Composite Web Search

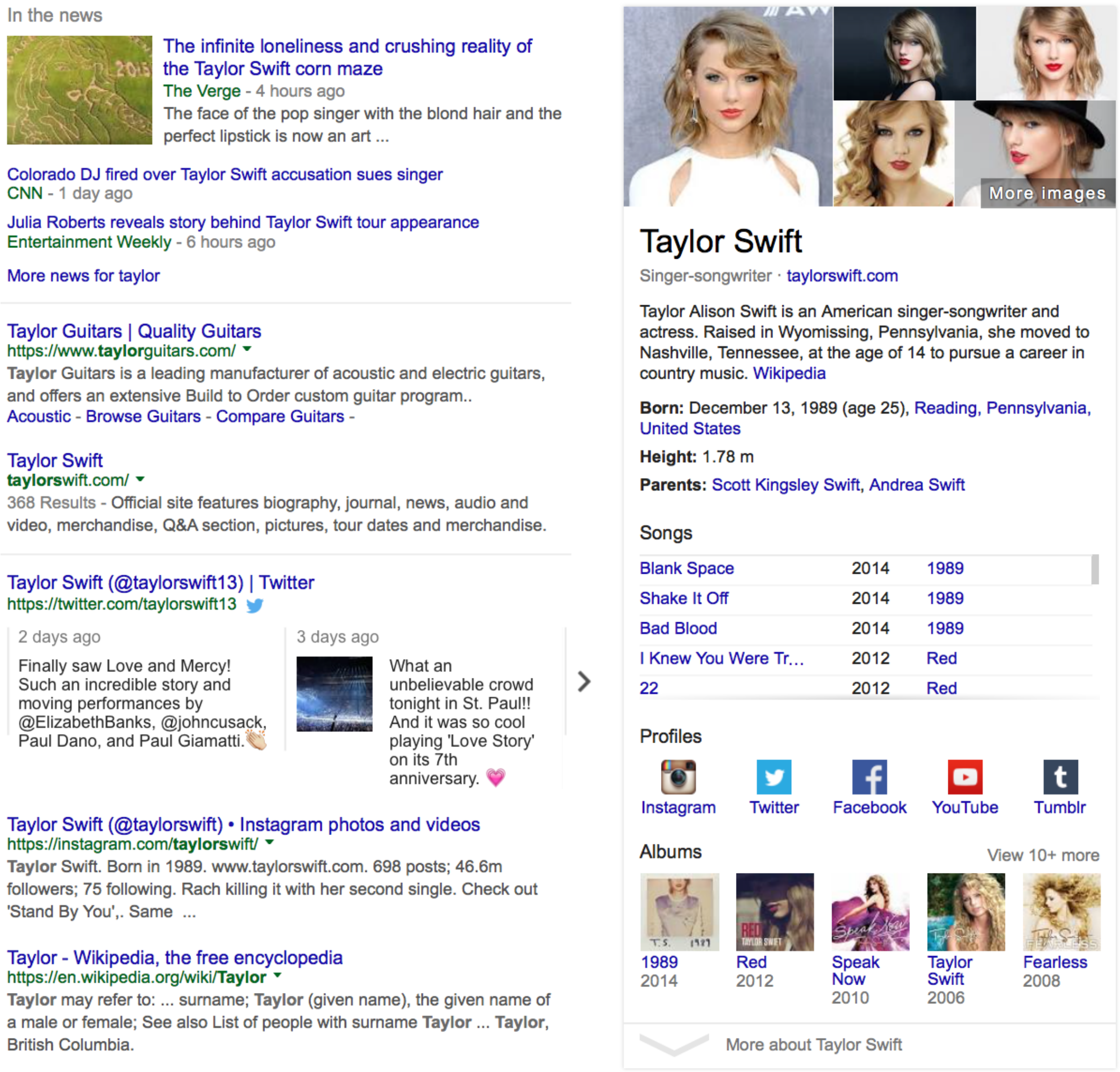
Horațiu S. Bota

Submitted in partial fulfilment of the requirements for the
Degree of Doctor of Philosophy

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

September, 2018

Abstract



In the news

The infinite loneliness and crushing reality of the Taylor Swift corn maze
The Verge - 4 hours ago
The face of the pop singer with the blond hair and the perfect lipstick is now an art ...

Colorado DJ fired over Taylor Swift accusation sues singer
CNN - 1 day ago

Julia Roberts reveals story behind Taylor Swift tour appearance
Entertainment Weekly - 6 hours ago

More news for taylor

Taylor Guitars | Quality Guitars
https://www.taylorguitars.com/ ▾
Taylor Guitars is a leading manufacturer of acoustic and electric guitars, and offers an extensive Build to Order custom guitar program..
Acoustic - Browse Guitars - Compare Guitars -

Taylor Swift
taylorswift.com/ ▾
368 Results - Official site features biography, journal, news, audio and video, merchandise, Q&A section, pictures, tour dates and merchandise.

Taylor Swift (@taylorswift13) | Twitter
https://twitter.com/taylorswift13

2 days ago
Finally saw Love and Mercy! Such an incredible story and moving performances by @ElizabethBanks, @johncusack, Paul Dano, and Paul Giamatti.👏

3 days ago
What an unbelievable crowd tonight in St. Paul!! And it was so cool playing 'Love Story' on its 7th anniversary. 💗

Taylor Swift (@taylorswift) • Instagram photos and videos
https://instagram.com/**taylorswift**/ ▾
Taylor Swift. Born in 1989. www.taylorswift.com. 698 posts; 46.6m followers; 75 following. Rach killing it with her second single. Check out 'Stand By You',. Same ...

Taylor - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Taylor** ▾
Taylor may refer to: ... surname; **Taylor** (given name), the given name of a male or female; See also List of people with surname **Taylor** ... **Taylor**, British Columbia.

