and goals.

Significance Testing Like before, we employ significance testing to determine whether the performance of a result summary level stopping strategy was significantly different from the other eleven trialled. All tests in this section utilise the two-tailed Student's t-test, where $\alpha = 0.05$. Our tests consider the best performing or approximating stopping strategy, and how they compare to the other eleven. Like before, we are interested in *statistical non-significance* (i.e. $\alpha > 0.05$), meaning that the compared stopping strategies are *similar* to one another in terms of performance or approximations.

8.3.2.1 Performance

Before reporting on the performance of each stopping strategy across our performance *(what-if)* simulations, we must first determine whether querying strategy QS13 delivers queries of expected performance across systems ND (non-diversified) and D (diversified). From the reporting of the user study, we know that the performance across the two systems (as shown in Table 8.3) was significantly different in terms of precision (at most ranks), *αDCG* and *AR*. Results from the user study showed consistently higher levels of *αDCG* and *AR* for system D, a result that was in line with expectations.

Table 8.11 reports various precision, *αDCG* and *AR* for queries generated by querying strategy QS13. Across both systems ND and D, measures (± standard deviations) are shown across all three stopping strategies, including QS13 and constituent querying

Table 8.11 Mean performance values (± standard deviations) of all generated queries issued for performance runs. Included are **P@k**, **αDCG@k** and **AR@k** values, incorporating aspectual retrieval measures. Values are reported across both the non-diversified (ND) and diversified (D) systems. Note the increasing trends in performance across all measures as QS1 → QS3, as well as the improved performance for diversification measures over system D.

	ND (Non-Diversified)			**D** (Diversified)		
	QS1	QS13	QS3	QS1	QS13	QS3
P@1	0.04 ± 0.20	0.18 ± 0.39	0.23 ± 0.42	0.04 ± 0.20	0.19 ± 0.39	0.23 ± 0.43
P@5	0.02 ± 0.07	0.14 ± 0.21	0.18 ± 0.23	0.05 ± 0.16	0.18 ± 0.26	0.22 ± 0.28
P@10	0.02 ± 0.06	0.11 ± 0.18	0.15 ± 0.19	0.03 ± 0.10	0.14 ± 0.20	0.17 ± 0.21
αDCG@5	0.00 ± 0.00	0.01 ± 0.03	0.02 ± 0.03	0.00 ± 0.00	0.02 ± 0.04	0.02 ± 0.04
αDCG@10	0.00 ± 0.00	0.01 ± 0.03	0.02 ± 0.03	0.00 ± 0.01	0.02 ± 0.04	0.03 ± 0.05
AR@5	0.01 ± 0.04	0.30 ± 0.87	0.40 ± 0.98	0.05 ± 0.24	0.51 ± 1.25	0.65 ± 1.40
AR@10	0.01 ± 0.04	0.20 ± 0.51	0.27 ± 0.58	0.03 ± 0.13	0.30 ± 0.69	0.38 ± 0.77

strategies QS1 (single term, poor queries) and QS3 (three term, good queries). A total of 213 unique queries were extracted from the simulated interaction logs. Queries were then categorised depending upon their term length, allowing us to deduce measures for each individual querying strategy.

Closer examination of Table 8.11 shows that as we move from QS1 to QS3, a consistent improvement in precision is achieved. Again, this finding is in line with intuition. Interleaved querying strategy QS13 delivers intermediary performance between the two. When we turn our attention to aspectual measures, we again see performance improvements as we move from QS1 to QS3. Considering systems ND and D, we also see improvements in performance. For example, *AR*@10 is reported as 0.27 ± 0.58 for system

ND, with 0.38 ± 0.77 for system D. This jump in mean AR once again demonstrates that the diversification algorithm presented in Figure/Algorithm 8.5 did indeed work as expected, where it returned a greater number of unique entities in its re-ranked search results.

Results from Table 8.11 therefore provide us with confidence that the two systems were indeed working as intended, yielding queries that performed in line with our intuition, and offered similar trends in performance compared to those issued by the real-world subjects of the user study. We also gain confidence in knowing that interleaved querying strategy QS13 also offered improvements in aspectual measures compared to QS1. As such, this should be later reflected in our examination of the *what-if* experiments when we consider aspectual measures.

Satisfied with the performance of the generated queries, we now turn our attention to an examination of the individual result summary level stopping strategies. Like our reporting in Section 7.3.2.1, we consider results primarily from the perspective of HL-RQ3a, which requires an examination of the performance of each stopping strategy. Before this, we consider the general trends that we observed across the performance simulations, examining whether these trends are consistent across the experimental conditions trialled.

Figure 8.7 presents twelve individual plots, one per result summary level stopping strategy. These plots represent the mean levels of performance attained over each experimental condition (of either D AS, ND AS, D AD or ND AD) at varying depths per query, averaged over the five individual topics and 50 individual trials. The mean depth per query is represented along each x axis, with the performance attained (represented as CG) represented along the y axes. Although strategies such as SS3-NC allowed simulated searchers to browse to depths greater than 25 on average, all plots were cut at this depth for consistency, and to highlight what occurs at lower depths per query. Each point on the respective lines for each condition represents one of the stopping threshold values used for each stopping strategy, as reported in Table 6.3 on page 178.

General trends across mean depths per query can be observed in each of the twelve plots in Figure 8.7. We can see across a majority of stopping strategies and experimental conditions that as the mean depth per query increases, simulated searchers attain greater of levels of CG on average before a peak point of CG attainment is reached. After this point, as searchers traverse result lists to greater depths, the mean level of CG begins to diminish. Peaks can be more profound in some stopping strategies (e.g. SS1-FIX and SS5-COMB) than with others (e.g. SS4-SAT). This trend can be observed across all four experimental conditions, where all four start at very similar levels of mean CG across shallow mean depths per query, before gaps begin to emerge between them. Indeed, it can be observed across the twelve plots that condition D AS consistently offers the best approximations across nearly all depths per query reported. This condition is very closely followed by D AD, with slightly lower mean levels of CG. Interestingly, we then observe a gap between these two conditions, and the remaining two conditions, ND AS and ND AD. This gap is largely present amongst all twelve stopping strategies and is more profound in stopping strategies that offer higher mean levels of CG. This is especially true after peak CG has been reached, and the mean depth per query begins to increase. The gaps between conditions clearly demonstrate a difference in performance between systems D and ND, with system D consistently offering improved mean levels of CG. Interestingly, gaps between tasks AD and AS are less profound, with negligible differences observed from an examination of the twelve plots.

While the plots in Figure 8.7 present the general trends in performance across mean depths per query, Table 8.12 reports on the *highest level of CG* attained by each result summary level stopping strategy, across the four experimental conditions trialled. The values in Table 8.12 correspond to the peaks shown in each of the plots in Figure 8.7. For each stopping strategy and condition, we report: the greatest level of mean CG attained (**CG**); the mean depth per query at which this value was reached (**DQ**); and the stopping threshold value(s) that were used to attain this value ($\mathbf{x}_n$). Highlighted are the stopping strategies that yielded the highest mean level of CG, with the highest value demonstrated for each condition. For

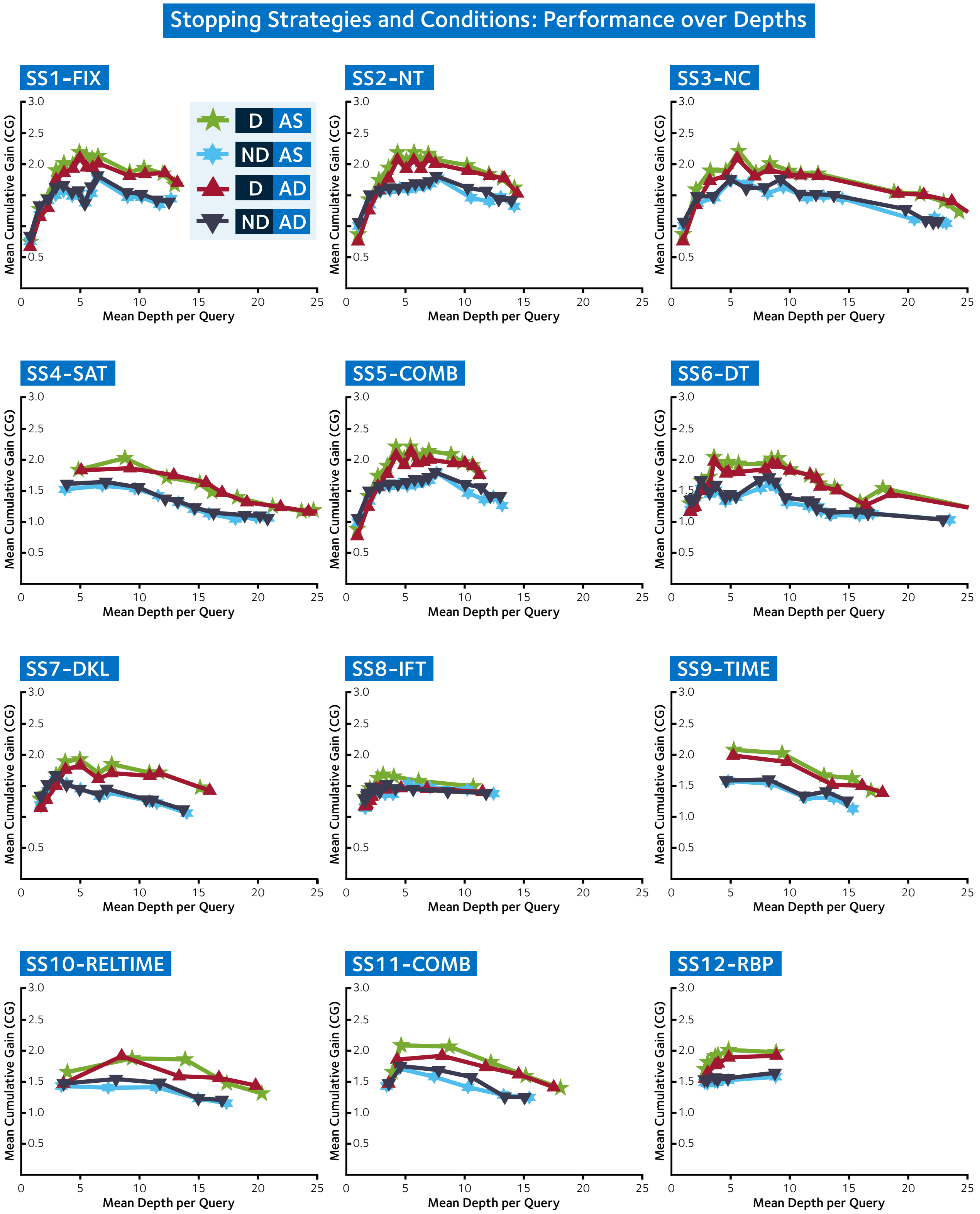


Figure 8.7 Plots showing the varying levels of performance, measured in CG, over the mean depth per query. Each result summary stopping strategy is shown on an individual plot, with each of the four experimental conditions shown within each plot. The depth per query reported on each *x* axis is cut at 25 to allow for an easier comparison between different stopping strategies.

conditions D AS, ND AS and D AD, combination stopping strategy SS5-COMB attains the highest mean level of CG, with values 2.21, 1.79 and 2.13 reached for the aforementioned conditions respectively. As these levels of CG are reached under a combination strategy, values $x_2 = 5$ and $x_4 = 3$, $x_2 = 10$ and $x_4 = 6$, and $x_2 = 7$ and $x_4 = 3$ were used to attain these values for conditions D AS, ND AS and D AD respectively. This means that under condition D AS, a searcher would examine a total of five non-relevant documents or save three relevant documents per query before stopping – whatever occurred first. Interestingly, our fixed-depth, baseline stopping strategy SS1-FIX reaches the highest level of CG for condition ND AD, at 1.81. SS5-COMB is however very close behind, with a mean CG of 1.80 reported. Like the results in Section 7.3.2.1, this again demonstrates that a fixed-depth strategy can be hard to beat in terms of attaining a high level of CG.

Indeed, it should be noted that the following stopping strategies reported maximum levels of CG close to the absolute maximum observed:

- SS2-NT and SS3-NC, the frustration stopping strategies (considering total and contiguous non-relevance);
- SS4-SAT, the satiation-based stopping strategy;
- SS9-TIME, the time-based strategy; and
- SS11-COMB, the patch-based combination strategy.

This was true across all four experimental conditions. For SS5-COMB, the relatively low mean depths per query at which the greatest level of CG was reached also demonstrates that the combination strategy was particularly robust at detecting a query of poor performance. Subsequently stopping at shallower depths, the strategy thus saved time for the simulated searcher. Taking this further, an interesting trend that was initially observed in Figure 8.7

shows that peak CG is attained at generally shallower mean depths per query under tasks using system D than when compared to system ND.

With performance levels attained by several stopping strategies reported to be very similar, we employed statistical testing to determine what strategies, if any, yielded *significantly different* performances from the best performing strategy over each condition. At $\alpha = 0.05$, this statistical significance would demonstrate that performance is significantly poorer than the best performing strategy (SS5-COMB for D AS, ND AS and D AD, with SS1-FIX for ND AD). Results of these tests can also be observed in Table 8.12, with cells highlighted denoting no statistical significance from the best performing strategy (i.e. $p > 0.05$). Therefore, cells without highlighting denote a statistically significant difference in terms of the CG values reported. Indeed, only SS8-IFT and SS10-RELTIME yielded significant differences across interfaces D AS, ND AS and D AD, suggesting that these particular stopping strategies were not effective. All other strategies reported no significant differences from the best performing strategies.

Given that subjects from the user study were asked in aspectual (AS) tasks to find relevant documents containing at least one new, unseen entity, we consider in tandem both mean CG and mean AR. Figure 8.8 again presents 12 different plots, each one representing the individual result summary level stopping strategy. Each line on the plot again represents one of the four experimental conditions that were trialled. While the mean depth per query is shown along the x axes, we instead plot these against the mean AR values along the y axes, denoting the mean number of documents containing unseen entities over the course of a session. Unsurprisingly, the plots follow similar trends to those shown in Figure 8.7, with the same mean depth per query values shown.

Indeed, a higher mean level of CG correlates strongly with a higher mean level of AR – an intuitive result. In order to attain gain, one must save documents, and by saving documents, a simulated searcher will also identify documents with unseen entities. Trends illustrate that like Figure 8.7, plots build up to a peak before slowly diminishing as the

Table 8.12 Results from the simulated *what-if* simulated performance runs, showing the highest levels of **CG** attained for each result summary level stopping strategy trialled (grouped by their type). $\mathbf{x_n}$ denotes the parameter threshold(s), with **DQ** denoting the depth per query at which the greatest CG value was attained at. For each condition, the stopping strategy which attained the highest level of CG is highlighted. Light blue highlighting denotes *no significant difference* from the best performing strategy, with **no highlighting** denoting a significant difference at α=*0.05*. For combination thresholds, x_2,x_4 are presented for SS5-COMB, with x_{10},x_4 for SS11-COMB.

		D-AS			ND-AS			D-AD			ND-AD		
		x_n	DQ	CG	x_n	DQ	CG	x_n	DQ	CG	x_n	DQ	CG
FIX	SS1	7	4.95	2.19	10	6.45	1.77	7	4.99	2.09	10	6.47	1.81
FRUS	SS2	5	4.29	2.18	10	7.71	1.76	9	6.93	2.08	10	7.55	1.80
	SS3	5	5.64	2.20	4	5.03	1.73	5	5.52	2.08	4	4.93	1.74
SAT	SS4	2	8.77	2.01	2	6.87	1.57	2	9.21	1.85	2	7.15	1.63
COM	SS5	5,3	4.21	2.21	10,6	7.65	1.79	7,3	5.47	2.13	10,6	7.50	1.80
DIFF	SS6	0.30	3.58	2.02	0.55	8.29	1.58	0.30	3.60	1.95	0.55	8.22	1.70
	SS7	5.5	4.91	1.92	6.5	2.87	1.59	5.5	4.95	1.82	6.5	2.86	1.66
IFT	SS8	0.008	3.18	1.68	0.008	5.30	1.53	0.006	4.68	1.46	0.010	3.48	1.52
TIME	SS9	30	5.26	2.08	30	4.79	1.58	30	5.22	1.99	60	8.21	1.61
	SS10	20	9.37	1.86	10	3.39	1.42	20	8.48	1.90	20	8.01	1.53
COM	SS11	10,4	4.62	2.08	10,6	4.44	1.71	20,4	8.08	1.91	10,6	4.56	1.75
RBP	SS12	0.95	4.82	2.01	0.99	8.81	1.57	0.99	8.87	1.91	0.99	8.75	1.63

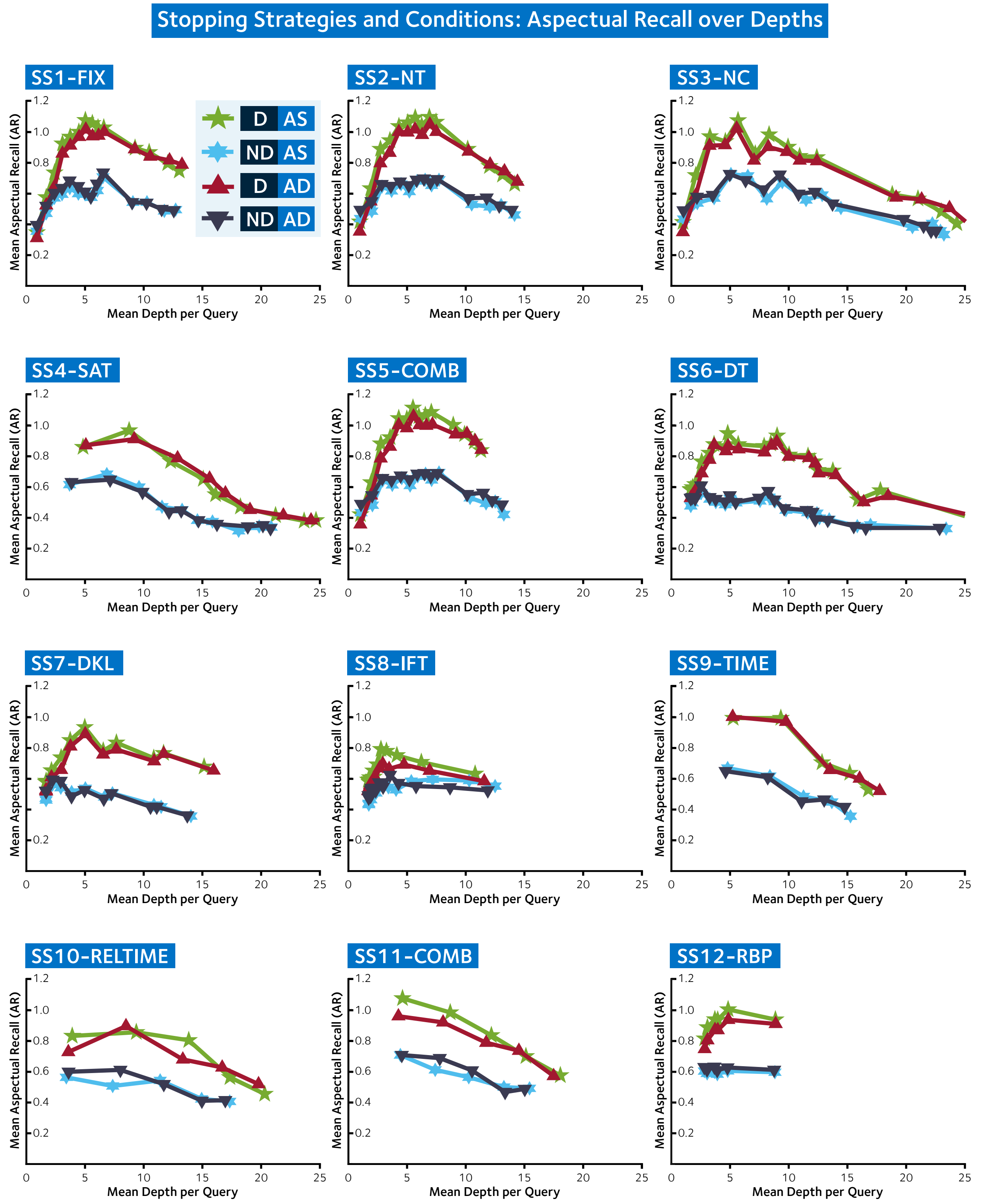


Figure 8.8 Aspectual recall (documents containing at least one new entity) over the mean depth per query for each result summary level stopping strategy, reported over each of the four experimental conditions. Note the profound gaps between the two tasks using diversified system D and non-diversified system ND. The depth per query reported on each *x* axis is cut at 25.

mean depth per query increases. We also observe that a much higher level of AR is attained by tasks using system D. For example, SS5-COMB attains a maximum AR of 1.04 at a mean depth per query of 4.21, and 0.69 at a mean depth per query of 7.65 under conditions D AS and ND AS respectively. These results provide further proof that subjects and simulated searchers enjoyed a more diverse set of results when the diversification algorithm was applied.

Turning our attention to the remaining stopping strategies, we first consider our fixed-depth baseline SS1-FIX, in addition to frustration-based strategies SS2-NT and SS3-NC. All three perform remarkably well, with SS1-FIX indeed offering the best performance for condition ND AD. Under condition D AS, for example, the three strategies offer a maximum CG of 2.19, 2.18 and 2.20, respectively. This is compared to the maximum reached by SS5-COMB of 2.21. In addition, the strategies all offer these levels of CG at similar mean depths per query, suggesting that such approaches are just as robust as the combination strategy. This perhaps is an unsurprising finding, as SS5-COMB utilises SS2-NT as its frustration-based component.

Considering our time-based stopping strategies SS9-TIME and SS10-RELTIME, we find that SS9-TIME consistently delivered the best performance across the four experimental conditions. This result is perhaps somewhat surprising and mirrors findings from Section 7.3.2.1 on page 219. An adaptive approach such as SS10-RELTIME would intuitively make more sense. Here, the searcher would then be able to adapt his or her stopping behaviour based upon the amount of relevant content found, rather than simply stopping after 30 or 60 seconds have elapsed from the point at which they would have begun their examination of results.

Considering our measure-based strategy SS12-RBP, the performance offered across the four conditions is somewhat poorer than the other strategies. However, no significant differences in performance were observed with this stopping strategy. Once notable trend regarding the results for this strategy is the relatively shallow mean depth per query that simulated searchers reached, despite the close proximity of the patience parameter to 1. For tasks employing system ND, results are particularly invariant – no obvious peak in performance is observed.

The difference threshold stopping strategies SS6-DT and SS7-DKL yielded similar performance when compared to SS12-RBP. However, simulated searchers subscribing to the difference strategies both traversed result lists to greater depths on average. Performance from these strategies perhaps was hindered by the querying strategy that we employed and demonstrates that they are not particularly robust at dealing with poor quality queries.

The IFT-based stopping strategy SS8-IFT however consistently performed the worst over all four experimental conditions. As can be observed from Figure 8.7, we can see that the aforementioned gap between the two systems was not present, with all four tasks bunched very closely together across the range of mean depths per query. Indeed, performance across tasks utilising system D was significantly different from the best performing strategy, SS5-COMB. CG reached 1.68 and 1.46 for tasks D AS and D AD respectively at relatively low mean depths per query. This suggests that the strategy was stopping too early, which may indicate an incorrectly set rate of gain. This, and other aspects of our findings are discussed in Section 10.2, beginning on page 338.

8.3.2.2 Real-World Comparisons

We now turn out attention to an examination of our real-world searcher comparison runs. These simulations provide a means for us to address HL-RQ3b, under the context of varying task types and goals. As a reminder, we *replayed* all of the queries issued by the real-world searchers, allowing us to make a direct comparison between simulated and real-world click depths, indicative of a searcher's stopping behaviours.

Figure 8.9 on page 297 provides a total of twelve plots, one per result summary level stopping strategy. Within the plots are four lines, one representing the approximations offered by each of the four conditions for a given stopping strategy. Along the x axes is the mean depth per query, with the y axes this time to denote the MSE of the real-world vs. simulated click depths for the given stopping strategy and experimental condition combination.[6] Each point on the plotted lines represents a stopping threshold parameter configuration. The closer the MSE tends to zero, the better the approximation to actual searcher stopping behaviours. For reference, we also include four dashed vertical lines on each of the plots in Figure 8.9. These lines represent the mean click depths achieved by the real-world searchers, over each experimental condition. The closer the lowest MSE was in a given simulation to the corresponding real-world click depth, the better the approximation. For example, the last point for condition D AS under stopping strategy SS1-FIX lied very close to the real-world click depth mean of 12.85, compared to the simulation result of 13.25. However, under SS7-DKL, the same condition obtained a relatively shallower depth compared to the mean of the corresponding real-world value. This results in a higher MSE.

From the plots in Figure 8.9, we can observe a number of notable trends. One of the first points of note is the consistent ordering of the different conditions across the twelve stopping strategies. Remembering that the *lower* the MSE value the better the approximation offered, we see that D AS consistently offers the best approximations, especially at lower mean depths per query. This is then followed by ND AD, D AD and finally ND AS. Note that unlike the results of the performance *(what-if)* simulations reported earlier in Section 8.3.2.1, there is a distinct lack of separation between the two experimental systems trialled – all four conditions are evenly spaced out. We also see smooth curves for stopping strategies such as SS1-FIX, SS2-NT and SS5-COMB that offered good levels of performance from simulations reported in Section 8.3.2.1. Conversely, strategies that performed poorly also appear more diffuse with widely spaced plots, such as those shown by SS7-DKL and SS8-IFT. These plots demonstrate that the approximations offered by the stopping strategies are relatively poor related to the mean depths per query, reflecting the presence of less exact approximations.

[6] Refer to Section 6.4.3.2 on page 182 for further information on how we computed the *Mean Squared Error (MSE)*.

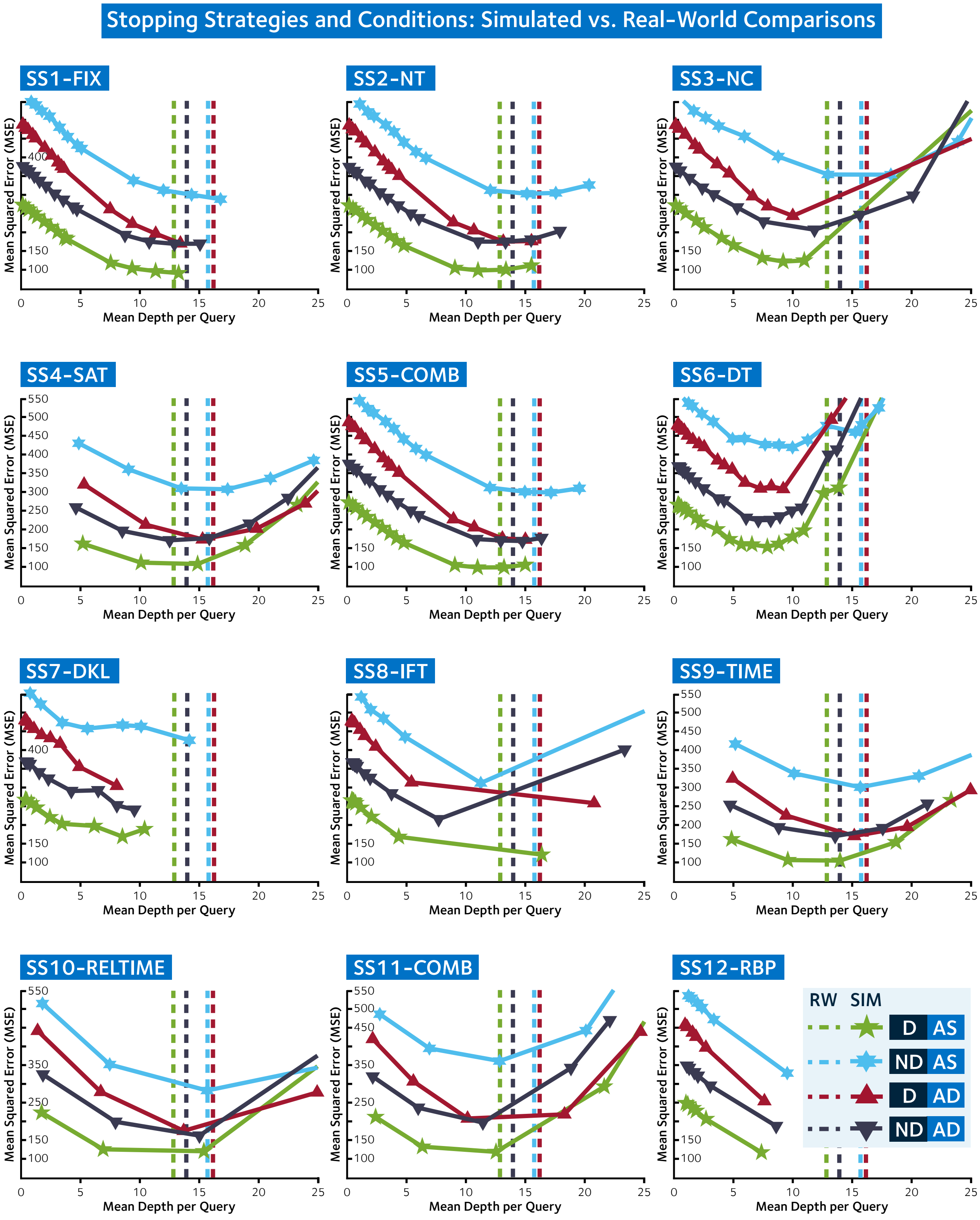


Figure 8.9 Plots reporting the comparison runs, reporting the MSE vs. the mean depth per query. Runs over each of the four experimental conditions are shown. Also included in the plots are a series of dashed lines denoting the mean depth per query reached by the real-world subjects of the user study — one for each experimental interface.

Given these trends, what stopping strategies offer the best approximations of actual searcher stopping behaviours? Table 8.13 presents for each stopping strategy (over each condition) the point on the corresponding plots in Figure 8.9 where the lowest MSE value is attained (**MSE**), together with the values of stopping parameter threshold(s) ($\mathbf{x_n}$) used for its calculation. Note that for condition D AS, generally lower MSE values are attained than under other conditions. The table also highlights the stopping strategy that yielded the lowest MSE for each condition. Interestingly, the baseline, fixed-depth stopping strategy SS1-FIX offered the best approximations for conditions D AS and D AD, both of which were using diversified system D. In comparison, conditions ND AS and ND AD were both found to show SS10-RELTIME as the strategy that yielded the best approximation to actual searcher stopping behaviour. Consistency is also observed across the stopping parameter thresholds that yielded the lowest MSE – for SS1-FIX, a depth of 24 is cited as yielding the best approximations, with the best approximations under SS10-RELTIME yielded at 30 seconds after the last relevant document had been saved.

Given the fact that we observed consistent results across the systems trialled, we also decided to examine whether a particular stopping strategy emerged as providing good approximations over all four experimental conditions when averaged together. Results from this analysis are shown in the **Average** grouping in Table 8.13. Statistical tests comparing the best approximating stopping strategy (again SS1-FIX @24) against the remaining eleven stopping strategies yielded no statistically significant differences. Indeed, this was true across all four experimental conditions when considered in isolation. This result highlights that all stopping strategies offered similar approximations. As in Chapter 7, no statistical differences emerged when considering the average of the four experimental conditions. Because of this, we do not explore this avenue any further.

Moving back to our per-condition examinations, Table 8.14 reports additional information relating to the best approximation offered by each stopping strategy. We report for each stopping strategy the: mean depth per query (**DQ**); the mean number of saved and TREC relevant documents (or interactive precision, **iP**); and the mean **AR**, denoting the number

Table 8.13 Results from the simulated comparison runs, showing the *lowest* MSE value reached over each result summary level stopping strategy trialled. $\mathbf{x_n}$ denotes the parameter threshold(s) that the lowest **MSE** was reached with. Results are presented across the four experimental conditions. For each condition, the stopping strategy that attained the lowest MSE is highlighted. For the combination stopping strategies, two parameters are presented, with x_2,x_4 presented for SS5-COMB and x_{10},x_4 presented for SS11-COMB. Significance testing yielded no significant differences between the stopping strategies at α=0.05.

		D-AS		ND-AS		D-AD		ND-AD		Average	
		x_n	MSE	x_n	MSE	x_n	MSE	x_n	MSE	x_n	MSE
FIX	SS1	24	90.27	24	287.50	24	169.62	21	168.39	24	179.03
FRUS	SS2	18	98.49	18	303.84	21	175.76	18	175.39	21	190.83
	SS3	9	123.04	9	354.20	10	244.59	8	207.71	8	247.90
SAT	SS4	3	107.38	4	307.00	3	173.16	3	170.68	3	190.14
COM	SS5	24,3	93.79	21,5	288.30	24,6	173.23	21,5	162.81	21,5	181.12
DIFF	SS6	0.70	154.08	0.70	418.69	0.85	307.94	0.65	224.88	0.70	280.81
	SS7	3.5	168.04	3.0	424.14	3.0	303.07	3.0	239.57	3.0	288.72
IFT	SS8	0.002	120.96	0.004	312.39	0.002	258.71	0.004	215.60	0.004	252.71
TIME	SS9	90	105.68	90	300.81	90	171.19	90	170.88	90	187.14
	SS10	30	120.43	30	282.86	30	174.96	30	162.19	30	185.11
COM	SS11	30,5	117.50	30,7	360.06	30,5	206.59	30,5	187.68	30,5	218.25
RBP	SS12	0.99	118.09	0.99	330.16	0.99	255.20	0.99	188.53	0.99	222.99

of saved, TREC relevant documents containing at least one new entity. The values were attained at the stopping threshold parameter(s) that yielded the lowest MSE, as reported in Table 8.13. Also included in Table 8.14 are the mean values for each measure that were attained by real-world subjects within the user study (represented by the **RW**) column, included for direct comparison. We once again also highlight the stopping strategy for each condition that yielded the lowest MSE values.

Results observed from Table 8.14 are largely as expected: stopping strategies that yielded close approximations to actual mean searcher behaviours offered mean depths per query that were similar to those of the real-world mean for the given condition. In contrast, we find that stopping strategies offering approximations with higher MSE values such as SS6-DT, SS7-DKL and SS8-IFT do so at lower or higher mean depths per query, although these differences are not statistically significant. Results from Table 8.14 also show that the simulated searchers on average saved fewer TREC relevant documents across all four experimental conditions, and, as a consequence, identified fewer TREC relevant documents with at least one new entity. For example, across conditions D AS, ND AS, D AD and ND AD, real-world subjects saved 2.63, 2.18, 2.51 and 2.22 TREC relevant documents respectively, on average. Across our fixed-depth baseline SS1-FIX, simulated searchers saved 1.94, 1.96, 1.67 and 1.66 TREC relevant documents. These comparatively low values may be an artefact of how we instantiated the simulations of interaction. We leave this discussion to Section 10.2.

With the comparatively low numbers of TREC relevant documents saved, this also corresponds to a relatively low level of CG when compared to the real-world means across conditions. Figure 8.10 reports the mean levels of CG that were attained for each of the stopping strategies, along with the mean real-world CG values for each experimental interface. These values were again computed from the stopping threshold parameter(s) that yielded the lowest MSE values, as reported in Table 8.13. Bar charts provide a visual representation of CG attained by each stopping strategy for each experimental condition, with the ranked values provided to the right of each chart. With real-world CG values of 4.09,

Table 8.14 Additional results from the searcher comparisons runs. We report the mean depth per query (**DQ**), the mean interactive precision (**iP**) — and from that, the mean number of saved documents that contain one or more new entities (**AR**). All values are reported at the configuration yielding the lowest MSE (refer to Table 8.13), indicating the best approximation to real-world stopping behaviours. We also include the mean real-world (**RW**) values over each condition for a direct comparison. Note that stopping strategies offering the lowest MSE are highlighted — cell colouring here does not denote the outcome of any significance testing.

		D-AS			ND-AS			D-AD			ND-AD		
		DQ	iP	AR	DQ	iP	AR	DQ	iP	AR	DQ	iP	AR
	RW	12.85	2.63	3.20	15.73	2.18	2.35	16.19	2.51	2.63	13.94	2.22	2.02
FIX	SS1	13.25	1.94	1.50	16.79	1.96	1.44	13.40	1.67	1.42	12.89	1.66	1.26
FRUS	SS2	11.01	2.04	1.58	15.15	1.91	1.43	13.15	1.75	1.48	13.34	1.77	1.33
	SS3	9.16	1.91	1.47	18.22	2.33	1.59	9.97	1.75	1.43	11.80	1.73	1.25
SAT	SS4	14.87	1.89	1.54	17.40	1.73	1.32	15.29	1.70	1.48	12.51	1.27	0.96
COM	SS5	11.62	1.81	1.48	15.43	1.86	1.43	14.98	1.80	1.49	14.16	1.75	1.32
DIFF	SS6	7.85	1.36	1.11	9.97	1.25	0.94	9.19	1.50	1.25	7.10	1.05	0.81
	SS7	8.50	1.35	1.12	14.12	1.40	1.05	8.02	1.29	1.14	9.50	1.30	0.99
IFT	SS8	16.35	1.16	0.83	11.21	1.04	0.74	20.76	1.16	0.86	7.66	0.95	0.71
TIME	SS9	13.96	1.92	1.49	15.66	1.76	1.33	15.15	1.71	1.46	13.55	1.49	1.12
	SS10	15.42	2.07	1.63	15.69	1.68	1.25	13.72	1.83	1.55	15.05	1.50	1.13
COM	SS11	12.54	1.75	1.44	12.81	1.48	1.14	10.14	1.58	1.37	11.71	1.35	1.03
RBP	SS12	7.36	1.54	1.30	9.54	1.29	1.00	7.61	1.39	1.25	8.61	1.19	0.92

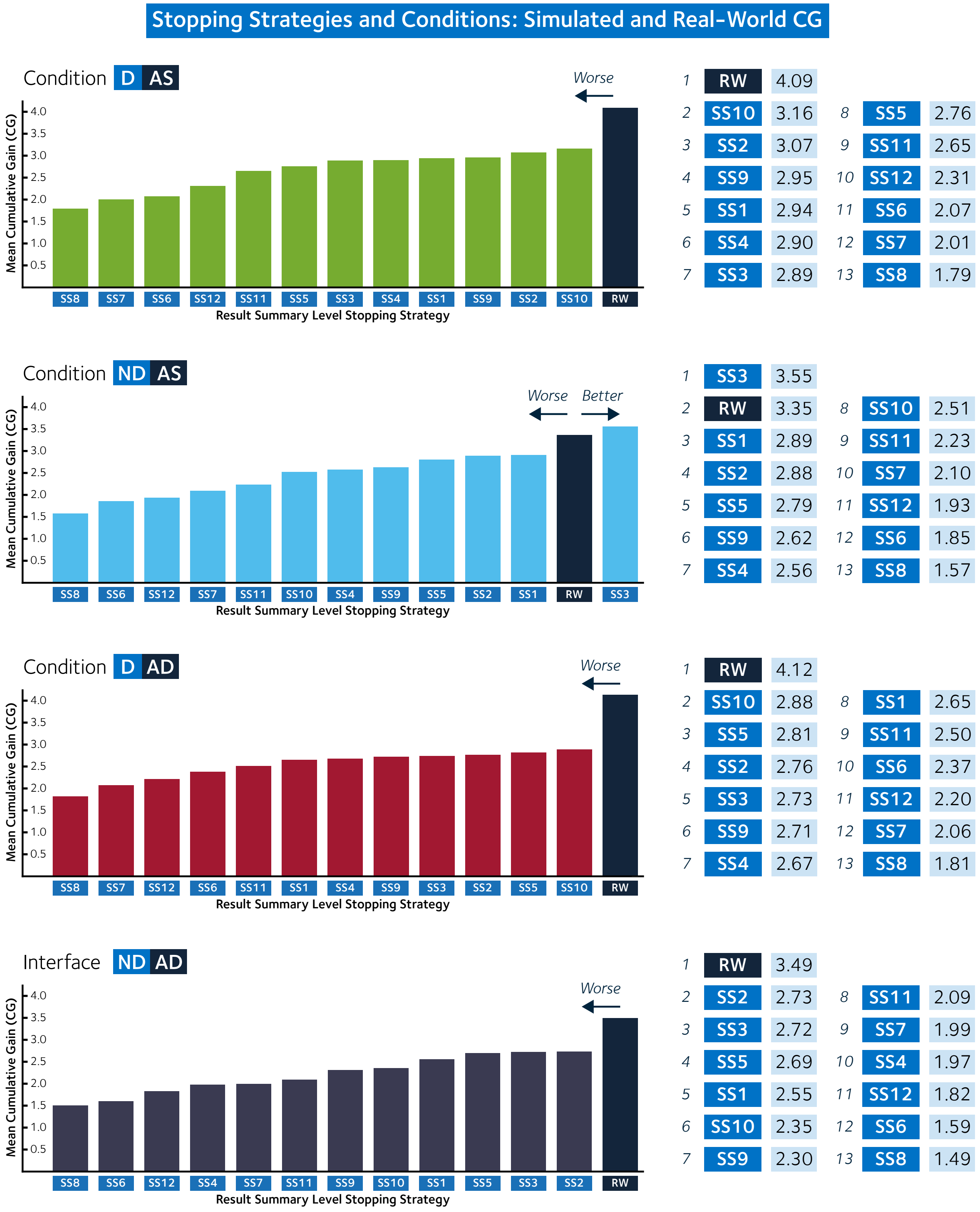
3.35, 4.12 and 3.49 attained for conditions D AS, ND AS, D AD and ND AD, we can see that in all but condition ND AS, the real-world searchers on average outperform all twelve stopping strategies. For condition ND AS, SS3-NC yielded a CG value of 3.55 compared to a real-world value of 3.35.

For conditions D AS and D AD, we find that stopping strategy SS10-RELTIME also yielded the highest level of CG after the real-world means. This is in contrast to stopping strategy SS1-FIX yielding the lowest MSE (approximating mean click depths). For condition ND AD, frustration-based stopping strategy SS2-NT was the stopping strategy offering the best CG (2.73), along with SS3-NC that offered the highest overall CG for condition D AS, as previously mentioned. This means that instead of following their actual behaviours, the real-world searchers would have on average attained a slightly higher level of CG if they rigidly followed SS3-NC @9. However, the difference is not significant.

Looking towards the lower end of the bar charts in Figure 8.10, we note that stopping strategies SS6-DT, SS7-DKL, SS8-IFT and SS12-RBP frequently appear. In particular, SS12-RBP consistently ranks last, with the lowest level of CG attained. For this stopping strategy, we can see that the plot in Figure 8.9 demonstrates that it underestimates the stopping depths, with searchers stopping at much lower mean depths per query. These values, being much lower than the mean real-world equivalents, result in higher MSE values. Underestimation can also be clearly seen for SS6-DT, with the lowest MSE values attained in the range of 7.10 to 9.97 across the four experimental interfaces. The same can be observed for SS7-DKL. Contrast this plot for example to that of SS4-SAT, with a smooth curve and the lowest point of each line close to the real-world means.