More images

Taylor Swift
Singer-songwriter · taylorswift.com

Taylor Alison Swift is an American singer-songwriter and actress. Raised in Wyomissing, Pennsylvania, she moved to Nashville, Tennessee, at the age of 14 to pursue a career in country music. Wikipedia

Born: December 13, 1989 (age 25), Reading, Pennsylvania, United States
Height: 1.78 m
Parents: Scott Kingsley Swift, Andrea Swift

Songs

Blank Space	2014	1989
Shake It Off	2014	1989
Bad Blood	2014	1989
I Knew You Were Tr...	2012	Red
22	2012	Red

Profiles

Instagram Twitter Facebook YouTube Tumblr

Albums View 10+ more

1989 2014	Red 2012	Speak Now 2010	Taylor Swift 2006	Fearless 2008

More about Taylor Swift

The figure above shows Google's results page for the query *"taylor swift"*, captured in March 2016. Assembled around the long-established list of search results is content extracted from various source — news items and tweets merged within the results ranking, images, songs and social media profiles displayed to the right of the ranking, in an interface element that is known as an entity card. Indeed, the entire page seems more like an assembly of content extracted from various sources, rather than just a ranked list of blue links.

Search engine result pages have become increasingly diverse over the past few years, with most commercial web search providers responding to user queries with different types of results, merged within a unified page. The primary reason for this diversity on the results page is that the web itself has become more diverse, given the ease with which creating and hosting different types of content on the web is possible today.

This thesis investigates the aggregation of web search results retrieved from various document sources (e.g., images, tweets, Wiki pages) within information "objects" to be integrated in the results page assembled in response to user queries. We use the terms "composite objects" or "composite results" to refer to such objects, and throughout this thesis use the terminology of Composite Web Search (e.g., result composition) to distinguish our approach from other methods of aggregating diverse content within a unified results page (e.g., Aggregated Search). In our definition, the aspects that differentiate composite information objects from aggregated search blocks are that composite objects ***(i)*** contain results from multiple sources of information, ***(ii)*** are specific to a common topic or facet of a topic rather than a grouping of results of the same type, and ***(iii)*** are not a uniform ranking of results ordered only by their topical relevance to a query.

The most widely used type of composite result in web search today is the entity card. Entity cards have become extremely popular over the past few years, with some informal studies suggesting that entity cards are now shown on the majority of result pages generated by Google. As composite results are used more and more by commercial search engines to address information needs directly on the results page, understanding the properties of such objects and their influence on searchers is an essential aspect of modern web search science.

The work presented throughout this thesis attempts the task of studying composite objects by exploring users' perspectives on accessing and aggregating diverse content manually, by analysing the effect composite objects have on search behaviour and perceived workload, and by investigating different approaches to constructing such objects from diverse results. Overall, our experimental findings suggest that items which play a central role within composite objects are decisive in determining their usefulness, and that the overall properties of composite objects (i.e., relevance, diversity and coherence) play a combined role in mediating object usefulness.

Declaration

I hereby declare that except where explicit reference is made to the work of others, the contents of this 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 thesis is the result of my own work, under the supervision of Dr. Ke Zhou, Dr. John Williamson and Prof. Joemon M. Jose and references work done in collaboration, specifically indicated in the text.

Horațiu S. Bota
September, 2018

To my family.

Contents

Abstract i

Declaration iii

1 Introduction 1

1.1 Preamble 1

1.2 Definition 7

1.3 Thesis Statement 8

1.4 Motivation 9

1.5 Contributions 17

1.6 Publications 18

1.7 Thesis Structure 19

2 Background 20

2.1 User Models of Search 20

2.1.1 Information Retrieval 22

2.1.2 Information Seeking 22

2.1.3 Information Objects 24

2.1.4 Relevance versus Usefulness 25

2.1.5 Other Models of Search 27

2.1.6 Other Types of Search 28

2.2 System Models of Search 30

2.2.1 Information Retrieval Models 30

2.2.2 Evaluation of Retrieval Models 39

2.2.3 Retrieval Across Collections 45

2.2.4 Clustering and Composition 50

2.3 Chapter Summary 55

3 Searching the Web 56

3.1 Search Interactions 56

3.1.1 Recording Search Interactions 57
3.1.2 Types of Interactions 59
3.2 Search Interfaces 62
3.2.1 Search Results Presentation 63
3.2.2 Assistance in Browsing Results 66
3.2.3 Aggregated Search Results 70
3.2.4 Entity Cards 73
3.2.5 Other Search Interface Elements 75
3.3 Chapter Summary 76

4 Exploring Result Composition from the Users' Perspective 77
4.1 Introduction 77
4.2 Prior Work 79
4.2.1 Composite Retrieval 79
4.2.2 User Behaviour in Information Access 79
4.3 Experimental Methodology 80
4.3.1 Task Design 81
4.3.2 Object-level Characteristics 82
4.3.3 Limitations 82
4.3.4 Experimental System 83
4.4 Experimental Results 84
4.4.1 Composite Object Contents 84
4.4.2 Composite Object Characteristics 91
4.5 Discussion and Conclusions 93
4.6 Chapter Summary 94

5 The Effect of Entity Cards on Search Behaviour and Workload 96
5.1 Introduction 97
5.2 Prior Work 100
5.3 Experimental Methodology 101
5.3.1 Overview 101
5.3.2 Input Variables 105
5.3.3 Output Variables 109
5.3.4 Study Details 110
5.3.5 Statistical Methodology 113
5.4 Experimental Results 117
5.4.1 Preliminary Results 117
5.4.2 Card Content 120

5.4.3 Card Coherence 127
5.4.4 Card Diversity 135
5.5 Discussion and Conclusions 140
5.6 Chapter Summary 143

6 Composite Retrieval of Heterogeneous Web Search 145
6.1 Introduction 146
6.2 Prior Work 148
6.3 Result Composition Framework 150
6.3.1 Problem Formulation 150
6.3.2 Challenges 152
6.4 Composite Object Selection & Ranking 153
6.4.1 Produce Composite Objects 154
6.4.2 Choose Objects 156
6.4.3 Evidence and Estimation 159
6.5 Experimental Setup 161
6.5.1 Data 161
6.5.2 Evaluation 162
6.5.3 Baseline Systems 164
6.6 Experimental Results 165
6.6.1 Homogeneous Composite Retrieval 165
6.6.2 Heterogeneous Composite Retrieval 167
6.7 Discussion and Conclusions 173
6.8 Chapter Summary 174