8.4 Chapter Summary

In this chapter, we have examined how varying search task type, goals and systems affect the behaviour, performance and user experience of searchers. This was conducted via a



Condition D AS

1	RW	4.09			
2	SS10	3.16	8	SS5	2.76
3	SS2	3.07	9	SS11	2.65
4	SS9	2.95	10	SS12	2.31
5	SS1	2.94	11	SS6	2.07
6	SS4	2.90	12	SS7	2.01
7	SS3	2.89	13	SS8	1.79

Condition ND AS

1	SS3	3.55			
2	RW	3.35	8	SS10	2.51
3	SS1	2.89	9	SS11	2.23
4	SS2	2.88	10	SS7	2.10
5	SS5	2.79	11	SS12	1.93
6	SS9	2.62	12	SS6	1.85
7	SS4	2.56	13	SS8	1.57

Condition D AD

1	RW	4.12			
2	SS10	2.88	8	SS1	2.65
3	SS5	2.81	9	SS11	2.50
4	SS2	2.76	10	SS6	2.37
5	SS3	2.73	11	SS12	2.20
6	SS9	2.71	12	SS7	2.06
7	SS4	2.67	13	SS8	1.81

Interface ND AD

1	RW	3.49			
2	SS2	2.73	8	SS11	2.09
3	SS3	2.72	9	SS7	1.99
4	SS5	2.69	10	SS4	1.97
5	SS1	2.55	11	SS12	1.82
6	SS10	2.35	12	SS6	1.59
7	SS9	2.30	13	SS8	1.49

Figure 8.10 Bar charts, one per experimental condition, demonstrating the mean level of CG attained by each result summary level stopping strategy. Ordered by CG, these values are reached using the threshold configurations yielding the best approximations to actual searcher behaviour, as shown in Table 8.13. Also included are the mean real-world searcher CG values for each interface.

crowdsourced user study, as reported in Section 8.2. From this user study, we then subsequently used interaction data to ground an extensive set of simulations of interaction, reported in Section 8.3. These simulations were used to determine how each of the twelve stopping strategies proposed in Chapter 5 performed and approximated the mean stopping behaviours of searchers. In turn, these findings provide answers to our two high-level research questions HL-RQ3a and HL-RQ3b, when tasks and goals were varied.

Considering the user study first, we trialled four different experimental conditions. These were D AS (diversified system, aspectual task), ND AS (non-diversified system, aspectual task), D AD (diversified system, ad-hoc task) and ND AD (non-diversified system, ad-hoc task). Under system ND, results were returned with the standard BM25 retrieval model ($\beta = 0.75$). For system D however, results were re-ranked according to a diversification algorithm. This helped ensure that a more diverse set of results for the given topic were returned. It used entities contained within each document to perform the re-ranking. Considering tasks, searchers were either asked to save four relevant documents (for task AD), or save at least four relevant documents that contained at least one new entity related to the topic being examined (for task AS).

While significant differences existed between the performance of systems ND and D, no significant differences were observed when considering searcher behaviours, although trends were observed. When using system D, subjects did offer improvements on average, saving more TREC relevant documents and TREC relevant documents with at least one new entity. Evidence was also found to support our IFT-based hypotheses. However, the lack of significant differences between the experimental conditions – and (perhaps) the small sample size – may suggest that bigger differences may be required between systems. Individuals may then be able to subjectively report on whether the two would yield different levels of performance.

From the user study, we then took the interaction data to derive a series of interaction probabilities and costs. In turn, these were used to ground an extensive set of simulations of interaction. Split across performance (addressing HL-RQ3a) and comparison runs (addressing HL-RQ3b), we examined each individual stopping strategy both in terms of overall performance and how well they approximated actual searcher stopping behaviours. These were run across each of the four experimental conditions. Findings for HL-RQ3a show that all twelve stopping strategies offered reasonable levels of CG. We found that stopping strategies SS1-FIX and SS5-COMB offered the best levels of CG (with SS1-FIX performing best under condition ND AD). Only stopping strategy SS8-IFT consistently yielded significantly poorer levels of performance than the best performing strategies, perhaps due to how the strategy was instantiated.

When considering HL-RQ3b, we found that under system D, SS1-FIX offered the best approximations to searcher behaviours, with SS10-RELTIME providing the best approximations for system ND. This is an interesting result – and perhaps can be attributed to the fact that system D offered better performance on average than system ND. Intuitively, the result makes sense. Given the higher levels of performance afforded by D, more relevant documents mean that following a fixed-depth approach on average would most likely yield greater benefits. On the contrary, with the poorer performance offered by ND, an adaptive approach would make more sense when fewer relevant documents were presented. However, when comparing the simulation results to the real-world means, we found that the simulated counterparts of the real-world subjects performed poorly across all conditions, and may be utilising improved stopping criteria than the twelve operationalised stopping strategies trialled. Further work will be required to examine this.

Along with results from Chapter 7, findings from this chapter will be discussed in detail in Chapter 10. From an examination of our result summary level stopping strategies, we now turn our attention towards our final contributory chapter. Here, we experiment with our new stopping decision point – *SERP level stopping*.

Chapter 9

Modelling SERP Level Stopping Behaviours

In Chapters 7 and 8, we reported on two user studies that examined the effects of searcher behaviour and performance under different search contexts. Interaction data from these studies were then used to ground simulations of interaction that examined how each of the twelve different stopping strategies performed and approximated real-world searcher stopping behaviours. We now turn our attention towards providing a complete answer to our first high-level research question, HL-RQ1.

- HL-RQ1 How can we improve searcher models to incorporate different stopping decision points?

In order to address this research question, we presented in Chapter 4 the *Complex Searcher Model (CSM)*, a conceptual, high-level model of the search process. The CSM introduced the SERP level stopping decision point, motivated by information scent. This new stopping decision point allows a simulated searcher to *abandon* a SERP if a general *overview* of the given SERP shows that the results did not appear to provide promising results. With the definition of the CSM partially satisfying HL-RQ1, in this we chapter provide the results of an empirical study using the new stopping decision point.

9.1 Motivation and Research Questions

We begin with the concept of SERP abandonment (discussed previously in Section 4.4.4 on page 117), before considering how *Information Foraging Theory (IFT)* provides strong theoretical motivation. The concept behind the new SERP level stopping decision point revolves around the notion of SERP abandonment, when a searcher fails to click on any of the results returned for a given query (Diriye et al., 2012; Hassan and White, 2013). This may occur for a variety of reasons, both good and bad. The primary motivator for this study considers the notion of bad abandonment, where searchers abandon a SERP because they are dissatisfied by the results returned (Hassan and White, 2013).

As we discussed in Section 3.3.1.1 on page 92, Pirolli and Card (1999) argue that information seekers are like animals foraging in the wild, and as such will follow a scent to find food. As discussed previously, information seekers have been shown to follow a series of *proximal cues* provided by SERP components such as hypertext links, titles, snippets and thumbnails to help locate relevant information (Pirolli and Card, 1995, 1999; Chi et al., 2001; Olston and Chi, 2003; Pirolli, 2007). For example, Card et al. (2001) found that when navigating through webpages, searchers were more likely to leave when the information scent provided on a page began to decline. Work by Wu et al. (2014) discussed a user study where low, medium and high scent SERPs were created by changing the number and distribution of relevant items on the page – thus altering the proximal cues provided. Those interacting with high scent SERPs examined more content and went to greater depths compared to those who utilised low scent SERPs. Further work by Ong et al. (2017) – and indeed the user study reported in Section 7.2 – all confirm that modifying the scent of a SERP does indeed alter a searcher's stopping behaviour.

For this chapter, we operationalise the information scent as the performance of a given SERP, examining how the new SERP level stopping decision point within the searcher model – as

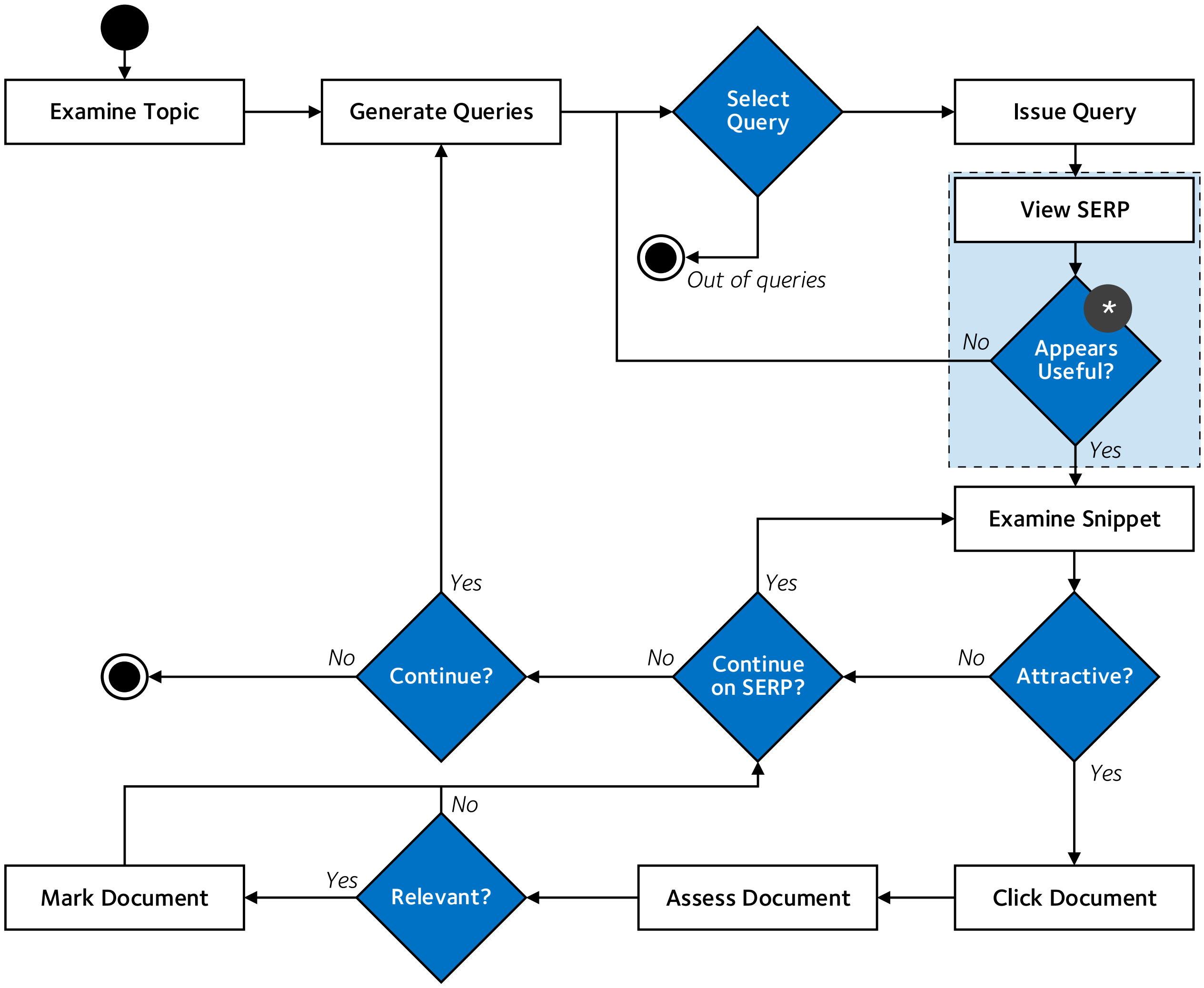


Figure 9.1 The *Complex Searcher Model (CSM)*, highlighting the stopping decision point (by an asterisk*, with the SERP examination component also highlighted within the blue rectangle) that is examined in detail in this chapter. Refer to Section 4.1 for an in-depth explanation of the model.

shown in Figure 9.1[1] – affects searcher, stopping and overall performance. This is achieved by enumerating a series of different SERP level implementations, allowing us to operationalise the new stopping decision point in several ways. As such, we pose two key research questions to be addressed in this chapter.

SERP-RQ1 Does the incorporation of a SERP level stopping decision point lead to improved overall performance?

[1] Further information on the *Complex Searcher Model (CSM)* can be found in Chapter 4, starting on page 107.

> **SERP-RQ2** Does the incorporation of a SERP level stopping decision point lead to improved approximations of searcher stopping behaviour?

Taken together, the answers to these research questions will provide us with a complete answer to high-level research question **HL-RQ1**. This is in conjunction with the CSM proposed in Chapter 4. In the next section, we outline the methodology undertaken to address the aforementioned research questions.

9.2 Methodology

In order to address the two research questions posed above, we followed general methodology. This is detailed in Section 6.4 on page 157. A variety of different simulation components that mapped to individual components of the CSM were left unchanged from the general methodology. One can assume that all components are left unchanged, save for changes to our experimental setup outlined here. The component of interest for the work in this chapter is the SERP level stopping decision point.

In this section, we outline:

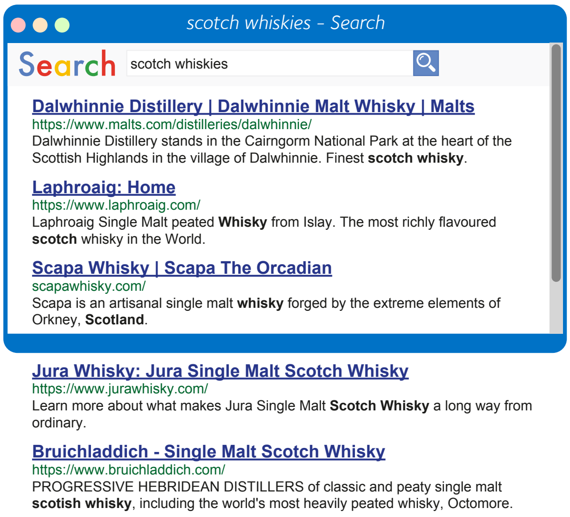
- the different SERP level stopping decision point implementations that were trialled, including the introduction of a new interaction probability concerning the likelihood of examining a SERP (Section 9.2.1); leading onto
- a discussion of the different interaction probabilities and costs that were used for this study (Section 9.2.2);
- an enumeration of the different result summary level stopping strategies trialled for this study (Section 9.2.3); and
- a summary of the other components of the CSM that departed from the general methodology (Section 9.2.4).

9.2.1 SERP Decision Making

This section discusses the various ways in which we implemented this new stopping decision point. As a searcher can only obtain an impression of the overall quality of a SERP from what he or she can see *at a glance*, we begin this section with a discussion on the size of the browser's viewport.

Considering Browser Viewport Size Real-world searchers are able to infer the quality (and perhaps relevance) of a given page or SERP through the examination of various proximal cues (Chi et al., 2001). Such cues are not considered in this work. Instead, we rely upon more simplistic means to implement the stopping decision point. One aspect of a SERPs presentation that we do consider in this chapter is the *size of the browser's viewport*. A SERP is typically larger than the viewport within which it is displayed, leading to the inclusion of scrollbars. Results can only be seen *above-the-fold*, or what is visible within the viewport.

We argue that a searcher can infer the quality of the SERP from the initial view with which they are presented, and thus incorporate a *viewport size* (v_{size}) variable in our simulations of interaction – a searcher can only judge what they can see. This variable can vary between the different interfaces we trialled. For example, longer snippet text resulted in fewer result summaries being displayed in the initial view. By using a fixed-size popup window in the two user studies (as discussed in Section 6.2.2), we were able to manually check the number of result summaries displayed within the popup window, and use these values to provide more extensive grounding to the new stopping decision point.


scotch whiskies - Search

Search scotch whiskies

Dalwhinnie Distillery | Dalwhinnie Malt Whisky | Malts
https://www.malts.com/distilleries/dalwhinnie/
Dalwhinnie Distillery stands in the Cairngorm National Park at the heart of the Scottish Highlands in the village of Dalwhinnie. Finest **scotch whisky**.

Laphroaig: Home
https://www.laphroaig.com/
Laphroaig Single Malt peated **Whisky** from Islay. The most richly flavoured **scotch** whisky in the World.

Scapa Whisky | Scapa The Orcadian
scapawhisky.com/
Scapa is an artisanal single malt **whisky** forged by the extreme elements of Orkney, **Scotland**.

Jura Whisky: Jura Single Malt Scotch Whisky
https://www.jurawhisky.com/
Learn more about what makes Jura Single Malt **Scotch Whisky** a long way from ordinary.

Bruichladdich - Single Malt Scotch Whisky
https://www.bruichladdich.com/
PROGRESSIVE HEBRIDEAN DISTILLERS of classic and peaty single malt **scotish whisky**, including the world's most heavily peated whisky, Octomore.

Figure 9.2 The SERP viewport threshold. In this example, three result summaries are visible, with two present but outwith the viewport. Therefore, $v_{size}=3$.

For simulations of interaction reported in Chapter 7, different values of v_{size} could be used over the four experimental interfaces trialled. This was because for each interface, result summary lengths varied. As we tended from interface T0 (no snippet text) to T4 (four snippet fragments), longer result summaries would impact upon the number of result summaries visible in the initial view. Longer result summaries would mean that fewer would be presented within a viewport of the same size, compared to shorter result summaries. However, result summary lengths were not modified under the experimental conditions trialled in Chapter 8. This means that v_{size} would (on average) remain constant, with fixed popup window dimensions resulting in two lines of snippet text per result summary.

Definition: Low vs. High Scent Given a SERP, would it constitute as *low scent* or *high scent?* For this chapter, we follow the work of Wu et al. (2014). In their study, the authors state that a low scent SERP offers little or no relevant content. Definitions by Wu et al. (2014) and Hassan and White (2013) define a poor scented SERP as $P@10 = 0.0$. We take this definition to delineate between *good* and *bad* SERPs, and extend it by also considering v_{size}. This leads to our definition of $P@v_{size} = 0.0$ for a poor quality SERP, meaning that a simulated searcher would then gauge the quality of a SERP by examining the average number of result summaries displayed within a fictional browser viewport for the given experimental interface or condition being trialled. The definition of low and high scented SERPs was also used for SS11-COMB, as defined in Section 5.5 on page 131.

Probability of Examination For this new stopping decision point, we introduce the *probability of examining a SERP*, or P(E). Given a SERP presented to a searcher, this probability determines the likelihood that the searcher will *enter* the SERP (based upon its information scent) and begin to examine result summaries in detail. Taking this concept further, we can then consider two further probabilities of interaction that incorporate the notion of a SERPs information scent, yielding:

- P(E|HS), the probability of examining a SERP perceived to give a high information scent (i.e. a good quality SERP); and

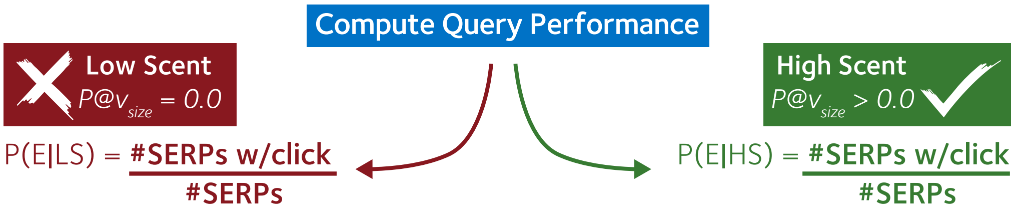

Compute Query Performance
Low Scent
P@v size = 0.0
High Scent
P@v size > 0.0
P(E|LS) = #SERPs w/click / #SERPs
P(E|HS) = #SERPs w/click / #SERPs

Figure 9.3 An illustration highlighting how the different SERP examination costs were computed. As an example, low scented SERPs offering $P@v_{size} = 0.0$ are selected, with the calculation for *P(E|LS)* then taking place. We consider both the probability of examining SERPs yielding both high and low information scents. The definitions for low and high scented SERPs using v_{size} are adapted from Wu et al. (2014) and Hassan and White (2013).

- **P(E|LS)**, the probability of examining a SERP offering what appears to be a low information scent (i.e. a poor quality SERP, or $P@v_{size} = 0.0$).

These values were computed from interaction log data, taken from the two user studies reported in Chapters 7 and 8. Computed values derived are not reported in this section; refer to Section 9.2.2 for the probabilities. Intuitively, one would expect a searcher demonstrating competency at searching for information to know when a query is returning good results and vice versa. As such, one would expect to see a higher probability for $P(E|HS)$ than when compared to $P(E|LS)$, and would provide evidence that searchers do indeed attempt to avoid low quality SERPs.

As illustrated in Figure 9.3, we took each query issued from the interaction log of each user study, and extracted for each the $P@v_{size}$ score (as per Wu et al. (2014)), considering $P@v_{size} = 0.0$ as our criterion for a SERP of poor scent. For the interactions recorded on each SERP, we could then count the number that recorded no clicks (meaning no result summaries were deemed to be attractive enough to examine further). We considered this as a definition of an **abandoned SERP**, as used in previous work by Hassan and White (2013). From these counts, we could then compute the probabilities of examining a SERP, as illustrated in Figure 9.3.

9.2.1.1 Decision Point Implementations

We trialled three different implementations of the SERP level stopping decision point, providing us with the ability to explore the effect of incorporating it within the CSM. These are enumerated below, with an explanation of each. The first can be considered our baseline approach.

- **SERP Always** Considered our baseline, a searcher subscribing to this implementation will always enter the SERP and examine at least one result summary – the exact number would be determined by the result summary level stopping strategy. This is the generally accepted approach as used in prior simulations of interaction. As such, we consider this to be our baseline implementation. As a reminder, this implementation was used in the simulations reported in Chapters 7 and 8.

From here, the remaining two strategies begin to consider a simulated searcher's judgements regarding the perceived quality of a SERP, and thus begin to use the new stopping decision point to abandon a SERP before examining individual result summaries in detail.

- **SERP Perfect** Here, a simulated searcher will only begin to examine a SERP in detail if $P@v_{size} > 0$ (considering the viewport size). If $P@v_{size} = 0$, the searcher will abandon the SERP, and proceed to the next action as dictated by the CSM. This is the upper bound in terms of performance for the stopping decision point, and is analogous to, as an example, the *ideal user* of Hagen et al. (2016).
- **SERP Average** This final implementation used a stochastic element to determine whether the simulated searcher should enter the SERP or not. Like above, the viewport size ($P@v_{size}$) of the SERP is computed. If the SERP is of high scent, $P(E|HS)$ is used to determine whether the searcher should enter the SERP. Conversely, if the SERP is considered to be of low scent, $P(E|LS)$ is used instead to determine the likelihood

of abandonment. We considered the probabilities of interaction for a given interface or condition, taking the average.

Given the three implementations, one would intuitively expect SERP Perfect to yield simulated searcher that attain the highest overall levels of CG. These searchers would not waste time examining poor scented SERPs, and instead spend their time examining SERPs that will return at least one relevant document. For the implementation that offers the best approximations of real-world stopping behaviours however, SERP Perfect may not be best. It depends how well real-world subjects were able to discern from good and poor scented SERPs. It is more likely that SERP Average will provide the better approximations of real-world behaviours.

By considering the three different approaches to implementing the new SERP level stopping decision point, we can then clearly identify whether improved performance and improved approximations of actual searcher stopping behaviours are offered. We also trialled each of the stochastic SERP level stopping decision components a total of 10 times, computing the average over the different trials. Given that the decision maker components of the **SimIIR** framework were run a total of 50 times each (responsible for determining the attractiveness of result summaries and relevancy of documents), this made the addition of a stochastic SERP level stopping decision point expensive in terms of the number of additional runs that were required.

9.2.2 Interfaces, Conditions, and Experimental Grounding

To determine whether the new SERP level stopping decision point implementations offered improvements, we conducted a series of simulations across interfaces and conditions trialled in the two user studies, reported earlier in Chapters 7 and 8.

From Chapter 7, the different experimental interfaces – whereby result summary lengths were manipulated – were considered. Namely, these were T0, T1, T2 and T4. From

Table 9.1 Probabilities of examining high (**P(E|HS)**) and low scented SERPs (**P(E|LS)**), along with $\mathbf{v_{size}}$ values for each of the experimental interfaces and conditions trialled in this chapter. Statistical tests between interfaces/conditions yielded no significant differences, at α=0.05. Probabilities that are used in the experiments reported in this chapter are highlighted. Refer to Tables 7.6 (page 217) and 8.10 (page 283) for other interaction probabilities and costs for the studies reported in Chapters 7 and 8.

		T0	T1	T2	T4
Chapter 7	P(E|HS)	0.76	0.79	0.78	0.78
	P(E|LS)	0.27	0.40	0.31	0.40
	v_{size}	10	9	7	6

		D-AS	ND-AS	D-AD	ND-AD
Chapter 8	P(E|HS)	0.76	0.76	0.73	0.75
	P(E|LS)	0.29	0.37	0.26	0.34
	v_{size}	7	7	7	7

Chapter 8, we also considered the four experimental conditions that manipulated the underlying system and searcher tasks: D AS, ND AS, D AD and ND AD.

From the interaction data of the two user study interaction logs, we could then compute the probabilities of subjects examining low scented ($P(E|LS)$) and high scented ($P(E|HS)$) SERPs. Values were computed as per the explanations provided in Section 9.2.1. The computed values are reported in Table 9.1, along with the corresponding v_{size} value for each interface or condition, denoting the number of result summaries visible within a simulated viewport. From examination of the table, we can see that the probabilities for both sets are very similar across all interfaces and conditions. Indeed, a two-tailed Student's t-test yielded no significant differences across the four interfaces or conditions where $\alpha = 0.05$. Given the close proximity of the probabilities (and the subsequent lack of differences that we would likely observe), we simplified our experimentation. We chose to run experiments for one interface and condition per study, selecting T2 and ND AD. These were selected as they represent a standard search interface and task.

Values reported in Table 9.1 should be considered in tandem with the interaction costs and probabilities reported in Tables 7.6 (page 217) and 8.10 (page 283). These tables report interaction costs (such as querying and document examination costs) and other probabilities (considering the probabilities of clicking on result summaries, $P(C|R)$ and $P(C|N)$ – and saving documents, $P(S|R)$ and $P(S|N)$).

9.2.3 Result Summary Level Stopping Strategies

Chapter 5 presented twelve different result summary level stopping strategies. To further reduce the complexity of the experimentation reported in this chapter, we decided to reduce the number of strategies that we considered. Doing so reduced the risk of potentially repeating the same results as shown before, while still demonstrating that when enabled, the SERP level stopping decision point yielded improved performance (and closer approximations) to actual searcher stopping behaviour – while still reporting over a range of configurations.

We only report the results of three result summary level stopping strategies in this chapter. These were selected as they offered good performance and approximations of actual searcher stopping behaviours in results presented in previous chapters. We used:

- SS1-FIX, the fixed-depth baseline result summary level stopping strategy;
- SS2-NT, the frustration stopping strategy, considering the total number of non-relevant summaries encountered; and

- SS5-COMB, the combination stopping strategy combining frustration-based strategy SS2-NT and satisfaction-based strategy SS4-SAT.

Definitions for each of these stopping strategies can be found in Chapter 5, starting on page 121. Note that these strategies were instantiated with the same x_n values *(stopping threshold values)* reported in the general methodology in Section 6.4.2.6 (page 173).

9.2.4 Remaining CSM Components

The remaining CSM components were configured as presented in Section 6.4.3.1. For comparison runs, queries issued by real-world subjects under T2 and ND AD were again *replayed* in the simulations, the process of which is outlined in Section 6.4.3.2 on page 182. Specific implementation details for interface T2 were the same as described in Section 7.3.1 (page 216), with the methodology in Section 8.3.1 (page 281) followed for ND AD.

9.3 Results

We now report the results of our simulations of interaction, involving the new SERP level stopping decision point. As in previous chapters, we discuss our findings over two subsections, considering:

- performance runs (Section 9.3.1), where we discuss the highest levels of performance attained by simulated searchers under different *what-if* scenarios; and
- comparison runs (Section 9.3.2), where we provide results of the simulations that were directly compared to actual mean searcher stopping behaviours.

Both of these runs provide answers for this chapter's two research questions, SERP-RQ1 and SERP-RQ2. Refer to Section 9.1 for information on the questions posed.

Significance Testing While the difference between individual result summary level stopping strategies is interesting, the main focus of the results reported in this chapter is the difference in performance and approximations offered across individual SERP level decision point implementations. Statistical tests in this chapter are therefore performed across different SERP level decision point implementations to determine if significant differences exist between them. All tests reported in this chapter utilise the two-tailed Student's t-test, where $\alpha = 0.05$. Unlike Sections 7.3 and 8.3, we this time examine results *for* significant differences between the implementations.

9.3.1 Performance

Figure 9.4 presents six individual plots, each representing one of the three different result summary level stopping strategies trialled (from SS1-FIX, SS2-NT and SS5-COMB). The three plots on the top correspond to results over interface T2 (Chapter 7), with the bottom three plots corresponding to condition ND AD (Chapter 8). Each plot represents the mean level of CG that is attained across the runs (represented on the y axes) across varying mean depths per query (along the x axes). In each plot, we represent the three individual SERP level stopping decision points, with our baseline SERP Always, SERP Average and SERP Perfect presented. This means that we can observe how performance varies across different depths, over different result summary level stopping strategies – and over a different interface/condition.

From an initial observation of the plots, we note a number of different (and consistently occurring) trends. As with the plots in Figures 7.6 (page 222) and 8.7 (page 289), we observe that at low mean depths per query, all three SERP level decision point implementations offer similar levels of CG. CG then steadily rises up to a peak as the mean depth per query increases. Once this peak has been reached, the performance then begins to slowly tail off, or remain relatively invariant across greater mean depths per query. Of particular relevance to SERP-RQ1 is the difference in CG attained by the three different SERP level stopping

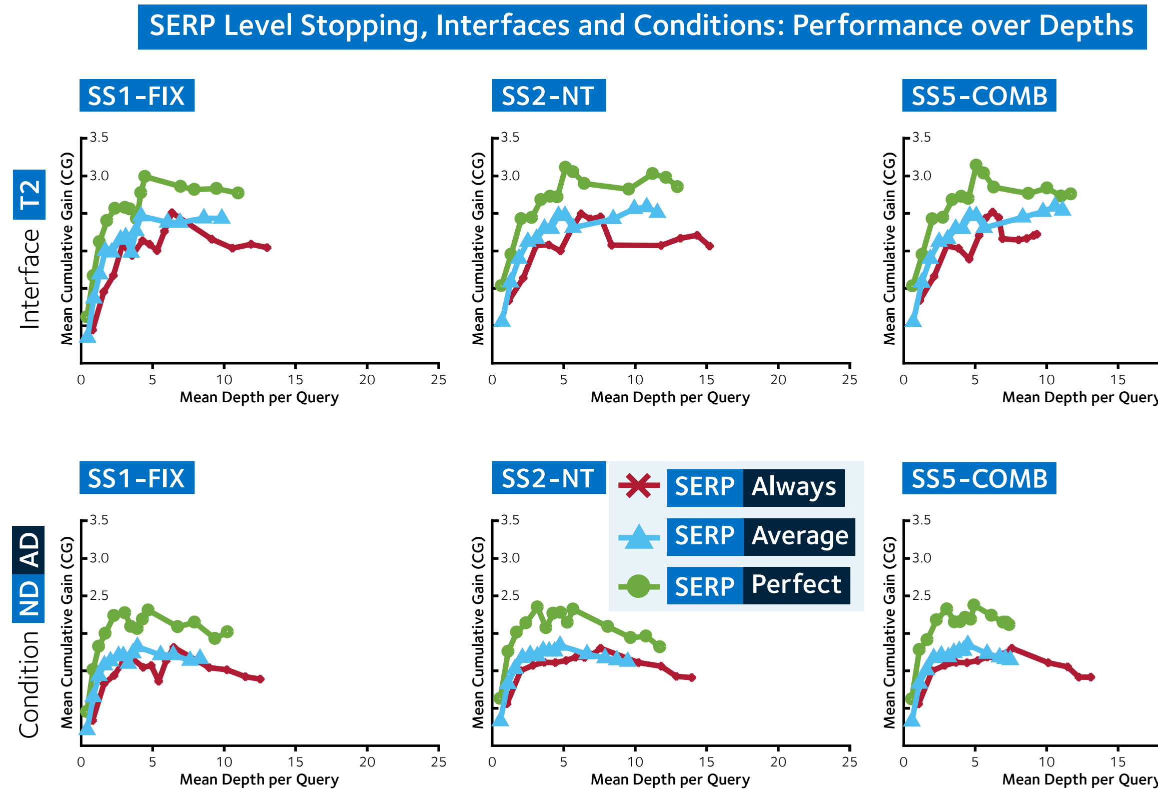


Figure 9.4 Plots demonstrating the varying levels of performance, measured in CG, over the mean depth per query. Plots are for the *what-if* simulated performance runs. Each result summary level stopping strategy is plotted separately, with the three different SERP level decision point implementations shown. Plots on the top relate to interface T2; plots on the bottom relate to condition ND AD.

decision point implementations. As previously mentioned, all three start from a similar point at shallow depths per query, across all stopping strategies and the interface/condition. However, as the mean depth per query increases, we observe that the performance across the three different implementations begins to diverge from one another.

We consistently find that as the mean depth per query increases, the SERP Always (baseline) implementation consistently offers the lowest levels of CG, and the SERP Perfect implementation consistently offers higher levels of mean CG. This is an intuitive result; avoiding SERPs that offer a poor scent means that you are likely to invest more time in issuing queries that offer better results, thus identifying and saving more relevant documents. This consistent improvement also provides evidence for addressing SERP-RQ1 – incorporating a SERP level stopping decision point does indeed lead to higher overall performance. The final implementation SERP Average presents further evidence to support this claim, with performance generally performing better than the baseline SERP Always implementation. However, it should be noted that there are certain points where SERP Always does outperform SERP Average. For example, examining the plot in Figure 9.4 for interface T2 over stopping strategy SS2-NT, one can see that SERP Always outperforms SERP Average at a mean depth per query from around 6 to 8. This may be because that a simulated searcher may decide to skip some queries judged to yield a poor scented SERP, and would therefore have the time to issue more queries later on in the session. As we do not consider previously examined content in the initial SERP judgement, a simulated searcher may enter subsequent SERPs without any additional relevant content to mark, lowering their overall mean CG. We discuss this potential limitation later in Chapter 10.

Given the evidence supporting SERP-RQ1, we now turn our attention to the absolute best performance that each of the stopping strategies yield, across each interface/condition and over the three SERP level stopping decision point implementations. Table 9.2 provides these values, with values for interface T2 reported first, and condition ND AD reported underneath. For each stopping strategy and SERP level stopping decision point combination, we report: the highest level of **CG** attained; the mean depth per query (**DQ**) that this was attained at, and the stopping threshold(s) ($\mathbf{x_n}$) that were used. For combination stopping strategy SS5-COMB, two stopping threshold values were used for x_2 and x_4. These are presented in Table 9.2 in this order.

From Table 9.2, we find results that complement the findings observed in Figure 9.4. In the table, we highlight the SERP level stopping decision point implementation and stopping strategy combination yielding the highest level of mean CG. Unsurprisingly, these are all obtained with the SERP Perfect SERP level stopping decision point implementation. Indeed, a general trend from the table can be observed – improvements in the highest levels of CG can be clearly seen as we tend from left to right, or SERP Always to

Table 9.2 Results from the simulated *what-if* simulated performance runs, showing the highest levels of CG attained for result summary level stopping strategies SS1-FIX, SS2-NT and SS5-COMB. These are reported over the SERP Always *(baseline)*, SERP Average and SERP Perfect SERP level stopping decision point implementations. $\mathbf{x_n}$ denotes the stopping parameter threshold(s), with **DQ** denoting the depth per query at which the greatest **CG** value was attained at. Note that for combination strategy SS5-COMB, x_2,x_4 are shown as $\mathbf{x_n}$ column. Stopping strategy/SERP decision point implementation combinations that yield the highest CG values are highlighted.

		Always			Average			Perfect		
		x_n	DQ	CG	x_n	DQ	CG	x_n	DQ	CG
Int. T2	SS1-FIX	10	6.33	2.50	10	4.17	2.45	10	4.45	2.98
	SS2-NT	7	6.16	2.49	21	10.77	2.58	8	5.07	3.11
	SS5-COMB	8,4	6.17	2.52	21,9	10.54	2.59	8,7	5.04	3.14
Cdn. ND AD	SS1-FIX	10	6.47	1.81	10	3.93	1.82	10	4.69	2.30
	SS2-NT	10	7.55	1.80	10	4.71	1.84	5	3.12	2.35
	SS5-COMB	10,6	7.55	1.80	10,4	4.43	1.85	10,3	4.87	2.37

SERP Perfect. The biggest increase in maximum CG that can be observed from the table is for SS5-COMB over interface T2, with mean CG rising from 2.52 and a mean depth per query of 6.17 to 3.14, at a mean depth per query of 5.04. Interestingly, the rises in CG are more profound over T2 generally when compared to the results obtained over condition ND AD.

Closer inspection of the values reported for the SERP Always SERP level stopping decision point implementation reported in Table 9.2 can also be undertaken. These values are considered as our initial baseline, and are essentially the values attained by the simulated searchers when instructed to browse every resultant SERP. As such, this is the configuration that is employed in the results of the simulations of interaction we presented in Chapters 7 and 8. Specifically, values of interface T2 reported in Table 7.8 on page 224 match those reported in Table 9.2 above. Indeed, this is also confirmed for condition ND AD, with results reported in Table 8.12 presented on page 292 again matching those in Table 9.2.

Considering the stopping threshold(s) and mean depths per query across the result summary level stopping strategies, we see that SS1-FIX @10 consistently offers the highest mean CG, with a slight drop in the mean depth per query at which the highest CG score was attained. Indeed, this trend can broadly be observed across the other two stopping strategies, and over both interface T2 and condition ND AD. An exception to this trend is observed over SS2-NT and SS5-COMB under the SERP Average SERP level stopping decision point implementation. Greater stopping thresholds are observed for x_2, with a resultant greater mean depth per query ($\approx$ 10.6 over interface T2, compared to $\approx$ 6.16 over SERP Always). Despite this, we see that a fixed depth approach holds up remarkably well, showing that when issuing a performant query, a fixed approach will offer good returns. However, overall, we find that adaptive strategies SS2-NT and SS5-COMB outperform SS1-FIX.

Evidence has thus far led to trends supporting SERP-RQ1 – the new SERP level stopping decision point implementation does indeed yield improvements in performance. However, are the improvements offered by SERP Average and SERP Perfect significant improvements over our baseline implementation, SERP Always? As outlined at the start of Section 9.3, we ran a series of two-tailed Student's t-tests to determine whether this was the case. Tests were run comparing the best performing implementation, SERP Perfect, against both SERP Average and SERP Always, over each of the result summary level stopping strategies, as well as over interface T2 and ND AD.

Between SERP Perfect and SERP Always and SERP Perfect and SERP Average, significant differences considering the levels of CG were observed. This was true across all result summary level stopping strategies, over both T2 and ND AD. Considering SS1-FIX over interface T2, we observed the following:

- SERP Perfect → SERP Average: $SD = 2.43, t(2748) = 3.25, p = 0.001$; and
- SERP Perfect → SERP Always: $SD = 2.46, t(498) = 2.19, p = 0.03$.

We also ran comparisons between SERP Average and SERP Always, to determine if a significant difference existed there. Unsurprisingly, no significant difference was observed, with $p = 0.78$ reported over SS1-FIX and interface T2. However, results clearly demonstrate that the upper bound SERP Perfect implementation yields significant performance improvements over the existing baseline approach, across all stopping strategies and the interface/condition. This solidifies our supporting evidence for SERP-RQ1.

9.3.2 Real-World Comparisons

From our *what-if* performance simulations, we now examine how closely each of the aforementioned stopping strategies compares to actual searcher behaviours. Therefore, this section provides an answer to SERP-RQ2. As a reminder, these simulations *replayed* all of the queries issued by real-world searchers, allowing us to compare real-world and simulated click depths. In turn, we could then see if the inclusion of the SERP level stopping decision point improved approximations.

Figure 9.5 presents six plots, one for each of the three result summary level stopping strategies: SS1-FIX; SS2-NT; and SS5-COMB. These are duplicated over interface T2 and condition ND AD. Each of the plots illustrates the mean depth per query to which searchers traversed result lists to (represented along the x axis). This is plotted against the MSE[2] of the real-world vs. simulated searcher click depths (thus considering stopping

[2] Refer to Section 6.4.3.2 on page 182 for further information on how we computed the *Mean Squared Error (MSE)* for our comparisons.

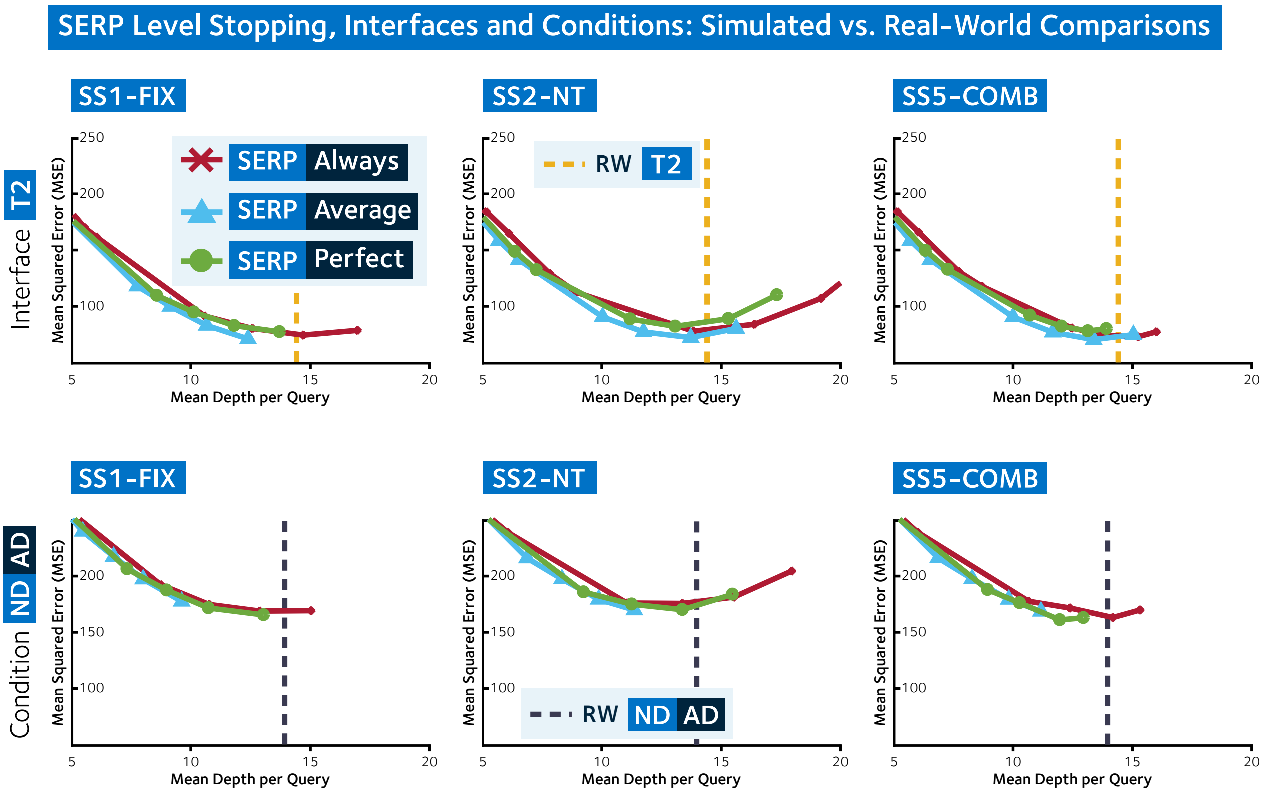


Figure 9.5 Plots reporting the comparison runs, reporting the MSE vs. the mean depth per query. Runs over interface T2 (top) and condition ND AD (bottom) are shown, for the three trialled result summary level stopping strategies. SERP level decision point implementations are also shown. Also included are dashed lines denoting the mean depth per query reached by the real-world subjects of the corresponding user study. Note that the mean depth per query is limited from *5* to *20* to highlight what happens around the mean real-world click depths.

behaviours). Each point on the plotted lines represents how close the click depth approximation was on average for a given stopping threshold parameter configuration. The closer the MSE value tends towards zero, the closer the simulated searcher's approximation to actual stopping behaviours. Each plot also presents one of the three trialled SERP level stopping decision points, from SERP Always, SERP Average, and SERP Perfect. The plots demonstrate how approximations from each of the implementations differ across interfaces/conditions and result summary level stopping strategies. We also include the mean real-world click depths for a straightforward visual comparison, represented as vertical dashed lines. These differ between interface T2 and ND AD, as shown in Fig-

Table 9.3 Results from the simulated comparison runs, showing the *lowest* **MSE** value reached over each of the three result summary level stopping strategies trialled. These are reported over the SERP Always *(baseline)*, SERP Average and SERP Perfect SERP level stopping decision point implementations. The stopping strategy yielding the lowest MSE for both T2 and ND AD are highlighted. Note that for combination strategy SS5-COMB, x_2,x_4 are shown as $\mathbf{x_n}$ column.