7 Click Graph Representation Learning for Heterogeneous Documents 176
7.1 Introduction 176
7.2 Prior Work 179
7.2.1 Click Graphs 180
7.2.2 Aggregated Search 180
7.3 Data and Methods 180
7.3.1 Click Log 180
7.3.2 Click Graphs 181
7.3.3 Propagation Algorithm 185
7.3.4 Experimental Design 187
7.4 Experimental Results 190
7.4.1 Graph Connectedness 190
7.4.2 Retrieval Effectiveness 193

7.4.3 Modifying Graph Structure 196
7.5 Chapter Summary 201

8 Discussion and Future Work 203
8.1 Discussion 203
8.2 Future Work 207
8.3 Chapter Summary 212

9 Conclusion 213
9.1 Summary of Contributions 214
9.2 Summary of Conclusions 215
9.3 Final Remarks 218

List of Tables

2.1 Types of information searching. 28

4.1 Object-level characteristics. 83
4.3 Subtopic agreement based on semantic similarity of titles assigned to composite objects by study participants. 86
4.5 Vertical origin of ornament documents. 90
4.6 Mean document relevance in multi-vertical composite objects. . . . 91
4.7 Correlation of criteria assessments with object pairwise preference 92
4.8 Linear correlation of composite object properties. 93

5.1 Queries and task descriptions used in card study. 106
5.2 Questions associated with individual dimensions of perceived task workload, as shown to study participants at the end of card study. . 111
5.3 Experimental conditions in card study. 111
5.5 Summary statistics for card study outcome measures. 119

6.1 Verticals used in creating the federated search collection used in the algorithmic result composition study. 161
6.3 Evaluation of result composition approaches in a heterogeneous environment. 166
6.4 Performance of various result composition approaches in a heterogeneous web environment. 168
6.5 Comparison of best-performing result composition approaches in a heterogeneous web environment against all baselines. 169
6.6 Robustness of result composition approaches. 172

7.1 Evaluation of direct ranking based on graph-learnt representations. 194
7.2 Evaluation of changes to graph structure and their impact on retrieval performance. 198
7.3 Top 9 terms learnt under different algorithm and graph configurations for the head query *"taylor swift"*. 200

List of Figures

1.1 Web search results over the past two decades. 3
1.2 A modern web search results page. 4
1.3 Web search functionality beyond the browser. 11

2.1 A model of search by Järvelin and Ingwersen (2004) 21
2.2 The cognitive communication system for information seeking by Ingwersen and Järvelin (2005) 24

3.1 Result categorisation interface used in Dumais et al. (2001). 67
3.2 Example search interface in-line answers. 75

4.1 Interface of system used to manually construct composite objects. . 81
4.2 Average number of documents in composite objects. 86
4.3 Number of unique verticals in all objects created by participants. . 87
4.4 Vertical structure of two vertical composite objects. 88
4.5 Vertical structure of three vertical composite objects. 88
4.6 User indicated most influential criterion for object preference. . . . 91

5.1 Results page assembled by Google in response to the query *"castro"*. 98
5.2 Search interface used in card study. 102
5.3 Structure of entity cards used in card experiment. 104
5.4 Entity card structure and different experimental manipulations of entity cards. 108
5.5 Example statistical analysis used in card study. 115
5.6 Mean effect size of content manipulations in card study. 126
5.7 Mean effect size of coherence manipulations in card study. 134
5.8 Mean effect size of diversity manipulations in card study. 139

7.1 Example disconnected bipartite click graph. 182
7.2 Example augmented bipartite click graph. 184
7.3 Estimating connectedness across graphs of various size. 191

7.4 Distribution of vertices across sub-graphs in the unmodified graph variant. 192
7.5 Distribution of heterogeneous documents across sub-graphs in the unmodified graph variant. 192
7.6 Distribution of query vertices across sub-graphs in different augmentations of the click graph. 197

8.1 High-level representation of a search engine's indexing process, reproduced from Croft et al. (2009, p. 15). 203
8.2 High-level representation of a search engine's query process, reproduced from Croft et al. (2009, p. 16). 204
8.3 High-level representation of a composite search engine. Top row illustrates the components of a composite search engine, whereas a subset of challenges associated with each component are illustrated below. Components and associated challenges that are explicitly addressed through work presented in this thesis are presented within a solid contour. 205

Chapter 1

Introduction

This thesis investigates the aggregation of web search results retrieved from various document sources (e.g., *images*, *tweets*, *Wiki pages*) within information *"objects"* to be integrated in the results page assembled in response to user queries.

1.1 Preamble

From Surfers to Searchers

Beyond accurate use of nomenclature and technical details, the *Internet*, the *World Wide web*, *search engines*, the *online* are, today, terms synonymous with information. Indeed, for many if not most users of the online, they are not only synonymous, but also become reality through the same outlet: Google's results page.

How has *"googling"* become a synonym for finding information? How is it that the web seems to contain *all* the information in the world? *"Just google it"*, *"let me google that for you"*, *"ask google for a restaurant nearby"*, *"do you google yourself?"* — but rarely, if ever, *"I couldn't find it on Google"*.

At 28 years of age (Berners-Lee et al., 1994), the web is now the default information repository in the world. Many factors have contributed to its phenomenal success, from advances in networking technology to the relatively low cost of becoming a web user. But among the many technological advances that have driven the web's global adoption, search is, arguably, the most influential. Search has reduced information access to the simplest interaction paradigm: in exchange for a few keywords, for most practical purposes, search engines scan the entirety of the web[1] and return a list of documents ordered by their relevance instantly. Through such simple and effective interaction, in contrast to following links or browsing

[1] In fact, search engines index less than 5% of pages available on the web, as seen at http://www.worldwidewebsize.com/.

categorised resources, search has established itself as the principal method of accessing online information today.

Even though influential, search is, however, a new method of accessing information on the web. The now dated idiom *"surfing the web"* is a reflection of how web browsing initially involved following information from link to link[2]. The web was originally set up as an unstructured but linked information repository; news websites might publish a list of news-worthy stories and the onus of locating relevant information, by following links, lay with the user. Although the option still remains, today, who *"presumes to surf the web"* any more? (Wilson et al., 2010) The reason for this shift from *surfing* to *searching* is that, over the past decade, the web has grown beyond what is navigable from one link to the next and *searching* has replaced surfing as the main path to information. And the online continues to grow: more people access the web across cultures; a new addressing system is being deployed to accommodate the flood of devices joining the internet[3]; with more devices, more content is created and hosted on the web, which in turn makes search indispensable, etc.