		Always		Average		Perfect	
		x_n	MSE	x_n	MSE	x_n	MSE
Int. T2	SS1-FIX	21	74.04	24	**70.68**	24	76.94
	SS2-NT	15	77.76	21	**72.12**	18	82.06
	SS5-COMB	21,6	72.81	21,10	**70.23**	21,7	77.89
Cdn. ND AD	SS1-FIX	21	168.39	24	176.98	24	**164.97**
	SS2-NT	18	175.39	24	**169.52**	21	169.85
	SS5-COMB	21,5	162.81	24,8	168.26	21,5	**160.69**

ures 7.8 (page 232) and 8.9 (page 297), respectively. Note also truncated x and y axes. As all plotted lines were close together, we altered the axes to better highlight the approximations when the MSE values were at their lowest.

At a glance, the small difference between MSE values demonstrates that there is not much of a difference between the three SERP level stopping decision point implementations. To aid in the reporting of our results, we also include a table of MSE values. Table 9.3 reports the lowest MSE values that were attained across each of the SERP level stopping decision point implementations and stopping strategies, over T2 and ND AD. Along with the MSE values are the stopping parameter threshold(s) that were used to attain the lowest MSE for each combination. For combination stopping strategy SS5-COMB, we once again report

x_2 and x_4 for the thresholds, in that order. We also highlight in the table the lowest MSE values over the three SERP level stopping decision point implementations, considering each result summary level stopping strategy in turn.

From Table 9.3, we note that over interface T2, the SERP Average SERP level stopping decision point consistently yielded the lowest MSE values over the three result summary level stopping strategies trialled. This is closely followed by our baseline approach, SERP Always, with SERP Perfect consistently offering the poorest approximations of mean real-world searcher approximations. This result is intuitive – real-world searchers would have been unlikely to correctly judge the scent of a SERP with 100% accuracy, and thus would mean that the upper bound SERP Perfect implementation would be furthest from mean real-world stopping behaviours. In contrast, a stochastic approach offered by SERP Average would intuitively yield better approximations. Although real-world searchers would not have rolled a die to determine whether a SERP is worth examining, they would have had the flexibility to abandon SERPs that they felt did not offer a good scent. This flexibility is provided to the SERP Average simulated searcher. An interesting observation for interface T2 is the higher stopping parameter threshold(s) that were found to offer the lowest MSE values. This demonstrates that the mean real-world searcher stopping behaviours over this interface is that of a tolerant searcher, who is willing to examine results to greater depths on average, before deciding to stop.

The trends that we observe for interface T2 in Table 9.3 can also be demonstrated by close examination of the corresponding (top) plots in Figure 9.5. Close examination shows that across varying mean depths per query, the general trends observed from Table 9.3 hold – SERP Perfect and SERP Always consistently offered poorer approximations than SERP Average, which consistently offered the lowest MSE values.

Results over interface ND AD are different from those of interface T2. Examining Table 9.3, we observe that SERP Average yielded the best mean searcher stopping approximation for SS2-NT @24. Interestingly however, we find that the lowest MSE values

Table 9.4 Additional results from the searcher comparison runs, reporting the mean depth per query (**DQ**), the mean level of **CG**, and mean interactive precision (**iP**). These values were attained using the configurations yielding the lowest MSE (refer to Table 9.3), indicating the best approximation to real-world stopping behaviours. We also include the mean real-world searcher values (**RW**) over T2 and ND AD for a direct comparison.

		Always			Average			Perfect		
		DQ	CG	iP	DQ	CG	iP	DQ	CG	iP
Interface T2	RW	14.39	2.36	2.47	14.39	2.36	2.47	14.39	2.36	2.47
	SS1-FIX	14.67	2.14	1.41	12.35	1.82	1.17	13.68	2.16	1.41
	SS2-NT	13.81	2.15	1.42	13.70	1.98	1.27	13.06	2.22	1.45
	SS5-COMB	15.22	2.08	1.34	13.35	1.93	1.25	13.11	1.99	1.30
Condition ND AD	RW	13.94	3.49	2.22	13.94	3.49	2.22	13.94	3.49	2.22
	SS1-FIX	12.89	2.55	1.66	9.61	1.84	1.20	13.06	2.59	1.69
	SS2-NT	13.34	2.73	1.77	11.33	2.12	1.38	13.35	2.78	1.84
	SS5-COMB	14.16	2.69	1.75	11.14	2.05	1.34	11.94	2.53	1.66

for SS1-FIX @24 and SS5-COMB @21,5 were yielded by the SERP Perfect SERP level stopping decision point implementation. This perhaps demonstrates that the change in task goals (from time-limited for T2 to find x for ND AD) influences the SERP level decision making of real-world searchers on average. Results show that under SS1-FIX and SS5-COMB, searchers under ND AD were better able to discern between poor and high quality SERPs. We provide a discussion into this result in Section 10.2.

Table 9.4 also provides additional information on the comparison runs. The table reports the mean depth per query (**DQ**), **CG** and interactive precision (**iP**) attained across each of the SERP level stopping decision point implementations and result summary level stopping strategies, again over interface T2 and condition ND AD. To allow for easy comparison, we also include the real-world mean depth per query, CG and interactive precision values (row **RW**) across the interface and condition examined.

Trends from Table 9.4 show that as we tend from SERP Always to SERP Average, we generally observed a *drop* in the CG attained by the simulated searchers. Over interface T2 for example, CG for SS2-NT @15 dropped from 2.15 for SERP Always to 1.98 for SS2-NT @21 for SERP Average. Unsurprisingly, we observed that under SERP Perfect, CG was generally higher than the other two SERP level stopping decision point implementations. Corresponding interactive precision rose and fell with the CG attained – an intuitive result. As a reminder, results in Table 9.4 demonstrate the levels of CG and interactive precision attained using the configurations that yielded the lowest MSE values. These were computed with respect to click depths. As such, these values do not correspond to the maximum level of performance that could be attained; rather, they demonstrate the best performance that would have been attained by a searcher had a given result summary level stopping strategy been rigidly followed. From the **RW** values for interface T2 and ND AD, we see that the simulated searcher CG and interactive precision values fall below the real-world equivalents.

To conclude our analysis of the comparison runs, we performed a series of statistical tests to demonstrate if significant differences in terms of approximations existed. Like our *what-if* performance runs described in Section 9.3.2 above, we considered the best-performing SERP level stopping decision point implementation and result summary level stopping strategy combination, comparing the MSE values attained there against the other two SERP level decision point implementations. Given the closeness of each of the SERP level decision point implementations, no significant differences were found over any combination of result summary level stopping strategy, or interface/condition. Therefore, results show that there is supporting evidence for SERP-RQ2, albeit not statistically significant. Interesting findings between interface T2 and ND AD will receive further discussion in Section 10.2.

9.4 Chapter Summary

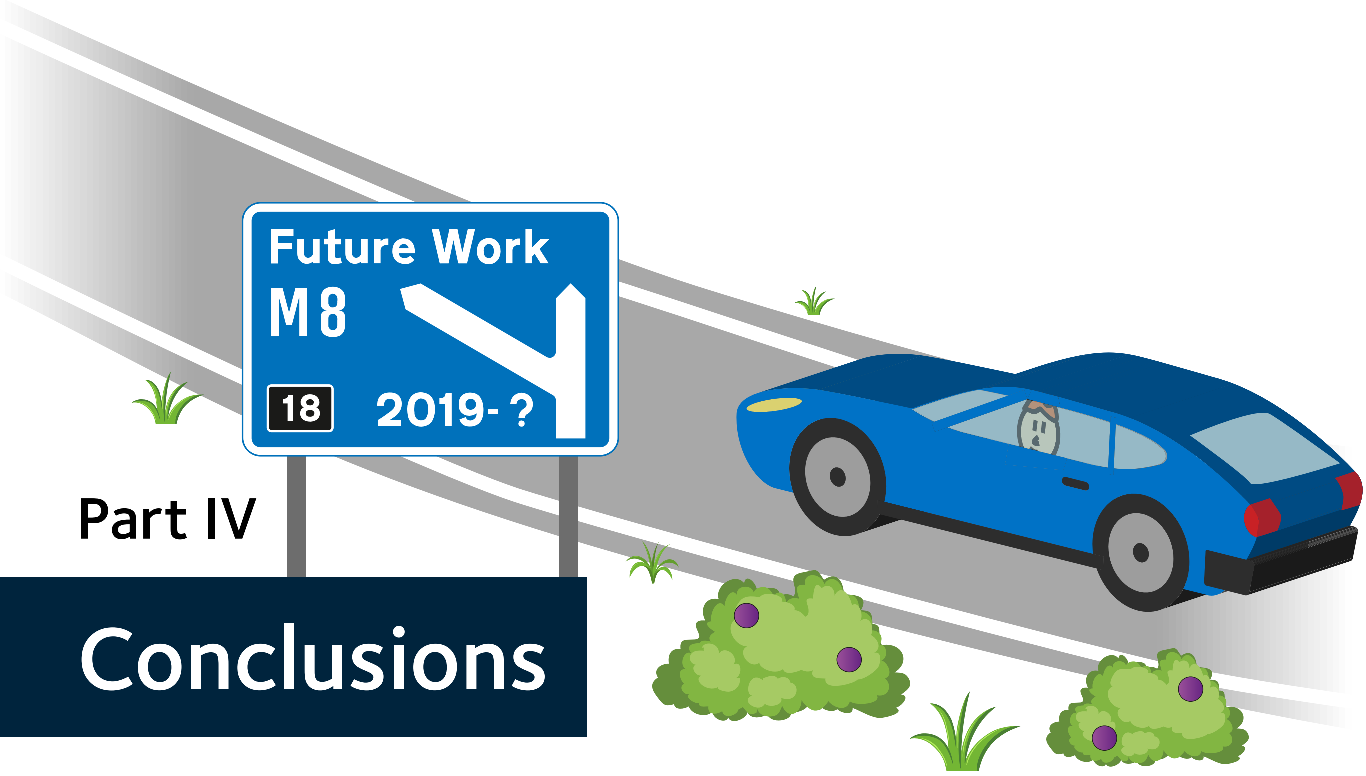
In this chapter, we conducted a series of simulations that empirically tested the CSM, complete with the new SERP level stopping decision point (as discussed in Section 4.3.1 on page 113). We observed improvements in:

- SERP-RQ1 overall performance; and
- SERP-RQ2 approximations to actual real-world searcher stopping behaviours.

Results were significantly improved in terms of overall CG attained between the upper-bound SERP Perfect implementation (only examine a SERP if it appears to yield a high scent), and the baseline SERP Always implementation (always examine a SERP, regardless of perceived quality). Improvements were consistent across our trialled interface and condition, as well as across different result summary level stopping strategies. Our stochastic implementation, SERP Always, consistently ranked between our baseline and upper-bound implementation.

Results pertaining to SERP-RQ2 were however not significant, with our findings differing across interface T2 and condition ND AD. Over interface T2, we found that the SERP Average implementation consistently yielded the better approximations over SERP Perfect and SERP Always – an intuitive result. Differences, as previously mentioned, were however very slight and not statistically significant – MSE approximations varied in the region of ten units. Over condition ND AD, we found that SERP Average and SERP Perfect offered the best approximations – an interesting result. This suggests that when task goals vary, a searcher's behaviour with respect obtaining an initial impression of a SERP also varies.

Overall, we find that including the new SERP level stopping decision point ultimately leads to better performing and more realistic simulations of interaction.



Part IV

Conclusions

This final part of the thesis summarises the findings from this research, provides a discussion of our results, and explores a number of potential areas for future work.

Chapter 10

Discussion and Future Work

The final chapter of this thesis summarises and discusses the results reported. In particular, we emphasise on the impact of our findings on IR and IIR research, discuss the limitations of our work, and outline several potential future research directions.

10.1 Thesis Summary

In this thesis, we examined how stopping behaviours vary under different search contexts. In particular, we conducted and reported on two user studies under the domain of news search, examining how ❶ result summary lengths and ❷ a variation of search tasks, goals and retrieval systems affected search behaviours. A total of eight different interfaces and conditions were used to examine how behaviours vary – as summarised in Table 10.1.

From the first user study reported in Chapter 7, results showed that as result summary lengths increased (from T0 → T4), searchers became more confident in the decisions they took pertaining to the relevance of documents encountered. However, this was not reflected empirically; their accuracy in identifying relevant content did not improve with

Table 10.1 A summary table of the different experimental interfaces and conditions that were trialled. These are based upon the work reported in Chapters 7 and 8. In total, eight different experimental interfaces and conditioned were employed, considering different result summary lengths, systems and tasks.

		Summary Length	System	Task
Chapter 7	T0	Title only	ND (Non Div.)	AD (Ad-hoc)
	T1	Title + 1 snippet	ND (Non Div.)	AD (Ad-hoc)
	T2	Title + 2 snippets	ND (Non Div.)	AD (Ad-hoc)
	T4	Title + 4 snippets	ND (Non Div.)	AD (Ad-hoc)
Chapter 8	D-AS	Title + 2 snippets	D (Div.)	AS (Aspectual)
	ND-AS	Title + 2 snippets	ND (Non Div.)	AS (Aspectual)
	D-AD	Title + 2 snippets	D (Div.)	AD (Ad-hoc)
	ND-AD	Title + 2 snippets	ND (Non Div.)	AD (Ad-hoc)

longer summaries. In terms of stopping behaviours, a downward trend was observed. As the length of summaries increased, subjects examined to shallower depths per query – an intuitive result, given the increased examination times required for longer summaries.

Considering variations of tasks, goals and systems as reported in Chapter 8, we found that when using diversified system D (i.e. BM25 and XQuAD (Santos et al., 2010)), subjects issued more queries, and stopped at comparatively shallower depths per query. This was in comparison to the non-diversified system ND (i.e. BM25 baseline), where subjects reported feeling less confident about their decisions. Despite the significant differences we observed regarding how the two systems performed, few significant differences were observed when examining changes in searcher behaviours. Most subjects reported difficulty in identifying differences in performance between the two systems.

Analysis of interaction data from these user studies was then used to ground an extensive set of simulations of interaction. These simulations were designed to test a total of twelve individual stopping strategies, derived from six different stopping heuristics[1] and the RBP IR measure. Our approach to cataloguing these heuristics – together with the subsequent operationalisation of them into stopping strategies – provided an answer to **HL-RQ2**. We then tested the overall performance and how closely the simulations matched up to real-world searcher behaviours (across the eight experimental interfaces and conditions). In turn, this allowed us to provide answers to both **HL-RQ3a** and **HL-RQ3b**. The simulations were modelled with the *Complex Searcher Model (CSM)*, a high-level, conceptual model of the search process. By incorporating a new SERP level stopping decision point into the CSM, complete with subsequent empirical evaluation (as presented in Chapter 9), we could then provide an answer to **HL-RQ1**.

Results show that when enabled, the new SERP stopping decision point led to significant improvements over the baseline implementation, with consistent improvements in overall performance (measured in CG) reported across a range of experimental conditions, interfaces and stopping strategies. Improvements in approximations of real-world searcher stopping behaviours were also achieved. However, statistical significance for these improvements was not obtained. Overall, these results provide compelling evidence to address **HL-RQ1**. The results also demonstrate a promising direction for future research in developing our understanding of the search process.

With respect to our simulated analyses of individual stopping strategies, we found several stopping strategies offered high levels of mean CG, and good approximations toward actual searcher stopping behaviours. For example, we found that with increased result summary length, **SS11-COMB** consistently offered the best performance. **SS1-FIX** and **SS4-SAT** offered the best real-world searcher approximations. Furthermore, **SS5-COMB** offered the best level of CG within the second user study, while **SS1-FIX** offered the best level

[1] Stopping heuristics for example considered a searcher's tolerance to non-relevance, or their *frustration* with observing non-relevant content (Kraft and Lee, 1979).

of performance across condition ND AD. However, SS1-FIX and SS10-RELTIME yielded the lowest MSE values. Despite several strategies performing well, no single strategy clearly emerged as offering significantly improved levels of performance or approximations when acting alone. On the contrary, several more complex stopping strategies offered poorer performance, such as SS6-DT and SS7-DKL. This was a common theme in our results: simple and combination-based stopping strategies generally provided the highest levels of performance. This includes the fixed-depth stopping strategy, SS1-FIX, which, counter to our intuition, consistently performed well.

10.2 Discussion

From the analysis of our simulations of interaction, a number of novel, interesting areas of discussion were revealed. In this section, we discuss our findings with an emphasis on examining the result summary level stopping strategies. In particular, our discussion is guided by our four high-level research questions. We repeat these below.

- HL-RQ1 How can we improve searcher models to incorporate different stopping decision points?
- HL-RQ2 Given the stopping heuristics defined in the literature, how can we encode these heuristics into a series of operationalised, programmable stopping strategies that can be subsequently incorporated into the searcher model and be evaluated?
- HL-RQ3a Given the aforementioned operationalised stopping strategies, how well does each one perform?
- HL-RQ3b How closely do the operationalised stopping strategies compare to the actual stopping behaviours of real-world searchers?

With these research questions pertaining to the simulations of interaction (along with the implemented stopping strategies), in-depth discussion of our user studies can be found in Sections 7.2.3 on page 214 and 8.2.3 on page 278. However, we do briefly touch on summarising statements relating to searcher behaviours in Section 10.2.3.

10.2.1 Searcher Models and Realism

Work in this thesis has reported advancements to modelling the IIR process – particularly with the inclusion of the new SERP level stopping decision point. The inclusion of the new stopping decision point led to significant improvements in terms of the level of CG that could be attained, together with improved approximations of real-world behaviours.

However, these significant improvements from the SERP Always baseline (as reported in Chapter 9) were only achieved with the SERP Perfect implementation. This is a limitation, as the implementation relied upon access to the TREC QRELs in order for the impression to be determined – although this implementation acted as a good upper bound. While improvements in performance and approximations were noted with the SERP Average implementation, these changes did not achieve a significant level of improved performance. We discuss this limitation of our simulations later in Section 10.2.4.

Of course, attaining access to the *gold standard* is wholly unrealistic. However, the present study demonstrates the *maximum performance* that can be reached with the inclusion of this new stopping decision point. The observed improvements demonstrate that more realistic simulations of interaction may be produced. With further work examining the proximal cues that searchers observe when forming an initial impression of the SERP, incorporating these findings into future models and simulations of the search process would arguably make them even more realistic.

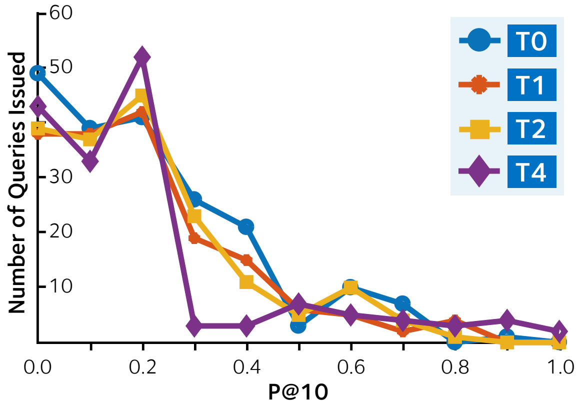
10.2.2 Stopping Strategy Operationalisation

In general, findings across all interfaces and conditions demonstrated that simple stopping strategies tended to yield better performance, and matched better with real-world searcher stopping behaviours. Stopping strategies SS2-NT, SS3-NC, SS4-SAT, SS9-TIME and SS10-RELTIME for example performed and approximated well. We consider these to be simple in the sense that the stopping criterion that they each encoded was straightforward to implement and subsequently measure. Examples included the consideration of aspects such as the number of non-relevant documents encountered, or the elapsed time spent searching since a query was issued.

In contrast, findings also demonstrated that the more complex stopping strategies tended to perform worse. They consistently offered poorer performance and approximations. Complexity was again denoted by the criterion/criteria that were considered by each of the stopping strategies, with more complex computations required in order to determine when the simulated searcher should stop. Given these general findings, *why did the more complex stopping strategies perform and approximate worse on average?* The present section of the discussion focuses primarily on this question, considering the difference-based strategies SS6-DT and SS7-DKL, the IFT-based strategy SS8-IFT, and the RBP-based strategy SS12-RBP. We also discuss the importance of more performant stopping strategies, such as SS5-COMB and SS11-COMB.

Difference Stopping Strategies Considering SS6-DT and SS7-DKL, we hypothesise that the performance of issued queries may be having an effect on the way in which these strategies perform. In other words, the stopping strategies *may not be very robust* to varying levels of query performance. Recall that for our *what-if* performance runs, we employed an interleaved querying strategy QS13, where single and three term queries were interleaved.[2] From empirical evidence, it was shown that single term queries offered poor

[2] Refer to Section 6.4.2.2 on page 164 for additional information on how the querying strategy was implemented.


Query Performance over Interfaces

Query Performance vs. Lengths (All Interfaces)

#Terms	#Queries	Mean P@10
1	12	0.16 ± 0.13
2	162	0.23 ± 0.18
3	287	0.24 ± 0.21

On average, as query length increases, P@10 scores also increase. Values reported represent the mean performance ± standard deviations.