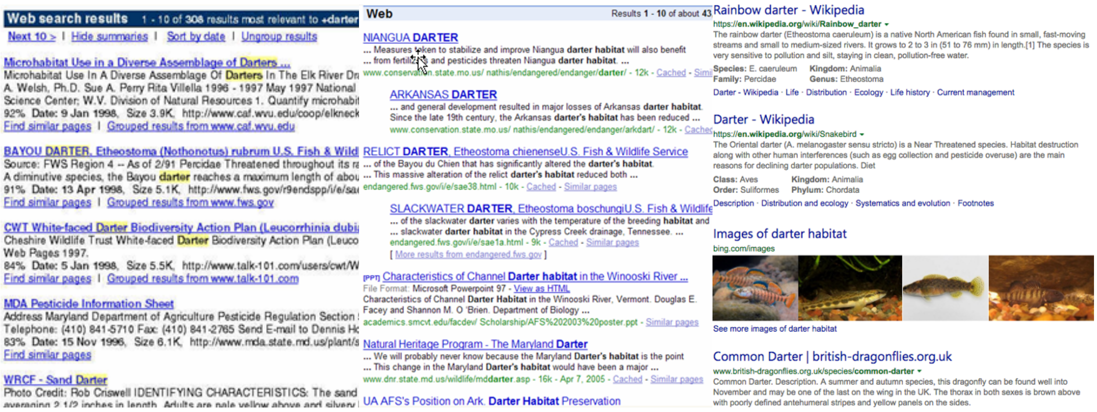
The unrelenting growth of the web feeds the need for constant web search improvement. The two are locked in a feedback loop: the easier it is for web users to locate and consume information through search, the more worthwhile it is for creators to publish their content on the web, *because* it can be easily found and consumed by vast populations of searchers. In the past few decades, hand in hand with the ever widening reach of the internet, this feedback loop has lead to an explosion of information on the web, in terms of both quantity and diversity of content. In turn, this vastness of diverse information creates the need for search systems that can not only *"find the needle in the haystack"* (i.e., locate and rank relevant documents), but also expose, in a unified interface, the many types of content generated on the web.

Search tension

Two broad areas of research are at the core of web search advancement: ***(i)*** algorithms (for retrieval, recommendation, etc.) and ***(ii)*** the use of search systems (search interfaces, user modelling, etc.). Retrieval algorithms have developed

[2] Interestingly, in my first language, Romanian, the most popular idiom used to describe web interactions is *"caută pe net"*, which literally means *"search the net"*. Similarly, in Spanish *"búscalo en internet"* is used. It seems cultures in which the internet became widely available only after search became the principal method of accessing information online were never exposed to this *"surfing"* of the web and did not even develop language to describe it.

[3] IPv6 is set to replace IPv4 as the new addressing system for devices connected to the internet because the IPv4 address space is too narrow to accommodate the sheer volume of devices.



Web search results 1 - 10 of 308 results most relevant to +darter

Next 10 > | Hide summaries | Sort by date | Ungroup results

Microhabitat Use in a Diverse Assemblage of Darters ...
Microhabitat Use In A Diverse Assemblage Of Darters In The Elk River Dr
A. Welsh, Ph.D. Sue A. Perry Rita Villella 1996 - 1997 May 1997 National
Science Center; W.V. Division of Natural Resources 1. Quantify microhabit
92% Date: 9 Jan 1998, Size 3.9K, http://www.caf.wvu.edu/coop/elkneck
Find similar pages | Grouped results from www.caf.wvu.edu

BAYOU DARTER, Etheostoma (Nothonotus) rubrum U.S. Fish & Wild
Source: FWS Region 4 -- As of 2/91 Percidae Threatened throughout its ra
A diminutive species, the Bayou darter reaches a maximum length of abou
91% Date: 13 Apr 1998, Size 5.1K, http://www.fws.gov/r9endspp/i/e/sa
Find similar pages | Grouped results from www.fws.gov

CWT White-faced Darter Biodiversity Action Plan (Leucorrhinia dubi
Cheshire Wildlife Trust White-faced Darter Biodiversity Action Plan (Leuco
Web Pages 1997.
84% Date: 5 Jan 1998, Size 5.5K, http://www.talk-101.com/users/cwt/W
Find similar pages | Grouped results from www.talk-101.com

MDA Pesticide Information Sheet
Address Maryland Department of Agriculture Pesticide Regulation Section
Telephone: (410) 841-5710 Fax: (410) 841-2765 Send E-mail to Dennis H
83% Date: 15 Nov 1996, Size 6.1K, http://www.mda.state.md.us/plant/s
Find similar pages

WRCF - Sand Darter
Photo Credit: Rob Criswell IDENTIFYING CHARACTERISTICS: The sand
averaging 2 1/2 inches in length. Adults are pale yellow above and silvery

Web Results **1** - **10** of about 43

NIANGUA **DARTER**
... Measures taken to stabilize and improve Niangua **darter habitat** will also benefit
... from fertilizers and pesticides threaten Niangua **darter habitat**. ...
www.conservation.state.mo.us/ nathis/endangered/endanger/**darter**/ - 12k - Cached - Simil

ARKANSAS **DARTER**
... and general development resulted in major losses of Arkansas **darter habitat**.
Since the late 19th century, the Arkansas **darter's habitat** has been reduced ...
www.conservation.state.mo.us/ nathis/endangered/endanger/arkdart/ - 12k - Cached

RELICT **DARTER**, Etheostoma chienenseU.S. Fish & Wildlife Service
... of the Bayou du Chien that has significantly altered the **darter's habitat**.
... This massive alteration of the relict **darter's habitat** reduced both ...
endangered.fws.gov/i/e/sae38.html - 10k - Cached - Similar pages

SLACKWATER **DARTER**, Etheostoma boschungiU.S. Fish & Wildlife
... of the slackwater **darter** varies with the temperature of the breeding **habitat** and
... slackwater **darter habitat** in the Cypress Creek drainage, Tennessee. ...
endangered.fws.gov/i/e/sae1a.html - 9k - Cached - Similar pages
[More results from endangered.fws.gov]

[PPT] Characteristics of Channel **Darter habitat** in the Winooski River ...
File Format: Microsoft Powerpoint 97 - View as HTML
Characteristics of Channel **Darter Habitat** in the Winooski River, Vermont. Douglas E.
Facey and Shannon M. O 'Brien. Department of Biology ...
academics.smcvt.edu/facdev/ Scholarship/AFS%202003%20poster.ppt - Similar pages

Natural Heritage Program - The Maryland **Darter**
... We will probably never know because the Maryland **Darter's habitat** is the point
... This change in the Maryland **Darter's habitat** would have been a major ...
www.dnr.state.md.us/wildlife/md**darter**.asp - 16k - Apr 7, 2005 - Cached - Similar pages

UA AFS's Position on Ark. **Darter Habitat** Preservation

Rainbow darter - Wikipedia
https://**en.wikipedia.org**/wiki/**Rainbow_darter** ▾
The rainbow darter (Etheostoma caeruleum) is a native North American fish found in small, fast-moving streams and small to medium-sized rivers. It grows to 2 to 3 in (51 to 76 mm) in length.[1] The species is very sensitive to pollution and silt, staying in clean, pollution-free water.
Species: E. caeruleum **Kingdom:** Animalia
Family: Percidae **Genus:** Etheostoma
Darter - Wikipedia · Life · Distribution · Ecology · Life history · Current management

Darter - Wikipedia
https://**en.wikipedia.org**/wiki/Snakebird ▾
The Oriental darter (A. melanogaster sensu stricto) is a Near Threatened species. Habitat destruction along with other human interferences (such as egg collection and pesticide overuse) are the main reasons for declining darter populations. Diet
Class: Aves **Kingdom:** Animalia
Order: Suliformes **Phylum:** Chordata
Description · Distribution and ecology · Systematics and evolution · Footnotes

Images of darter habitat
bing.com/images
See more images of darter habitat

Common Darter | british-dragonflies.org.uk
www.british-dragonflies.org.uk/species/**common-darter** ▾
Common Darter. Description. A summer and autumn species, this dragonfly can be found well into November and may be one of the last on the wing in the UK. The thorax in both sexes is brown above with poorly defined antehumeral stripes and yellow panels on the sides.

(a) Infoseek (1997) (b) Google (2007) (c) Bing (2017)

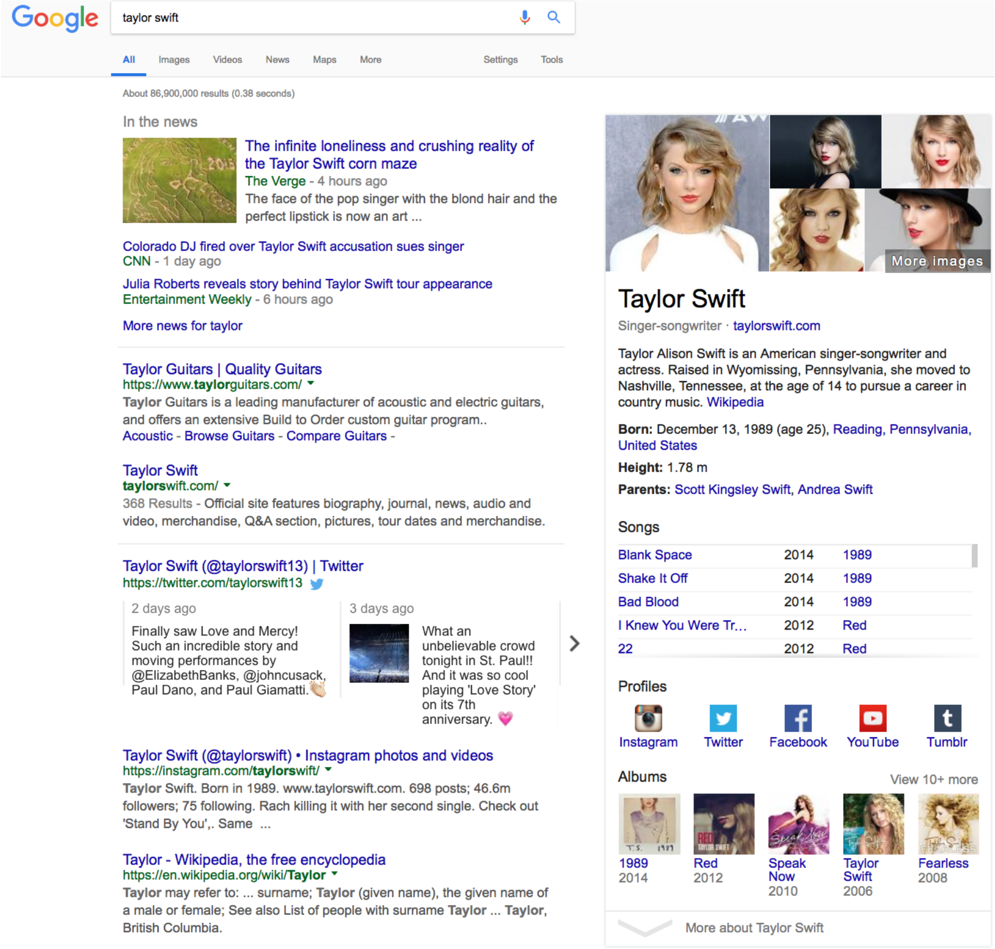
Figure 1.1: Search result listings for the same query (*"darter habitat"*) as returned by popular web search engines at various times over the past two decades. Figures 1.1a and 1.1b reproduced from *Search User Interfaces* (Hearst, 2009, p. 2, Figure 1.1). Figure 1.1c captured in December, 2017.

in various ways over the years, from incorporating users' historical preferences into document ranking, to tailoring search results to users' geographical context. More pertinent to this thesis, in addition to traditional, text-heavy resources, retrieval algorithms have been developed to also rank different types of documents, such as *images* or *videos*, as these documents became abundant on the web.