Figure 10.1 Plot demonstrating the performance of queries (**P@10**) across the four experimental interfaces trialled in the user study we report on in Chapter 7. On the right, a table highlights the varying levels of performance (averaged over all four experimental interfaces) in relation to query term lengths. As query term length increases, so too does the mean P@10 score. Similar findings were observed for the study reported in Chapter 8.

performance (in terms of *P@k*) when compared to three term queries – single term queries offered higher levels of query ambiguity compared to the three term queries.[3] As such, using a fixed threshold across queries of varying performance would not necessarily make sense. A low threshold for SS6-DT and SS7-DKL would mean that searchers would stop too early for single term queries, and examine to excessive depths for three term queries. A higher threshold would mean that searchers would examine to excessive depths generally. This means for example that a low threshold would be too stringent for single term queries, and suggests that poor levels of gain would be achieved.

As such, we hypothesise that for stopping strategies based upon the difference-based heuristic, thresholds should likely be query specific – perhaps dependent upon the length of the query issued. Given queries issued by the real-world subjects in Chapter 7, we also observed a large variation in performance for the queries that were issued. We report this in Figure 10.1, with a plot showing the number of queries issued across each of the four in-

[3] For example, consider the queries `piracy' and `piracy china sea'. The first single term query returned a majority of documents pertaining to software piracy, along with piracy at sea. In contrast, the three term query returned a majority of its matched documents to instances of piracy on the South China Sea, relevant to the TREC piracy topic.

terfaces, plotted against the performance of the queries. A table also provides evidence to support our hypothesis, showing that as the number of terms in the queries increased, so too did the mean level of query performance. Similar findings were observed in the user study reported in Chapter 8.

IFT Stopping Strategy Next, we consider the poor performance and approximations afforded by SS8-IFT. Evidence has shown that IFT has been proven to be good at predicting search behaviours (Ong et al., 2017; Azzopardi et al., 2018). In Section 8.2.2.5 on page 274, we demonstrated that our IFT-based hypotheses matched closely to empirical evidence. So, why did SS8-IFT consistently offer poorer performance and approximations when compared to more simplistic stopping strategies? We hypothesise that this comparative lack of performance can be attributed to how the *rate of gain* was operationalised, which serves as the stopping criterion for SS8-IFT. This is an inherently difficult value to compute, with limitations relating to the rate of gain considered from two angles:

❶ the *per-topic* rate of gain; and

❷ *how the rate of gain is estimated* by searchers in the first instance.

Considering point ❶ first, we note that the same gain stopping threshold values (for x_8) were trialled over all five topics in the reported simulations of interaction. Table 6.1 on page 140 demonstrated that the number of TREC relevant documents for each of the five topics varies considerably. As such, one would expect that the computed rate of gain would also vary considerably on a per-topic basis. This way, expectations of gain can be kept in check – a rate of gain threshold computed over a performant TREC topic with many relevant documents would perform much worse under a topic for which it is much harder to find relevant documents for (i.e. a comparatively smaller number of TREC relevant documents). This variation in the number of relevant documents over topics (amongst other factors, such as the retrieval system used) is illustrated in Figure 10.2. Using interface T2 (left) and condition ND AD (right) over SS3-NC, the two plots illustrate how performance varies

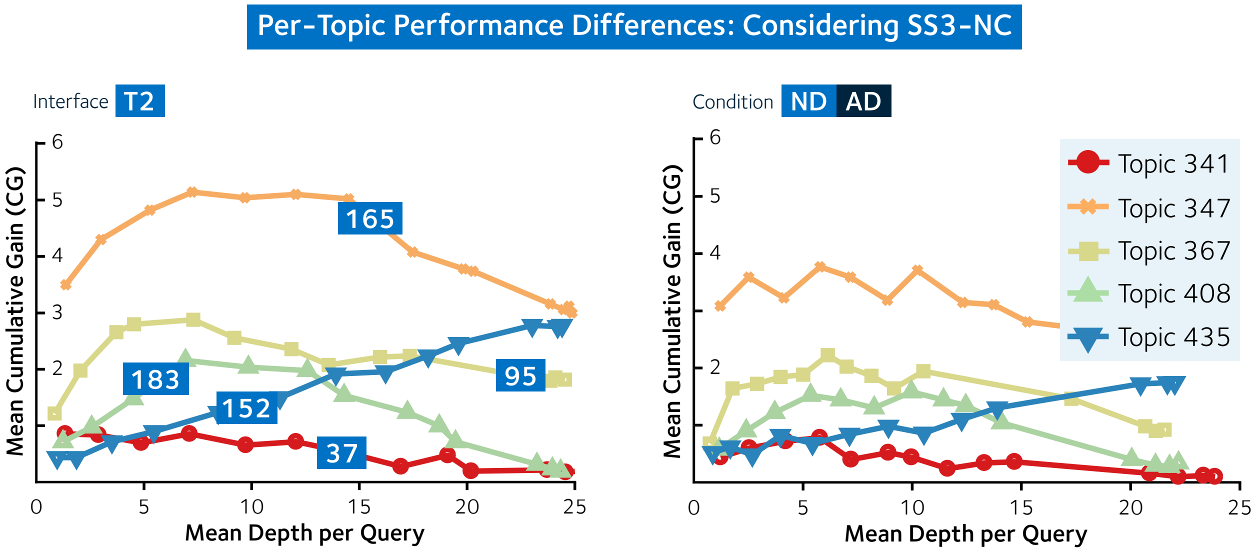


Figure 10.2 Plots demonstrating the wide per-topic variance over the *what-if* performance simulations. On the left, performance over interface T2 is shown — ND AD is shown on the right. Stopping strategy SS3-NC is used for this demonstration. Similar observations were observed across other interfaces, conditions and stopping strategies. Also highlighted on the left plot is the number of TREC relevant documents for each topic. Note the general performance improvement as the number of TREC relevant documents increases for a topic.

across the five topics. We also note a general trend of higher performance for a topic in the plots if a greater number of TREC relevant documents are present.

We also consider how the rate of gain is computed, as per point ❷. *How do searchers estimate a rate of gain threshold?* This is a difficult question to answer, with further study required to address this. However, one would be pressed to believe that from an initial impression of a SERP, a searcher would undertake a series of computations in their head to reach an estimation for a rate of gain threshold value. It is much easier to believe that searchers would rather employ a simpler stopping criterion in this instance, such as *stopping after observing k non-relevant result summaries* (i.e. the frustration-based heuristic). This can be simplified with the trivial example of an individual throwing a ball in the air, as illustrated in Figure 10.3. It would be easier to believe that the ball thrower would think of how to catch the ball in relation to how it is falling through the air, with feedback from their visual

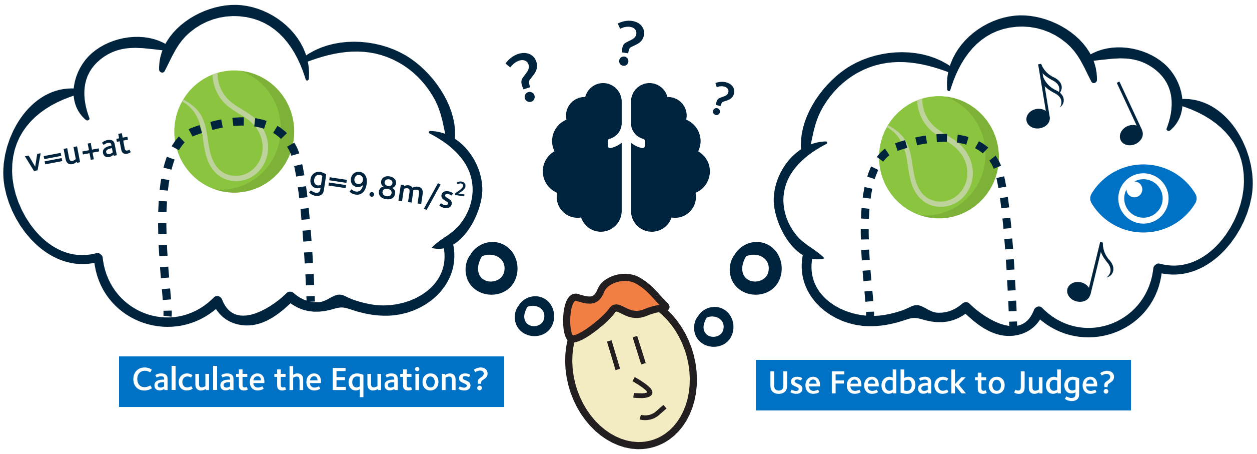


Figure 10.3 *How would you catch a ball in the air?* Would you consider all of the equations required to work out when and where in space the ball will be for you to catch it, or would you rely on your visual/proprioception systems to guide you? If you are human, it'll be the latter.

and proprioception systems. This is opposed to believing that the thrower of the ball may catch it by calculating the equations relating to the physics of the falling ball to predict the optimal point in space at which to intercept it.

However, even if we were able to provide better values for the rate of gain, would we see improvements in real-world approximations? While IFT says that we will, individuals may be behaving in a suboptimal way. A body of literature in ecology suggests that when foraging for food in the wild, animals *do* behave in a suboptimal way. Janetos and Cole (1981) and Krebs et al. (1983) state that animals may employ some *rule of thumb* that is less than perfect, with an example cited as *'take the largest thing you can eat'*. This is some ways analogous to the more simplistic stopping strategies we trialled. Krebs et al. (1983) also argue that these simplistic approaches are actually an optimisation within a wide number of constraints, such as sensory limitations. This may be true of searchers, too – with limited working memory, a more simplistic approach may, in reality, be the better, optimal choice even though the theory may suggest otherwise.

RBP Stopping Strategy We also observed that SS12-RBP, the RBP-based stopping strategy, also generally failed to provide a good approximation. Performance in the *what-if* simulations was generally significantly different from the best performing stopping strategies, although instances such as T2 did not demonstrate any significant difference. While performance matchups might have been relatively good, the depths to which simulated searchers examined content using this stopping strategy were weak to a considerable degree. Refer to plot SS12-RBP in Figure 7.8 on page 232 for an example. Recall the patience parameter p of RBP, that dictated how deep down a list of ranked results a searcher would be prepared to go. The point at which the searcher would decide to stop was modelled stochastically. In the real-world, searchers do not roll a dice to determine when to stop, but rather rely upon some form of an intuitive informational cue, as have been previously shown to affect search behaviours. However, it may also be the case that this way of representing human behaviour is also correct at times – humans can often behave irrationally.

Considering more Performant Stopping Strategies Both the combination-based stopping strategies SS5-COMB and SS11-COMB performed and approximated searcher stopping behaviours well. Formed of more simplistic stopping strategies (e.g. SS2-NT), results seem to suggest that searchers do not consider a single criterion when determining the point at which they should stop examining results – an interesting finding.

This interesting conclusion can be corroborated by other recent studies. Work by Zhang et al. (2017a) used the *Bejeweled Player Model (BPM)* to model a searcher's stopping behaviours, where they would stop when:

> *"he/she either has found sufficient useful information, or no more patience to continue."*
>
> Zhang et al. (2017a)

Findings from this study demonstrated improvements in the correlations between searcher satisfaction and existing IR evaluation measures. This was also corroborated in a recent study by Azzopardi et al. (2018). Central to this argument is the similarity of the BPM to SS5-COMB, that considered a combination of the satiation (SS4-SAT) and frustration (SS2-NT or SS3-NC) stopping strategies. This provides evidence that empirically validates the inclusion of both satiation and frustration-based stopping heuristics within the searcher model. The evidence is clearly showing that multiple criteria are being considered when a stopping point is decided, and future work should consider the development of measures that support both criteria.

The Fixed Depth Fallacy Overall, a majority of stopping strategies performed well and produced approximations that were very close to one another, with few significant differences. One particularly surprising result was that of SS1-FIX. The fixed-depth, non-adaptive baseline approach consistently offered good performance and approximations. This is counter-intuitive, as it would make sense for more adaptive strategies to offer improved approximations. It is likely that different subjects would have employed different stopping strategies, or a variety of different strategies depending upon the situation (i.e. as demonstrated by SS11-COMB). In this regard, next steps should consider stopping behaviours on an individual level. However, from the perspective of averaging over a population, many of the stopping strategies trialled, and when tuned appropriately (i.e. would SS1-FIX @24 really be considered as realistic? It is unlikely!), offer good approximations and performance. This provides a rationale as to how the fixed depth strategies consistently performed and approximated so well across our results.

10.2.3 Searcher Behaviours

From the reported user studies, it is clear that the interfaces and conditions that we trialled do affect the behaviours of searchers. In terms of stopping behaviours, we did observe differences, but differences often were not significant. We hypothesise that due to the high levels of variance that we observed, larger sample sizes over each study would be required in order to tease out significant differences and to provide data for further examination.

Understanding stopping behaviours is difficult. What findings from our studies do suggest is that variations in interfaces, tasks, goals and systems do impact upon performance. For example, as we increased result summary lengths, stopping depths became shallower (i.e. from T0 → T4). More extreme interfaces and conditions would likely amplify the effect. Factors such as how the prior topic knowledge that a subject possesses were also not considered, and would likely play a role in stopping behaviours.

10.2.4 Simulations of Interaction

In this thesis, we have presented significant advancements in terms of modelling and understanding the IIR process. We developed an extensive framework that allowed us to change components of the underlying CSM. Given this framework, we could then formulate the search problem more precisely, and explore the impact that each of the component variations had on the wider search process. As components were changed, we were able to demonstrate improvements in the approximations of human searcher behaviours. Given the limitations of our user studies with the risk of an insufficient amount of interaction data, simulations of interaction allowed us to generate more data at a much lower cost.

One particularly novel contribution concerning the simulations of interaction was addressing the issue of comparing results across different configurations. Being stochastic in nature, the simulations relied upon the roll of a dice to determine whether a simulated searcher would click on a result summary link (if deemed sufficiently attractive to warrant further examination), or save a document as relevant (if deemed relevant to the given information need). These were grounded on the TREC relevance judgements and interaction probabilities extracted from the user studies. Across different configurations however, outcomes of the dice roll would have resulted in different decisions being taken – which in turn ensured that when examining two configurations, their outcomes would not be comparable.

Section 6.4.2.3 on page 167 outlined a *pre-rolled judgements* technique that rolled the dice *a priori* 50 times, with 50 being the number of trials that were run per configuration. This then meant that during the simulations, the decision maker components of the **SimIIR** framework essentially became deterministic, extracting the judgement for a particular trial from a pre-rolled *action judgement file.* In turn, this addressed the issue of comparability between different configurations. With the same judgements, comparisons became fairer. Of course, a larger number of trials would always be more desirable as a means of teasing out further differences that perhaps would otherwise not have been observed.

We also note limitations of the approach taken in conducting our simulations of interaction. Most notably, we considered the *most optimistic outcome* at several points in our simulations, mainly pertaining to the perceived quality of a SERP. A prime example of this was highlighted in Section 10.2.1, with the SERP Perfect SERP level stopping decision point implementation highlighted as the implementation yielding significant improvements in performance, yet attaining this with access to TREC QREL judgements.

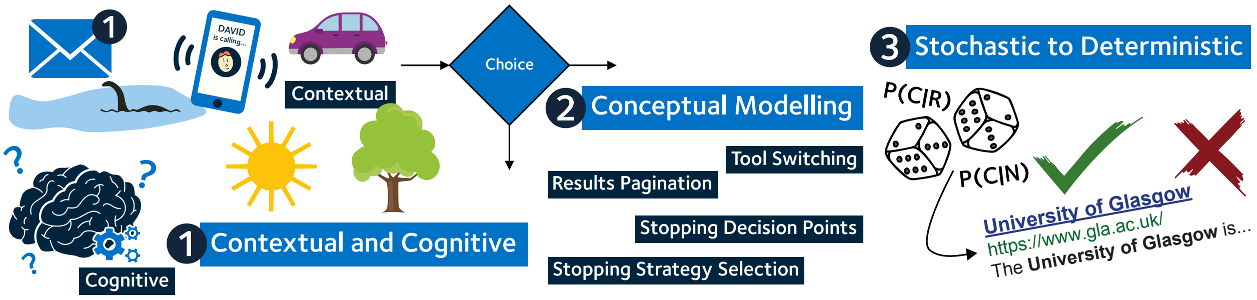
A further example of this approach was demonstrated with combination stopping strategy SS11-COMB . Similar to the SERP level stopping decision point, this strategy took an initial impression of a presented SERP, and used this impression to select an appropriate constituent stopping strategy – with either SS4-SAT for a SERP yielding relevant documents early in the rankings, or SS10-RELTIME for a SERP of dubious quality. Under these conditions, such strategies do intuitively make sense. However, the decision was again made with access to TREC QRELs – *P*@1 was used to determine if the SERP yielded relevant content at shallow ranks. If, for example, a stochastic approach were to be implemented in determining what stopping strategy to employ, it may mean that even better approximations of real-world searcher behaviours could be achieved.

10.3 Future Research Directions

From the summary and discussion of our empirical results, a number of potential avenues for future work may be considered. In this section, we consider: how to improve the realism of simulations of interaction further; stopping heuristics and strategies; simulation trials and topics; and the modelling of stopping from the level of individual searchers.

10.3.1 Improving Simulation Realism

In this thesis, we presented the CSM, a high-level, conceptual searcher model. It encapsulates many of the different activities and decision points that searchers would contend with across informational search tasks. With the inclusion of the new SERP level stopping decision point, improvements were made to the realism of the simulations that were executed with the CSM. However, *what changes could we subsequently make to the CSM and related infrastructure in future work that would aid in advancing the realism of these simulations further?* As illustrated below, we consider this open question from three main research strands.



Contextual and Cognitive Our first strand considers **contextual** and **cognitive** factors. All experimentation in this thesis was conducted under the domain of news search, with subjects of the user studies asked to imagine that they were newspaper reporters, having being given a task to find documents that they thought were relevant to a particular topic. However, this scenario is very specific. If we performed studies with the same methodology, but under a different search context, would we find similar results? Arguably, behaviours will change – general web search and a detailed examination of content under the context we employed will result in different outcomes, for example. Different tasks can also be considered. Aspectual and ad-hoc tasks were considered as we believed they would offer the greatest difference in terms of stopping behaviours. Would other retrieval tasks offer even bigger differences in terms of searcher behaviours?

Other factors such as the location at which the search is undertaken, the device upon which the search is undertaken and other external pressures will also undoubtedly influence the outcome of the results obtained. Crowdsourced subjects whose behaviours are reported in this thesis conducted our experiments on a desktop or laptop computer. They were instructed to be in a comfortable, quiet location, free from major distractions. In reality, individuals are less likely to search in such conditions. Perhaps time pressures would influence their behaviours – a student under pressure to finish a draft of her paper will behave differently to one who is not under the same pressure. With the proliferation of mobile devices such as smartphones, searching on such devices must also be considered. A recent study by Ong et al. (2017) demonstrated that search behaviours, for example, do differ between individuals using desktop computers and smartphones.

Much work remains to determine how we can try to understand and subsequently model the cognitive processes and factors that influence how individuals behave when searching. Individuals are products of their prior experiences, and are therefore unique; behaviours will undoubtedly differ from person to person. Within the modelling process, novel techniques can be applied that could possibly improve the realism of simulations. For example, within the **SimIIR** framework, the search context component tracks a list of queries issued, documents examined (and saved), along with other measures. Could this component be manipulated in such a way as to better mimic the behaviours of a human? Rather than maintaining a perfect list of everything that has been examined, a simulated searcher could be programmed to become 'forgetful' in remembering what they have examined, with cues within a document reminding them that they previously examined it. Other factors, such as prior topic knowledge (as alluded to in Section 10.2.3) ought to be considered, as such aspects would likely impact upon the stopping behaviours of searchers.

As alluded to in Chapter 8, further work could also be undertaken in relation to the decision making components of the **SimIIR** framework. This work would consider how simulated searchers would judge the attractiveness of result summaries and relevance of documents to a given topic. Decision makers were implemented primarily with ad-hoc retrieval in mind, considering only the probability of clicking or saving with respect to the TREC *Query Relevance Judgement (QREL)* judgement. For aspectual retrieval tasks, further work would consider whether the result summary or document contains a discussion of new entities for the topic, such as a previously unseen species of animal.

Conceptual Modelling We next consider a number of further enhancements to the CSM that could improve the realism of simulations further. Examples in the illustration above consider potential areas for future improvement. One such example, **tool switching**, (demonstrated by Thomas et al. (2014)) would be considered at the beginning of the search process. It would enable a searcher to determine what tool (or retrieval system) would be better suited to help them satisfy their information need. This is opposed to the current CSM as presented in this thesis that assumes a retrieval system has been selected *a priori*. A study by White and Dumais (2009) has shown that predicting tool switching is feasible. They reported that sufficiently consistent behaviours exhibited by searchers in relation to this phenomenon led to accurate predictions of tool switching events.

Results pagination is also listed in the illustration above. Here, a simulated searcher will be presented with SERPs that are split across a number of different pages, rather than examining a continuous ranked list of results. This would involve the notion of extracting additional grounding data from interaction logs, perhaps such as the likelihood of a searcher continuing to the next SERP page. This would likely impact upon the realism of simulations, as a study by Jansen and Spink (2005) showed a sharp decrease in content examined after the first page of results. Further examination of modelling stopping behaviours within the CSM is also considered; refer to Section 10.3.2 for further details.

Stochastic to Deterministic Decisions pertaining to the attractiveness of result summaries and the relevance of documents within our simulations of interaction were determined *stochastically*, or by a roll of the dice. While a simplifying assumption that has been used in many other studies employing simulations of interaction, this is an unrealistic approach. If implemented correctly, a more *deterministic* solution would offer more realistic simulations, where simulated searchers would be able to *learn* as they traverse through content, improving their decision making abilities based upon the content observed, rather than the outcome of a roll of a dice. Advancements in understanding the *information triage* process would undoubtedly lead to improved realism. In addition, the inclusion of *variable interaction costs* would also benefit the realism of future simulations.[4]

10.3.2 Stopping Heuristics and Strategies

In this thesis, we considered a total of twelve different stopping strategies, operationalised from a total of seven different stopping heuristics. We showed how each of the different strategies perform over a number of different experimental interfaces and conditions. During the methodological design stage, it became apparent that approaches taken for the operationalisation of our stopping strategies were just one of many. *What if we implemented our stopping strategies in different ways? Why did we select these strategies?* Here, we consider these questions with insight into what might happen if they were to be addressed.

Stopping Decision Points Following on with the theme of improving the underlying CSM, additional stopping decision points could be included. These would provide searchers subscribing to the CSM with greater flexibility regarding when they stop examining content. Additional stopping decision points could, for example, include one for tool switching. In this example, as we discussed earlier, a searcher, after spending some period of time on one retrieval system, could decide to stop using it after certain criteria are met. After this point has been reached, they will then switch to a different retrieval system. A further interesting research question would be whether the result summary level stopping strategies trialled in this thesis would work at different stopping decision points. For example, at a session level, would these strategies make sense? Would using them at that decision point lead to a better matchup with real-world stopping behaviours?

Stopping Strategy Selection From here, we can also consider a further decision point that could be encoded within the CSM. Inspired by **SS11-COMB**, consideration must be

[4] As discussed previously in this thesis, *time-biased gain* (Smucker and Clarke, 2012) is an example of such an approach.

taken into deciding *why* and *when* a particular stopping strategy could be employed. As we demonstrated in Figure 5.3 on page 132, SS11-COMB employs both the frustration and give-up time-based stopping heuristics – but not at the same time. Rather, a decision is made pertaining to the quality of the presented SERP (much like the SERP level stopping decision point). The outcome of this decision then dictates what stopping strategy is employed for the remainder of the query. Further refinements to this approach could, for example, include additional stopping strategies and a wider range of conditions for employing them. Empirical evidence could be extracted from interaction logs to determine if, under certain circumstances, searchers would favour one approach over another.

Stopping Strategy Operationalisation An open question arising from the work in this thesis considers: *how do you operationalise the stopping heuristics?* Clearly, from the outline of the twelve stopping strategies in Chapter 5 on page 121 (and implementation methodology in Section 6.4.2.6 on page 173), there are a large number of different ways in which the stopping strategies can be implemented. While we provided a means and justification for the approaches that we took in this thesis, we have reason to believe that some of the stopping strategies – especially SS6-DT, SS7-DKL and SS8-IFT – performed poorly, perhaps because of our implementations (refer to Section 10.2.2). For example, the rate of gain for SS8-IFT could have been computed on a per topic basis. Further work will be required in order to determine if different implementations would lead to performance improvements.

Considering Additional Stopping Heuristics Of course, the seven stopping heuristics that we considered do not constitute the entirety of the heuristics defined in the literature. We selected these heuristics as they offered interesting differences between one another, were *relatively* straightforward to implement, and would likely be discernible across complex informational search tasks (though this was not necessarily proved). Unused heuristics such as the mental list heuristic (considering different criteria that must be met, as outlined by Nickles (1995) and detailed in Section 3.2.2.3 on page 87) would have been much more challenging to operationalise and implement – and even so, would such a heuristic be suitable for the task at hand? The ability for a searcher to create a series of bullet points about a topic would imply he or she has some sound idea of their objective. The searcher's knowledge of a topic may be so limited that such a heuristic would be unsuitable. Linking back to contextual factors above, considering additional search contexts (perhaps with searchers of astute and limited knowledge of a topic) would be interesting to examine.

Towards Future IR Measures Given the above, findings from this research provide motivation for further work considering the inclusion of stopping heuristics within the measures that are used within IR research. For example, stopping strategy SS5-COMB demonstrated good overall and performance considering a searcher's satisfaction and tolerance toward non-relevant material. This has also been shown in the BPM (Zhang et al., 2017a).

10.3.3 Simulation Trials and Topics

We also consider future work in terms of *how* the simulations of interaction could be run. While 50 trials were selected because of the fact that approximately 50 subjects partook in each user study, there are likely trends and significant differences that exist that we simply did not observe because of a lack of experimental power. This limitation was also imposed with an insufficient amount of processing power to complete the experiments in a reasonable timeframe.[5] With more powerful computer hardware, scaling up the experiments with more trials would have become a more realistic prospect.

We also consider using five topics for our performance *(what-if)* experiments to be a limiting factor. While the decision to use five topics was justified due to a lack of data (considering entities across the remaining 45 topics in Chapter 8) – and to ensure that comparisons between interfaces and conditions were fair – 50 topics would have been preferred (refer to Figure 10.2). If (aspectual) data were available for the remaining 45 topics, we could then trial additional performance runs, which may also lead to the observation of other trends and potential significant differences.

[5] Using the experimental setup detailed in this thesis, all simulations of interaction took approximately 38 days of processing time.

10.3.4 Individual Searcher Stopping Behaviours

Our final consideration for future work revolves around the notion of *individual searcher stopping behaviours*. In this thesis, we considered searcher stopping behaviours, reported across ≈ 50 subjects, over each interface and condition that was trialled. This provided us with a rough approximation as to what strategies work best, with similar findings reported across interfaces and conditions. However, research has shown that individual searcher behaviours may differ to a significant degree. If we considered individual searchers, what trends would we then observe? We may see a decrease in how well the fixed depth stopping strategy SS1-FIX fares, given that we hypothesised in Section 5.1 on page 123 that such an approach would work well on average. If we examined behaviours on a per-searcher basis (or even at a session level), how would the strategy then fare?

Examining behaviours on a per-searcher level will avoid watering down results through averaging over a particular cohort, exposing more interesting results. For example, could we perform a classification of searcher stopping behaviours? Such an approach was followed, for example, by Smucker (2011), who devised a classification of searchers when examining documents – with searchers being categorised into one of either *fast and liberal* or *slow and neutral*. This is undoubtedly one key area of future work that we must consider in order to develop a deeper understanding of the stopping behaviours that searchers employ.

10.4 Final Remarks

Stopping during the search process is a difficult phenomenon to understand and model effectively. A wide range of different factors influence the internal decision-making process of searchers. We have shown in this thesis that a number of simple stopping strategies can offer improved performance and approximations of real-world searcher behaviours. We also provide novel evidence to motivate the fact that multiple stopping criteria need to be considered in the development of future IR evaluation measures, along with the inclusion of additional stopping decision points to improve the realism of future searcher models. The development of the CSM has also been positive, with a solid baseline provided for future work in developing ever more realistic simulations of interaction.

Despite the inherently difficult task that understanding and modelling stopping behaviours represent, we believe that the potential benefits of further exploration in this area will undoubtedly aid the searchers and researchers of future retrieval systems.

Bibliography

Ageev, M., Lagun, D., and Agichtein, E. (2013). Improving search result summaries by using searcher behavior data. In *Proceedings of the 35th ACM SIGIR*, pages 13–22.

Agosto, D. E. (2002). Bounded rationality and satisficing in young people's web-based decision making. *Journal of the Association for Information Science and Technology*, 53(1):16–27.

Ali, H., Scholer, F., Thom, J. A., and Wu, M. (2009). User interaction with novel web search interfaces. In *Proceedings of the 21st OZCHI*, pages 301–304.

Ashkan, A., Clarke, C. L., Agichtein, E., and Guo, Q. (2009). Classifying and characterizing query intent. In *Proceedings of the 31st ECIR*, pages 578–586.

Athukorala, K., Oulasvirta, A., Glowacka, D., Vreeken, J., and Jacucci, G. (2014). Narrow or broad?: Estimating subjective specificity in exploratory search. In *Proceedings of the 23rd ACM CIKM*, pages 819–828.

Azzopardi, L. (2009). Query side evaluation: An empirical analysis of effectiveness and effort. In *Proceedings of the 32nd ACM SIGIR*, pages 556–563.

Azzopardi, L. (2011). The economics in interactive information retrieval. In *Proceedings of the 34th ACM SIGIR*, pages 15–24.

Azzopardi, L., de Rijke, M., and Balog, K. (2007). Building simulated queries for known-item topics: An analysis using six European anguages. In *Proceedings of the 30th ACM SIGIR*, pages 455–462.

Azzopardi, L., Järvelin, K., Kamps, J., and Smucker, M. D. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44(2):35–47.

Azzopardi, L., Kelly, D., and Brennan, K. (2013). How query cost affects search behavior. In *Proceedings of the 36th ACM SIGIR*, pages 23–32.

Azzopardi, L., Thomas, P., and Craswell, N. (2018). Measuring the utility of search engine result pages: An information foraging based measure. In *Proceedings of the 41st ACM SIGIR*, pages 605–614.

Azzopardi, L. and Vinay, V. (2008). Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM CIKM*, pages 561–570.

Azzopardi, L. and Zuccon, G. (2015). An analysis of theories of search and search behavior. In *Proceedings of the 1st ACM ICTIR*, pages 81–90.

Baillie, M., Azzopardi, L., and Crestani, F. (2006). Adaptive query-based sampling of distributed collections. In *Proceedings of the 13th SPIRE*, pages 316–328.

Banks, J., Carson, J., and Nelson, B. (1996). *Discrete-event System Simulation*. Prentice-Hall international series in industrial and systems engineering. Prentice Hall.

Baron, J., Beattie, J., and Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, 42(1):88–110.

Bartlett, F. C. and Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, 3(2):187–192.

Baskaya, F., Keskustalo, H., and Järvelin, K. (2012). Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th ACM SIGIR*, pages 105–114.

Baskaya, F., Keskustalo, H., and Järvelin, K. (2013). Modeling behavioral factors in interactive information retrieval. In *Proceedings of the 22nd ACM CIKM*, pages 2297–2302.

Bast, H. and Celikik, M. (2014). Efficient index-based snippet generation. *ACM Transactions on Information Systems*, 32(2):6:1–6:24.

Bates, M. (1984). The fallacy of the perfect thirty-item online search. *RQ*, 24(1):pp. 43–50.

Bates, M. J. (1989a). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424.

Bates, M. J. (1989b). Training and education for online. chapter Information Search Tactics, pages 96–105. Taylor Graham Publishing.

Belkin, N. (1980). Anomalous states of knowledge as a basis for information retrieval. In *Canadian Journal of Information Science*, volume 5, pages 133–143.

Benov, D. M. (2016). The manhattan project, the first electronic computer and the monte carlo method. *Monte Carlo Methods and Applications*, 22(1):73–79.

Benson, P. G., Curley, S. P., and Smith, G. F. (1995). Belief assessment: An underdeveloped phase of probability elicitation. *Management Science*, 41(10):1639–1653.

Berners-Lee, T., Dimitroyannis, D., Mallinckrodt, A. J., McKay, S., et al. (1994). World wide web. *Computers in Physics*, 8(3):298–299.

Berryman, J. (2006). What defines "enough" information? how policy workers make judgements and decisions during information seeking: Preliminary results from an exploratory study. *Information Research: An International Electronic Journal*, 11(4):4.

Boole, G. (1847). *The Mathematical Analysis of Logic*. Philosophical Library.

Borlund, P. (2000). Evaluation of interactive information retrieval systems. Unpublished doctoral dissertation, Åbo Akademi University.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(5).

Borlund, P. and Schneider, J. W. (2010). Reconsideration of the simulated work task situation: A context instrument for evaluation of information retrieval interaction. In *Proceedings of the 3rd IIiX*, pages 155–164.

Browne, G. J. and Pitts, M. G. (2004). Stopping rule use during information search in design problems. *Organizational Behavior and Human Decision Processes*, 95(2):208 – 224.

Browne, G. J., Pitts, M. G., and Wetherbe, J. C. (2005). Stopping rule use during web-based search. In *Proceedings of the 38th HICSS*, pages 271b–271b.

Brutlag, J. (2009). Speed matters for Google Web Search. `http://goo.gl/t7qGN8` (retrieved on March 14th, 2018).

Busemeyer, J. R. (1982). Choice behavior in a sequential decision-making task. *Organizational Behavior and Human Performance*, 29(2):175 – 207.

Busemeyer, J. R. and Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, 32(2):91–134.

Callan, J., Allan, J., Clarke, C. L. A., Dumais, S., Evans, D. A., Sanderson, M., and Zhai, C. (2007). Meeting of the minds: An information retrieval research agenda. *SIGIR Forum*, 41(2):25–34.

Card, S., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R., Schraedley, P., and Boshart, J. (2001). Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability. In *Proceedings of the 19th ACM CHI*, pages 498–505.

Carr, N. (2008). Is google making us stupid? *Yearbook of the National Society for the Study of Education*, 107(2):89–94.

Carterette, B. (2011). System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th ACM SIGIR*, pages 903–912.

Carterette, B., Bah, A., and Zengin, M. (2015). Dynamic test collections for retrieval evaluation. In *Proceedings of the 5th ACM ICTIR*, pages 91–100.

Carterette, B., Kanoulas, E., and Yilmaz, E. (2011). Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM CIKM*, pages 611–620.

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM CIKM*, pages 621–630.

Charles-Dominique, P. and Martin, R. (1972). *Behaviour and Ecology of Nocturnal Prosimians: Field Studies in Gabon and Madagascar*. Advances in ethology. P. Parey.

Charnov, E. (1976). Optimal foraging, the Marginal Value Theorem. *Theoretical Population Biology*, 9(2):129–136.

Chen, D., Chen, W., Wang, H., Chen, Z., and Yang, Q. (2012). Beyond ten blue links: Enabling user click modeling in federated web search. In *Proceedings of the 5th ACM WSDM*, pages 463–472.

Chen, H. and Dumais, S. (2000). Bringing order to the web: Automatically categorizing search results. In *Proceedings of the 18th ACM CHI*, pages 145–152.

Chen, M. C., Anderson, J. R., and Sohn, M. H. (2001). What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. In *Proceedings of the 19th ACM CHI Extended Abstracts*, pages 281–282.

Chen, P. P.-S. (1976). The entity-relationship model – toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36.

Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using information scent to model user information needs and actions and the web. In *Proceedings of the 19th ACM CHI 2001*, pages 490–497.

Chierichetti, F., Kumar, R., and Raghavan, P. (2011). Optimizing two-dimensional search results presentation. In *Proceedings of the 4th ACM WSDM*, pages 257–266.

Chuklin, A., Markov, I., and de Rijke, M. (2015). *Click Models for Web Search*. Morgan & Claypool.

Chuklin, A. and Serdyukov, P. (2012). Good abandonments in factoid queries. In *Proceedings of the 21st WWW*, pages 483–484.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st ACM SIGIR*, pages 659–666.

Clarke, C. L. A., Agichtein, E., Dumais, S., and White, R. W. (2007). The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th ACM SIGIR*, pages 135–142.

Cleverdon, C., Mills, J., and Keen, M. (1966). *Factors Determining the Performance of Indexing Systems*, volume 1:2 of *Factors Determining the Performance of Indexing Systems*.

Cleverdon, C. W. (1991). The significance of the cranfield tests on index languages. In *Proceedings of the 14th ACM SIGIR*, pages 3–12.

Collins-Thompson, K., Callan, J., Terra, E., and Clarke, C. L. (2004). The effect of document retrieval quality on factoid question answering performance. In *Proceedings of the 27th ACM SIGIR*, pages 574–575.

Collins-Thompson, K., Hansen, P., and Hauff, C. (2017). Search as learning (dagstuhl seminar 17092). In *Dagstuhl Reports*, volume 7.

Cooper, W. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37.

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41.

Cooper, W. S. (1973a). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100.

Cooper, W. S. (1973b). On selecting a measure of retrieval effectiveness part ii. implementation of the philosophy. *Journal of the American Society for Information Science*, 24(6):413–424.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 1st ACM WSDM 2008*, pages 87–94.

Crescenzi, A., Capra, R., and Arguello, J. (2013). Time pressure, user satisfaction and task difficulty. In *Proceedings of the 76th ASIS&T*.

Crescenzi, A., Kelly, D., and Azzopardi, L. (2016). Impacts of time constraints and system delays on user experience. In *Proceedings of the 1st ACM CHIIR*, pages 141–150.

Croft, B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA.

Cutrell, E. and Guan, Z. (2007). What are you looking for?: An eye-tracking study of information usage in web search. In *Proceedings of the 25th ACM CHI*, pages 407–416.

Cutts, Q., Connor, R., Michaelson, G., and Donaldson, P. (2014). Code or (not code): Separating formal and natural language in cs education. In *Proceedings of the 9th WiPSCE*, pages 20–28.

Das Sarma, A., Gollapudi, S., and Ieong, S. (2008). Bypass rates: Reducing query abandonment using negative inferences. In *Proceedings of the 14th ACM KDD*, pages 177–185.

Dewey, M. (1891). Decimal classification and relative index for libraries, clippings, notes, etc. 240(41):407–593.

Diriye, A., White, R., Buscher, G., and Dumais, S. (2012). Leaving so soon?: Understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM CIKM*, pages 1025–1034.

Dolamic, L. and Savoy, J. (2010). When stopword lists make the difference. *Journal of the Association for Information Science and Technology*, 61(1):200–203.

Dostert, M. and Kelly, D. (2009). Users' stopping behaviors and estimates of recall. In *Proceedings of the 32nd ACM SIGIR*, pages 820–821.

Dumais, S., Cutrell, E., and Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of the 19th ACM CHI*, pages 277–284.

Edwards, A., Kelly, D., and Azzopardi, L. (2015). The impact of query interface design on stress, workload and performance. In *Proceedings of the 37th ECIR*, pages 691–702.

Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11):989–1003.

Eliot, S. and Rose, J. (2009). *A Companion to the History of the Book*. Blackwell Companions to Literature and Culture. John Wiley & Sons.

Farquhar, P. H. and Pratkanis, A. R. (1993). Decision structuring with phantom alternatives. *Management Science*, 39(10):1214–1226.

Feild, H., Jones, R., Miller, R., Nayak, R., Churchill, E., and Velipasaoglu, E. (2010). Logging the search self-efficacy of amazon mechanical turkers. In *Proceedings of the CSE SIGIR Workshop*, pages 27–30.

Fischhoff, B. (1977). Cost benefit analysis and the art of motorcycle maintenance. *Policy Sciences*, 8(2):177–202.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2):330.

Fishwick, P. A. (1995). Computer simulation: The art and science of digital world construction. Technical report, University of Florida.

Fox, C. (1992). Information retrieval. chapter Lexical Analysis and Stoplists, pages 102–130.

Francis, W. and Kučera, H. (1979). *Manual of Information to Accompany A Standard Corpus of Present-day Edited American English, for Use with Digital Computers*. Brown University, Department of Lingustics.

Francis, W. and Kučera, H. (1985). Frequency analysis of english usage: Lexicon and grammar. *Journal of English Linguistics*, 18(1):64–70.

Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265.

Fuhr, N. and Lalmas, M. (2006). Advances in xml retrieval: The inex initiative. In *Proceedings of IWRIDL*, page 16.

Gettys, C. F. and Fisher, S. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, 24(1):93–110.

Gibb, J. A. (1958). Predation by tits and squirrels on the eucosmid ernarmonia conicolana (heyl.). *Journal of Animal Ecology*, 27(2):375–396.

Gigerenzer, G. and Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart*, pages 75–95. Oxford University Press.

Green, R. (1984). Stopping rules for optimal foragers. *The American Naturalist*, 123(1):30–43.

Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y., and Faloutsos, C. (2009). Click chain model in web search. In *Proceedings of the 18th WWW*, pages 11–20.

Hagen, M., Michel, M., and Stein, B. (2015). What was the query? generating queries for document sets with applications in cluster labeling. In *Natural Language Processing and Information Systems*, pages 124–133.

Hagen, M., Michel, M., and Stein, B. (2016). Simulating ideal and average users. In *Proceedings of the 12th AIRS*, pages 138–154.

Harman, D. (1993). Overview of the first trec conference. In *Proceedings of the 16th ACM SIGIR*, SIGIR '93, pages 36–47.

Harman, D. (2010). Is the cranfield paradigm outdated? In *Proceedings of the 33rd ACM SIGIR*, page 1.

Harper, D. J. and Kelly, D. (2006). Contextual relevance feedback. In *Proceedings of the 1st ACM IIiX*, pages 129–137.

Harrower, M. and Brewer, C. A. (2003). Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37.

Harvey, M. and Pointon, M. (2017). Searching on the go: the effects of fragmented attention on mobile web search tasks. In *Proceedings of the 40th ACM SIGIR*, pages 155–164.

Hassan, A., Shi, X., Craswell, N., and Ramsey, B. (2013). Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd CIKM*, pages 2019–2028.

Hassan, A. and White, R. (2013). Personalized models of search satisfaction. In *Proceedings of the 22nd ACM CIKM*, pages 2009–2018.

He, J., Duboue, P., and Nie, J.-Y. (2012). Bridging the gap between intrinsic and perceived relevance in snippet generation. In *Proceedings of COLING 2012*, pages 1129–1146.

Hearst, M. (2009). *Search user interfaces*. Cambridge University Press.

Hearst, M. A. (1995). Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the 13th ACM SIGCHI*, pages 59–66.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

Heermann, D. W. (1990). *Computer-Simulation Methods*, pages 8–12.

Heine, M. D. (1981). Simulation, and simulation experiments. In Spärck Jones, K., editor, *Information Retrieval Experiments*, pages 197–198. Butterworth-Heinemann.

Hiemstra, D. (2009). *Information Retrieval Models*, pages 1–19.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2):79–102.

Huang, J., White, R. W., and Dumais, S. (2011). No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the 29th ACM CHI*, pages 1225–1234.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer Publishing Company, Incorporated.

Iwasa, Y., Higashi, M., and Yamamura, N. (1981). Prey distribution as a factor determining the choice of optimal foraging strategy. *The American Naturalist*, 117(5):710–723.

Iwata, M., Sakai, T., Yamamoto, T., Chen, Y., Liu, Y., Wen, J.-R., and Nishio, S. (2012). Aspectiles: Tile-based visualization of diversified web search results. In *Proceedings of the 35th ACM SIGIR*, pages 85–94.

Jahoda, G. (1961). Electronic searching. volume 4 of *The state of the library art*, pages 139–320. Graduate School of Library Service, Rutgers University.