The workings of a search engine — in particular, its retrieval algorithm — are opaque to end users: how exactly the ranked list of documents is assembled is, for most people, a mystery. Changes to the machinery that ranks documents are, perhaps, less obvious to searchers. In contrast, web search interfaces are, from the user's point of view, the entirety of a search system and any change brought to their configuration has immediate impact on how users perceive or interact with the system. Given that modern web search engines serve a broad user base, expressing extremely diverse information needs, any *"improvement"* to the search interface is likely to be just as puzzling as it is helpful to different groups of searchers. As such, search interfaces are notoriously resistant to change. Figure 1.1 shows three web search interfaces observed over the past two decades; they are very similar, if not *"nearly identical"* (Hearst, 2009). [4]

However, these diverging concerns — on one hand, a rapidly changing, more and more diverse web (with retrieval algorithms better at organising this diversity); on the other hand, search interfaces resistant to change, not least with respect to the types of content shown on the main page of results, so as not to

[4] This is not to say that research on user interfaces for search is lacking – Hearst (2009) provides an expert review of historic and recent developments in the area – but that its transition to mainstream web search is, perhaps, slower-moving than in the case of retrieval algorithms.



About 86,900,000 results (0.38 seconds)

In the news

The infinite loneliness and crushing reality of the Taylor Swift corn maze
The Verge - 4 hours ago
The face of the pop singer with the blond hair and the perfect lipstick is now an art ...

Colorado DJ fired over Taylor Swift accusation sues singer
CNN - 1 day ago

Julia Roberts reveals story behind Taylor Swift tour appearance
Entertainment Weekly - 6 hours ago

More news for taylor

Taylor Guitars | Quality Guitars
https://www.**taylor**guitars.com/
Taylor Guitars is a leading manufacturer of acoustic and electric guitars, and offers an extensive Build to Order custom guitar program..
Acoustic - Browse Guitars - Compare Guitars -

Taylor Swift
taylorswift.com/
368 Results - Official site features biography, journal, news, audio and video, merchandise, Q&A section, pictures, tour dates and merchandise.

Taylor Swift (@taylorswift13) | Twitter
https://twitter.com/taylorswift13

2 days ago
Finally saw Love and Mercy! Such an incredible story and moving performances by @ElizabethBanks, @johncusack, Paul Dano, and Paul Giamatti.👏

3 days ago
What an unbelievable crowd tonight in St. Paul!! And it was so cool playing 'Love Story' on its 7th anniversary. 💗

Taylor Swift (@taylorswift) • Instagram photos and videos
https://instagram.com/**taylorswift**/
Taylor Swift. Born in 1989. www.taylorswift.com. 698 posts; 46.6m followers; 75 following. Rach killing it with her second single. Check out 'Stand By You',. Same ...

Taylor - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Taylor**
Taylor may refer to: ... surname; **Taylor** (given name), the given name of a male or female; See also List of people with surname **Taylor** ... **Taylor**, British Columbia.

Taylor Swift
Singer-songwriter · taylorswift.com

Taylor Alison Swift is an American singer-songwriter and actress. Raised in Wyomissing, Pennsylvania, she moved to Nashville, Tennessee, at the age of 14 to pursue a career in country music. Wikipedia

Born: December 13, 1989 (age 25), Reading, Pennsylvania, United States

Height: 1.78 m

Parents: Scott Kingsley Swift, Andrea Swift

Songs

Blank Space	2014	1989
Shake It Off	2014	1989
Bad Blood	2014	1989
I Knew You Were Tr...	2012	Red
22	2012	Red

Profiles

Albums View 10+ more

1989 2014
Red 2012
Speak Now 2010
Taylor Swift 2006
Fearless 2008

More about Taylor Swift

Figure 1.2: Search page assembled and returned for the query *"taylor swift"* by the Google web search engine, displaying access to vertical search engines (top, below query box), aggregated search results merged within the ranked list of results, and entity card. Based on our definition, in this example, the entity card displayed on the right of the results listing is a type of composite object. Captured in March, 2016.

confuse certain segments of the user base — generated a certain tension underlying web search development. With more people creating, hosting and searching for diverse content on the web, even the inflexible, *"nearly identical"*, web search interface had to change to expose this diversity.

Modern web search engines address this tension by offering several solutions: ***(i)*** they provide specialised services to access diverse content (e.g., *image*, *news* or *map* search, often referred to as *vertical search services* or just *verticals*) available within a tabbed interface, on the main page of results; ***(ii)*** they merge blocks of documents from different sources (i.e., verticals) into a unified results page (a paradigm that has become known in the research literature as *aggregated search*) ***(iii)*** they integrate complex information aggregates (also known as *entity cards* or *knowledge graph results*) containing items extracted from different sources, on the results page. Figure 1.2 shows all three elements that give searchers direct access to diverse content on a unified results page. It seems, then, at least in certain cases or for certain queries, that the search interface has changed significantly from the ranked list of blue links, a change driven primarily by the need to accommodate diverse content within a unified results page.

Diversity in unity

Do all results pages need to display heterogeneous content? Perhaps not: for instance, Bing report[5] that roughly 30% of the queries it receives are navigational — meaning a searcher is trying to find a particular website like Facebook or BBC News, in which case, it is likely that a single result is of interest to the searcher. However, it remains that a majority of queries have informational or transactional intent, in which case presenting heterogeneous content on the results page might be useful. In fact, Arguello et al. (2009) indicate that 74% of queries (labelled by reviewers) require results from at least one other source in addition to general web search, suggesting that merging heterogeneous results into a unified page is an important aspect of improving search experience for a majority of queries.

The two methods of merging diverse content within a unified results page, aggregated search and entity cards, are relatively new additions to web search interfaces. Although aggregated search has been available since the early 2000s (Murdock and Lalmas, 2008), it is, perhaps, Google that introduced it into mainstream web search in 2007, under the name of *universal search*:

> "Google's vision for universal search is to ultimately search across all its content sources, compare and rank all the information in real time, and deliver a single, integrated set of search results that offers users precisely what they are looking for. Beginning today, the company will incorporate information from a variety of previously separate sources — including videos, images, news, maps, books, and websites — into a single set of results. At first, universal search results may be subtle. Over time users will recognize additional types of content integrated into their search results as the company advances toward delivering a truly comprehensive search experience." [6]

[5] Accessed December 2017: https://blogs.bing.com/search/2011/02/10/making-search-yours

[6] Accessed December 2017: https://googlepress.blogspot.com/2007/05/google-begins-move-to-universal-search_16.html

Why is there a need to integrate different types of documents available on the web directly into a unified results page, when isolated access to specific resources is available? The argument for a more diverse presentation of results can be made from multiple points of view — Arguello (2017) details this argument in-depth. A results page merging documents from different sources can: ***(i)*** show searchers that relevant content exists in particular verticals; ***(ii)*** provide searchers with easy access to different sources of information; ***(iii)*** bring attention to diversity in the information space, which can benefit searchers in understanding their information need, in current or future searches; and ***(iv)*** show search results in a diverse context (e.g., images next to the news article they were extracted from, on the same results page) which can help searchers assess their relevance. Perhaps the strongest argument for aggregated search is made by existing search engines (Google, Bing, Yandex, DuckDuckGo, Seznam, Naver, Baidu and others) which, over the past decade, have all adopted it as a way of merging and presenting diverse content on a unified results page.

Most aggregated search systems are based on a workflow that broadly involves two main tasks: *vertical selection* and *vertical presentation*. Vertical selection is concerned with predicting which sources of information are relevant to a given query, whereas vertical presentation is concerned with predicting where, merged in the ranked list of results, to place blocks of heterogeneous content. Typically, the results contained within each heterogeneous block are simply the top few results returned by their corresponding vertical in response to the searcher's query.

Even though widely adopted by modern search engines, aggregated search is limited in several aspects. To date, prior work has not investigated in detail which results from a particular vertical, rather than just the top few, to display on the results page (Arguello, 2017). This is an important aspect of merging heterogeneous content within a unified page as prior studies have shown that results from one source can influence user engagement with results from other sources (Arguello and Capra, 2016). As such, understanding the interactions between heterogeneous items can be informative for selecting which result to extract from verticals, rather than returning the top few. Secondly, limited effort has been put into understanding methods for the *display* of heterogeneous content on the results page (Arguello and Capra (2014) is one such effort). It remains unclear, for instance, why, post-retrieval, heterogeneous content needs to be aggregated by type (e.g., a block of image results or a block of video results), rather than topic or other features, or structured in any other way. Indeed, work on aggregated search frequently mentions presentation aspects as core directions for future work (Arguello, 2017), where additional research effort is required.

The other method of merging heterogeneous content within a unified results page — through the use of entity cards — addresses some of the limitations of aggregated search. Introduced in mainstream web search in 2012[7], entity cards bring together content extracted from various sources within a singular object displayed on the results page. Besides providing the benefits of aggregated search (see above), entity cards also provide a summary of relevant content directly on the results page, they help users navigate topically diverse results by highlighting one or multiple facets reflected in the results page, and they support exploratory search by highlighting entities related to a given query. Unlike aggregated search, entity cards merge together results of *different* types, rather than a single type, structured around a common topic and are typically displayed contextually, not integrated in the ranked list of web results[8]. Figure 1.2 shows an entity card incorporating images, Wikipedia facts, social media profiles and other types of results within one object displayed to the right of the ranked web results.

Much like aggregated search, most popular web search engines today have adopted the display of entity cards. However, prior research effort regarding entity cards has focused primarily on the underlying representation mechanism (i.e., knowledge graphs or knowledge bases) rather than understanding what types of results searchers expect or how heterogeneous results interact within such objects, or how entity cards influence searcher behaviour. Even more, it is unclear whether entity cards are the only type of complex result aggregate that can be effectively merged within the results page. Our work aims to address these understudied aspects of merging heterogeneous content within a unified results page by studying the *"composition"* of results retrieved from various sources. The definition of result composition, and how it relates to existing areas of heterogeneous web search, is discussed in the following section.

1.2 Definition

Our work investigates the composition of web search results retrieved from various document sources (e.g., *images, tweets, Wiki pages*) within singular objects to be integrated in the results page assembled in response to user queries. In our definition, composite objects contain complementary web search results that together achieve a common informational goal (i.e., address a specific aspect of a

[7] Accessed December 2017: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

[8] In desktop search.

user's query). We use the terms *"composite objects"* or *"composite results"*[9] to refer to such objects, and we refer to the process through which they are constructed, either by searchers or algorithms, as *"result composition"*.[10]

We use the terminology of *Composite Web Search* to distinguish the approach discussed in this thesis from *Aggregated Web Search*. We emphasise that although both approaches are concerned with merging heterogeneous results within a unified results page, they differ with respect to the way they organise and expose diverse content on the results page within *composite objects* or *aggregated search blocks* respectively. Specifically, the aspects differentiating composite objects from aggregated search blocks are that composite objects ***(i)*** contain results from multiple sources, ***(ii)*** are specific to a common topic or facet of a topic, and ***(iii)*** are not a uniform ranking of results ordered only by their topical relevance to a query. In contrast to composite objects, aggregated search blocks contain results from a single source of information (e.g., a block of image results), which are not organised around the various facets of a topic, and are typically a ranking of items ordered by their topical relevance, without considering item complementarity. As an example, in on our definition, an entity card is an instance of a composite object, because its constituent items originate from different sources, are focused on a common topic or entity, and are organised to provide complementary information that addresses a common search goal.

We use similar terminology as aggregated search when referring to specialised search services for different types of media (e.g., images, videos, news) or search tasks (e.g., search for local business, scientific articles), namely the term *verticals*. We use *heterogeneous* or *different* sources of information as synonyms for verticals, to highlight the origin of diverse results, in contrast with general web search results. In this context — and throughout this thesis, unless explicitly mentioned otherwise — the use of the term *diversity* refers to vertical diversity rather than the explicit topical diversity of web search results (Santos, 2012).

Given this definition, we clarify the statement of this thesis next.

1.3 Thesis Statement

Integrating heterogeneous documents into a unified results page is difficult. Aggregated search — the current paradigm for selecting and displaying results from

[9] We also use *complex results aggregate* or *complex aggregates* to refer to composite objects, without additional meaning, primarily to avoid excessive repetition.