Janetos, A. C. and Cole, B. J. (1981). Imperfectly optimal animals. *Behavioral Ecology and Sociobiology*, 9(3):203–209.

Jansen, B. J., Booth, D. L., and Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44(3):1251–1266.

Jansen, B. J. and Spink, A. (2005). Analysis of document viewing patterns of web search engine users. In *Web mining: Applications and techniques*, pages 339–354.

Jansen, B. J. and Spink, A. (2006). How are we searching the world wide web? a comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263.

Järvelin, K. and Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd ACM SIGIR*, pages 41–48.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM KDD*, pages 133–142.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th ACM SIGIR*, pages 154–161.

Joho, H. and Jose, J. M. (2006). A comparative study of the effectiveness of search result presentation on the web. In *Proceedings of the 28th ECIR*, pages 302–313.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing and Management*, 36(6):779–808.

Jordan, C., Watters, C., and Gao, Q. (2006). Using controlled query generation to evaluate blind relevance feedback algorithms. In *Proceedings of the 6th ACM/IEEE-CS JCDL*, pages 286–295.

Kaisser, M., Hearst, M. A., and Lowe, J. B. (2008). Improving search results quality by customizing summary lengths. In *Proceedings of the 46th ACL*, pages 701–709.

Kammerer, Y. and Gerjets, P. (2010). How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In *Proceedings of the Symposium on Eye-Tracking Research & Applications*, pages 299–306.

Kando, N., Eguchi, K., and Kuriyama, K. (1999). Construction of a large scale test collection: Analysis of the test topics of the NTCIR-1. In *Proceedings of IPSJ Annual Meeting*, pages 3–107.

Kanungo, T. and Orr, D. (2009). Predicting the readability of short web summaries. In *Proceedings of the 2nd ACM WSDM*, pages 202–211.

Kato, M. P., Sakai, T., and Tanaka, K. (2012). Structured query suggestion for specialization and parallel movement: Effect on search behaviors. In *Proceedings of the 21st WWW*, pages 389–398.

Kazai, G., Kamps, J., Koolen, M., and Milic-Frayling, N. (2011). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th ACM SIGIR*, pages 205–214.

Keenan, S., Smeaton, A. F., and Keogh, G. (2001). The effect of pool depth on system evaluation in trec. *Journal of the Association for Information Science and Technology*, 52(7):570–574.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224.

Kelly, D., Arguello, J., Edwards, A., and Wu, W.-C. (2015). Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 1st ACM ICTIR*, pages 101–110.

Kelly, D. and Azzopardi, L. (2015). How many results per page?: A study of SERP size, search behavior and user experience. In *Proceedings of the 38th ACM SIGIR*, pages 183–192.

Kelly, D. and Gyllstrom, K. (2011). An examination of two delivery modes for interactive search system experiments: Remote and laboratory. In *Proceedings of the 29th ACM CHI*, pages 1531–1540.

Kelly, D., Gyllstrom, K., and Bailey, E. W. (2009). A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd ACM SIGIR*, pages 371–378.

Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T., and Lykke, M. (2009). Test collection-based IR evaluation needs extension toward sessions — A case of extremely short queries. In *Proceedings of the 5th AIRS*, pages 63–74.

Khabsa, M., Crook, A., Awadallah, A. H., Zitouni, I., Anastasakos, T., and Williams, K. (2016). Learning to account for good abandonment in search success metrics. In *Proceedings of the 25th ACM CIKM*, pages 1893–1896.

Kim, J., Thomas, P., Sankaranarayana, R., and Gedeon, T. (2012). Comparing scanning behaviour in web search on small and large screens. In *Proceedings of the 17th ADCS*, pages 25–30.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., and Yoon, H.-J. (2014). Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology*.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., and Yoon, H.-J. (2016). Pagination versus scrolling in mobile web search. In *Proceedings of the 25th ACM CIKM*, pages 751–760.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., and Yoon, H.-J. (2017). What snippet size is needed in mobile web search? In *Proceedings of the 2nd ACM CHIIR*, pages 97–106.

Kiseleva, J., Kamps, J., Nikulin, V., and Makarov, N. (2015). Behavioral dynamics from the serp's perspective: What are failed serps and how to fix them? In *Proceedings of the 24th ACM CIKM*, pages 1561–1570.

Koch, S., Bosch, H., Giereth, M., and Ertl, T. (2009). Iterative integration of visual insights during patent search and analysis. In *Visual Analytics Science and Technology*, pages 203–210.

Kogut, C. A. (1990). Consumer search behavior and sunk costs. *Journal of Economic Behavior and Organization*, 14(3):381–392.

Kraft, D. and Lee, T. (1979). Stopping rules and their effect on expected search length. *IPM*, 15(1):47 – 58.

Krebs, J. (1973). *Behavioral Aspects of Predation*, pages 73–111. Springer US, Boston, MA.

Krebs, J., Ryan, J., and Charnov, E. (1974). Hunting by expectation or optimal foraging? a study of patch use by chickadees. *Animal Behaviour*, 22, Part 4:95–964.

Krebs, J. R., Stephens, D. W., Sutherland, W. J., and Myers, J. P. (1983). Perspectives in optimal foraging. *Perspectives in Ornithology*, pages 165—222.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th ACM SIGIR*, pages 191–202.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th ACM SIGIR*, pages 68–73.

Lancaster, F. (1968). *Information retrieval systems: characteristics, testing, and evaluation*. Information sciences series. Wiley.

Landauer, T., Egan, D., Remde, J., Lesk, M., Lochbaum, C., and Ketchum, D. (1993). Enhancing the usability of text through computer delivery and formative evaluation: the superbook project. *Hypertext: A psychological perspective*, pages 71–136.

Leal-Bando, L., Scholer, F., and Turpin, A. (2015). Query-biased summary generation assisted by query expansion. *JASIST*, 66(5):961–979.

Li, Q. and Chen, Y. P. (2010). Personalized text snippet extraction using statistical language models. *Pattern Recognition*, 43(1):378–386.

Li, Y. and Hu, D. (2013). Interactive retrieval using simulated versus real work task situations: Differences in sub-facets of tasks and interaction performance. In *Proceedings of the 76th ASIS&T*, pages 41:1–41:10.

Linden, G. (2006). *Marissa mayer at web 2.0*. `http://glinden.blogspot.com/2006/11/marissa-mayer-at-web-20.html` (retrieved on August 10th, 2018).

Lo, R. T.-W., He, B., and Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. In *Proceedings of the 5th Dutch-Belgian IR Workshop*, pages 17–24.

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., and Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052.

Loumakis, F., Stumpf, S., and Grayson, D. (2011). This image smells good: Effects of image info. scent in search engine results pages. In *Proceedings of the 20th ACM CIKM*, pages 475–484.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Luo, J., Zhang, S., Dong, X., and Yang, H. (2015). *Proceedings of the 37th ECIR*, chapter Designing States, Actions, and Rewards for Using POMDP in Session Search, pages 526–537.

Luo, J., Zhang, S., and Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th ACM SIGIR*, pages 587–596.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*.

Mansourian, Y. and Ford, N. (2007). Search persistence and failure on the web: a "bounded rationality" and "satisficing" analysis. *Journal of Documentation*, 63(5):680–701.

March, J. G. (1994). *Primer on decision making: How decisions happen*. Simon and Schuster.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press.

Marchionini, G., Dwiggins, S., Katz, A., and Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*, 15(1):35–69.

Marcos, M.-C., Gavin, F., and Arapakis, I. (2015). Effect of snippets on user experience in web search. In *Proceedings of the 16th HCI*, pages 47:1–47:8.

Marshall, C. C. and Shipman, F. M. (1997). Spatial hypertext and the practice of information triage. In *Proc. 8th ACM Hypertext*, pages 124–133.

Maxwell, D. (2016). Building realistic simulations for interactive information retrieval. In *Proceedings of the 1st ACM CHIIR*, pages 357–359.

Maxwell, D. and Azzopardi, L. (2014). Stuck in traffic: How temporal delays affect search behaviour. In *Proceedings of the 5th IIiX*, pages 155–164.

Maxwell, D. and Azzopardi, L. (2016a). Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM CIKM*, pages 731–740.

Maxwell, D. and Azzopardi, L. (2016b). Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proceedings of the 39th ACM SIGIR*, pages 1141–1144.

Maxwell, D. and Azzopardi, L. (2018). Information scent, searching and stopping: Modelling SERP level stopping behaviour. In *Proceedings of the 40th ECIR*, pages 210–222.

Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015a). An initial investigation into fixed and adaptive stopping strategies. In *Proceedings of the 38th ACM SIGIR*, pages 903–906.

Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015b). Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM CIKM*, pages 313–322.

Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of the 40th ACM SIGIR*, pages 135–144.

Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2019). The impact of result diversification on search behaviour and performance. *Information Retrieval Journal*. In press.

McBryan, O. A. (1994). GENVL and WWWW: Tools for taming the web. In *Proceedings of the 1st WWW*.

McDonald, J., Ogden, W., and Foltz, P. (1998). Interactive information retrieval using term relationship networks. *NIST Special Publication*, pages 379–384.

McMinn, A. J. (2018). *Real-Time Event Detection using Twitter*. PhD thesis, University of Glasgow.

McNair, J. N. (1982). Optimal giving-up times and the marginal value theorem. *The American Naturalist*, 119(4):511–529.

McNamee, P. (2006). Exploring new languages with haircut at clef 2005. In *Proceedings of the 6th CLEF*, pages 155–164.

Mitra, B. and Craswell, N. (2017). Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*.

Moffat, A., Thomas, P., and Scholer, F. (2013). Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM CIKM*, pages 659–668.

Moffat, A. and Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):2:1–2:27.

Mooers, C. (1950). *The theory of digital handling of non-numerical information and its implications to machine economics*. Zator technical bulletin.

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8):114.

Muralidharan, A., Gyongyi, Z., and Chi, E. (2012). Social annotations in web search. In *Proceedings of the 21st ACM CHI*, pages 1085–1094.

Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A., and Smola, A. (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd WWW*, pages 953–964.

Nickles, K. (1995). *Judgment-based and reasoning-based stopping rules in decision making under uncertainty*. PhD thesis, University of Minnesota.

Ofoghi, B., Yearwood, J., and Ghosh, R. (2006). A semantic approach to boost passage retrieval effectiveness for question answering. In *Proceedings of the 29th ACSC*, pages 95–101.

Olston, C. and Chi, E. (2003). Scenttrails: Integrating browsing and searching on the web. *ACM Transactions on Computer-Human Interactions*, 10(3).

Ong, K., Järvelin, K., Sanderson, M., and Scholer, F. (2017). Using information scent to understand mobile and desktop web search behavior. In *Proceedings of the 40th ACM SIGIR*, pages 295–304.

Oulasvirta, A., Hukkinen, J., and Schwartz, B. (2009). When more is less: The paradox of choice in search engine use. In *Proceedings of the 32nd ACM SIGIR*, pages 516–523.

Over, P. (1998). Trec-6 interactive track report. pages 73–82.

Over, P. (2001). The trec interactive track: an annotated bibliography. *Information Processing and Management*, 37(3):369–381.

Pääkkönen, T., Järvelin, K., Kekäläinen, J., Keskustalo, H., Baskaya, F., Maxwell, D., and Azzopardi, L. (2015). Exploring behavioral dimensions in session effectiveness. In *Proceedings of the 6th CLEF*, pages 178–189.

Paek, T., Dumais, S., and Logan, R. (2004). Wavelens: A new view onto internet search results. In *Proceedings of the 22nd ACM CHI*, pages 727–734.

Pedersen, J., Cutting, D., Tukey, J., et al. (1991). Snippet search: A single phrase approach to text access. In *Proceedings of the 1991 Joint Statistical Meetings*.

Perkins, D. N., Allen, R., and Hafner, J. (1983). Difficulties in everyday reasoning. *Thinking: The expanding frontier*, pages 177–189.

Peters, C. and Braschler, M. (2001). European Research Letter: Cross-language System Evaluation: The CLEF Campaigns. *Journal of the Association for Information Science and Technology*, 52(12):1067–1072.

Pirolli, P. (2007). *Information Foraging Theory: Adaptive interaction with information*. Human Technology Interaction Series. Oxford University Press, USA.

Pirolli, P. and Card, S. (1995). Information foraging in information access environments. In *Proc. 13th ACM SIGCHI*, pages 51–58.

Pirolli, P. and Card, S. K. (1999). Information foraging. *Psychological Review*, 106:643–675.

Pirolli, P., Schank, P., Hearst, M., and Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the 14th ACM CHI*, pages 213–220.

Pitts, M. G. and Browne, G. J. (2004). Stopping behavior of systems analysts during information requirements elicitation. *Journal of Management Information Systems*, 21(1):203–226.

Pitz, G., Reinhold, H., and Geller, E. S. (1969). Strategies of information seeking in deferred decision making. 4:1–19.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Prabha, C., Connaway, L., Olszewski, L., and Jenkins, L. (2007). What is enough? Satisficing information needs. *Journal of Documentation*, 63(1):74–89.

Reisberg, D. (1997). *Cognition: Exploring the science of the mind.* WW Norton & Co.

Resnick, M. L., Maldonado, C., Santos, J. M., and Lergier, R. (2001). Modeling on-line search behavior using alternative output structures. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 45, pages 1166–1170.

Robertson, S. (2008). On the history of evaluation in ir. *Journal of Information Science*, 34(4):439–456.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1995). Okapi at trec–3. In *Overview of the Third Text REtrieval Conference (TREC–3)*, page 109–126.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304.

Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th WWW*, pages 13–19.

Rose, D. E., Orr, D., and Kantamneni, R. G. P. (2007). Summary attributes and perceived search quality. In *Proceedings of the 16th WWW*, pages 1201–1202.

Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the 11th ACM CHI*, pages 269–276.

Ruthven, I. (2001). *Abduction, Explanation and Relevance Feedback.* PhD thesis, University of Glasgow.

Ruthven, I. (2008). Interactive information retrieval. *Annual Review of Information Science and Technology*, 42(1):43–91.

Saad, G. and Russo, J. (1996). Stopping criteria in sequential choice. *Organizational Behavior and Human Decision Processes*, 67(3):258 – 270.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375.

Sanderson, M. and Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE*, 100:1444–1451.

Sandstrom, P. E. (1994). An optimal foraging approach to information seeking and use. *The Library Quarterly: Information, Community, Policy*, 64(4):414–449.

Santos, R. L., Macdonald, C., and Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th WWW*, pages 881–890.

Savenkov, D., Braslavski, P., and Lebedev, M. (2011). Search snippet evaluation at yandex: lessons learned and future directions. *Multilingual and Multimodal Information Access Evaluation*, pages 14–25.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schwartz, B. (2005). *The Paradox of Choice: Why More Is Less*. Harper Perennial.

Shafir, E. and Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive psychology*, 24(4):449–474.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Simon, H. A. (1971). Decision making and organizational design. *Organizational Theory*, pages 189–212.

Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

Smith, G. F., Benson, P. G., and Curley, S. P. (1991). Belief, knowledge, and uncertainty: A cognitive perspective on subjective probability. *Organizational Behavior and Human Decision Processes*, 48(2):291–321.

Smucker, M. (2011). An analysis of user strategies for examining and processing ranked lists of documents. In *Proceedings of the 5th HCIR*.

Smucker, M., Guo, X., and Toulis, A. (2014). Mouse movement during relevance judging: Implications for determining user attention. In *Proceedings of the 37th ACM SIGIR*, pages 979–982.

Smucker, M. D. and Clarke, C. L. (2012). Time-based calibration of effectiveness measures. In *Proceedings of the 35th ACM SIGIR*, pages 95–104.

Soper, H. (1918). Means for compiling tabular and statistical data. U.S. Patent US00135169231-1920.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Spirin, N. V., Kotov, A. S., Karahalios, K. G., Mladenov, V., and Izhutov, P. A. (2016). A comparative study of query-biased and non-redundant snippets for structured search on mobile devices. In *Proceedings of the 25th ACM CIKM*, pages 2389–2394.

Spool, J. and Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. In *Proc. 19th ACM CHI Extended Abstracts*, pages 285–286.

Stephens, D. and Krebs, J. (1986). *Foraging Theory*. Monographs in Behavior and Ecology. Princeton University Press.

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing and Management*, 28(4):503–516. Special Issue: Evaluation Issues in Information Retrieval.

Sundar, S., Knobloch-Westerwick, S., and Hastall, M. (2007). News cues: Info. scent and cognitive heuristics. *Journal of the Association for Information Science Technology*, 58(3):366–378.

Svore, K. M., Teevan, J., Dumais, S. T., and Kulkarni, A. (2012). Creating temporally dynamic web search snippets. In *Proceedings of the 35th ACM SIGIR*, pages 1045–1046.

Syed, R. and Collins-Thompson, K. (2017). Retrieval algorithms optimized for human learning. In *Proceedings of the 40th ACM SIGIR*, pages 555–564.

Teevan, J., Cutrell, E., Fisher, D., Drucker, S. M., Ramos, G., André, P., and Hu, C. (2009). Visual snippets: Summarizing web pages for search and revisitation. In *Proceedings of the 27th ACM CHI*, pages 2023–2032.

Thomas, P., Moffat, A., Bailey, P., and Scholer, F. (2014). Modeling decision points in user search behavior. In *Proceedings of the 5th IIiX*, pages 239–242.

Tocher, K. (1963). *The art of simulation*. Electrical engineering series. English Universities Press.

Tombros, A. and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st ACM SIGIR*, pages 2–10.

Toms, E. G. and Freund, L. (2009). Predicting stopping behaviour: A preliminary analysis. In *Proceedings of the 32nd ACM SIGIR*, pages 750–751.

Tran, V., Maxwell, D., Fuhr, N., and Azzopardi, L. (2017). Personalised search time prediction using markov chains. In *Proceedings of the 3rd ACM ICTIR*, pages 237–240.

Turpin, A., Scholer, F., Jarvelin, K., Wu, M., and Culpepper, J. S. (2009). Including summaries in system evaluation. In *Proceedings of the 32nd ACM SIGIR*, pages 508–515.

Turpin, A., Tsegay, Y., Hawking, D., and Williams, H. E. (2007). Fast generation of result snippets in web search. In *Proceedings of the 30th ACM SIGIR*, pages 127–134.

Umemoto, K., Yamamoto, T., and Tanaka, K. (2016). Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proceedings of the 39th ACM SIGIR*, pages 405–414.

van Rijsbergen, C. (1979). *Information Retrieval*. Butterworth-Heinemann.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management*, 40(4):677–691.

Veerasamy, A. and Belkin, N. J. (1996). Evaluation of a tool for visualization of information retrieval results. In *Proceedings of the 19th ACM SIGIR*, pages 85–92.

Veerasamy, A. and Heikes, R. (1997). Effectiveness of a graphical display of retrieval results. In *Proceedings of the 20th ACM SIGIR*, pages 236–245.

Verberne, S., Sappelli, M., Järvelin, K., and Kraaij, W. (2015). User simulations for interactive search: Evaluating personalized query suggestion. In *Proceedings of the 37th ECIR*, volume 9022, pages 678–690.

Villa, R., Cantador, I., Joho, H., and Jose, J. M. (2009). An aspectual interface for supporting complex search tasks. In *Proceedings of the 32nd ACM SIGIR*, pages 379–386.

Voorhees, E. (2006). Overview of the trec 2005 robust retrieval track. In *Proceedings of TREC-14*.

Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. In *Proceedings of the 7th CLEF Initiative*, pages 355–370.

Voorhees, E. M. and Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*.

Wald, A. (1948). Sequential analysis. *Social Forces*, 27(2):170–171.

White, R. W. and Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM CIKM*, pages 87–96.

White, R. W., Jose, J. M., and Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management*, 39(5):707–733.

Wilkie, C. and Azzopardi, L. (2017). Algorithmic bias: Do good systems make relevant documents more retrievable? In *Proceedings of the 26th ACM CIKM*, pages 2375–2378.

Wilson, M. L., Kules, B., Schraefel, M., and Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97.

Woodruff, A., Rosenholtz, R., Morrison, J. B., Faulring, A., and Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for web search tasks. *Journal of the American Society for Information Science and Technology*, 53(2):172–185.

Wu, W.-C. (2012). How far will you go?: Using need for closure and information scent to model search stopping behavior. In *Proceedings of the 4th IIiX*, pages 328–328.

Wu, W.-C. and Kelly, D. (2014). Online search stopping behaviors: An investigation of query abandonment and task stopping. *Proceedings of the 77ASIST*, 51(1):1–10.

Wu, W.-C., Kelly, D., and Sud, A. (2014). Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th ACM SIGIR*, pages 557–566.

Yang, P. and Fang, H. (2017). Can short queries be even shorter? In *Proceedings of the 40th ACM SIGIR*, pages 43–50.

Yates, J. (1990). Judgment and decision making. *Journal of Behavioral Decision Making*, 4(1):76–78.

Yates, J. and Carlson, B. (1982). Toward a representational theory of decision making. *Ann Arbor, MI: University of Michigan, Working Paper*.

Yilmaz, E., Shokouhi, M., Craswell, N., and Robertson, S. (2010). Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM CIKM*, pages 1561–1564.

Zach, L. (2005). When is "enough" enough? modeling the information-seeking and stopping behavior of senior arts administrators: Research articles. *Journal of the Association for Information Science and Technology*, 56(1):23–35.

Zhang, F., Liu, Y., Li, X., Zhang, M., Xu, Y., and Ma, S. (2017a). Evaluating web search with a bejeweled player model. In *Proceedings of the 40th ACM SIGIR*, pages 425–434.

Zhang, Y., Liu, X., and Zhai, C. X. (2017b). Information retrieval evaluation as search simulation: A general formal framework for ir evaluation. In *Proceedings of the 3rd ACM ICTIR*, pages 193–200.

Zhang, Y., Park, L. A. F., and Moffat, A. (2010). Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69.

Zipf, G. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*.

Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J. M., and Azzopardi, L. (2013). Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval*, 16(2):267–305.