[10] We also use *aggregation of heterogeneous content within composite objects* as a synonym for *result composition*.

different sources on a singular page — is limited to block-level merging of rankings. While more complex aggregates, containing and highlighting the connections between heterogeneous documents (i.e., *composite results*) can be constructed and displayed effectively on a unified results page, their usefulness is constrained by ***(a)*** the relevance of a document (or group of documents) that play a central role within the aggregate and ***(b)*** a complex interplay between the overall properties (i.e., relevance, diversity and coherence) of the aggregate. This statement will be defended through work which seeks to answer the following groups of research questions:

- Given a search task, how do users construct composite objects containing results extracted from different source? What properties of these manually constructed objects are important to searchers? What roles do individual documents play within composite objects constructed by searchers?

- What effect does the display of composite objects on a unified results page have on search behaviour? How do the properties of composite objects moderate this effect?

- When constructing composite objects algorithmically, how can object properties (e.g., topical diversity or vertical diversity) be manipulated? What roles do individual documents play within composite objects? How do manipulations of composite object properties interact with object effectiveness?

- Given that documents extracted from different sources have varying structures (e.g., videos vs. text-based documents), can vertical-agnostic document representations be learnt within a common feature space which enable more effective assessments of document similarity across verticals?

These research questions attempt to investigate two key aspects of creating complex aggregates containing documents from different sources: the interplay between search interactions and composite object properties, and how algorithmic formulations of composite objects can effectively operationalise object properties.

Answers to these research questions will provide practitioners of web search with a foundation on which to study and develop the integration of complex aggregate structures info unified result pages.

1.4 Motivation

The World Wide Web has been a fertile research area ever since its creation — and so has searching the web. Topics ranging from search interface design to user interactions with web search systems or the implementation and evaluation of systems specialised for particular collections have been studied extensively over the past decades (Sanderson and Croft, 2012). In light of these comprehensive research efforts aimed at understanding web search, and recognising that commercial businesses define almost all aspects of what web search *is* today, several questions need to be answered before developing our research topic further:

1. Why study web search more?
2. What could users *"possibly need besides Google to search the web"*? – Wilson et al. (2010)
3. Why is the composition of heterogeneous web results an important topic?

Answers to these questions motivate the research efforts presented here and provide some insight as to why understanding complex result aggregates can be valuable for the future of web search.

Why study web search more?

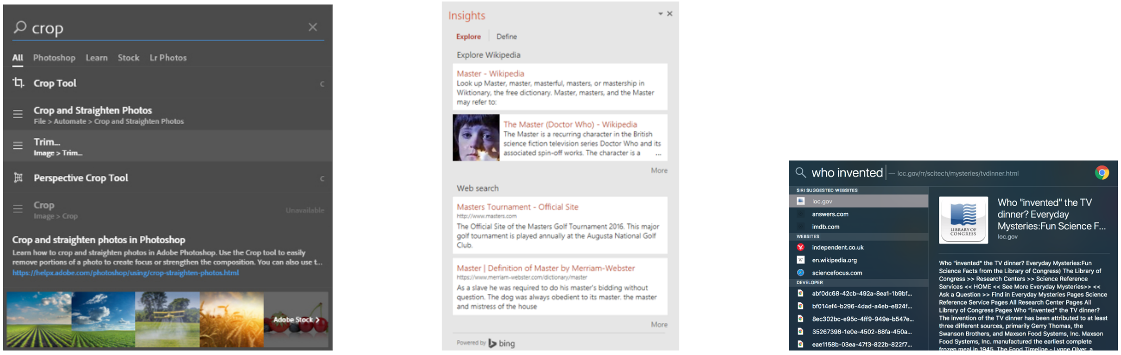
Although the question may seem trivial, given that keeping web search technology synchronised with a ubiquitous, dynamic and evolving web undoubtedly requires research effort, it is important to briefly clarify how extensively the science and technology behind web search impacts our world. In a 2011 attempt at quantifying the value of web search technology on the global economy, McKinsey & Company report[11] that:

> "The size of search can be hard to conceive. More than one trillion unique, worldwide URLs were indexed by Google alone by 2010. Some 90 percent of online users use search engines, and search represents 10 percent of the time spent by individuals on the web, totaling about four hours per month. Knowledge workers in enterprises spend on average five hours per week, or 12 percent of their time, searching for content. [...] A conservative estimate of the global gross value created by search was $780 billion in 2009."

Without dwelling on this particular point, it is enough to point out that improvements brought to the technology that drives web search have extremely far reaching effects both in terms of the number of people they affect, and the value they generate, thus justifying the further study of web search, in all its forms.

[11] Accessed December 2017: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/measuring-the-value-of-search

In addition to the extensive impact of web search in its current, browser-based form, there are many emerging directions in the development of search that are understudied and which motivate further research in the field of web search. One of the emerging directions is the inclusion of web search functionality directly within specialised applications, such as Microsoft Office, Adobe Creative Suite or Windows and macOS. Figure 1.3 shows several examples of such modern in-application search experiences. These examples illustrate how web search is expanding beyond its traditional browser-bound environment and becoming embedded within various areas of human-computer interaction. These novel and understudied directions exemplify some of the many forms web search can take. These examples also highlight the interplay between accessing (e.g., finding a document online) and creating information (e.g., writing an article within a text editor) and how web search is permeating this boundary explicitly — for example, in productivity software with embedded web search functionality. It is, perhaps, at this boundary between retrieval and creation or assembly of information that understanding the relationship between different types of documents is important, and where our work may prove to be insightful as well.



Crop and straighten photos in Photoshop

Learn how to crop and straighten photos in Adobe Photoshop. Use the Crop tool to easily remove portions of a photo to create focus or strengthen the composition. You can also use t...

https://helpx.adobe.com/photoshop/using/crop-straighten-photos.html

Master - Wikipedia

Look up Master, master, masterful, masters, or mastership in Wiktionary, the free dictionary. Master, masters, and the Master may refer to:

The Master (Doctor Who) - Wikipedia

The Master is a recurring character in the British science fiction television series Doctor Who and its associated spin-off works. The character is a ...

Masters Tournament - Official Site

http://www.masters.com

The Official Site of the Masters Golf Tournament 2016. This major golf tournament is played annually at the Augusta National Golf Club.

Master | Definition of Master by Merriam-Webster

https://www.merriam-webster.com/dictionary/master

As a slave he was required to do his master's bidding without question. The dog was always obedient to its master. the master and mistress of the house

who invented — loc.gov/rr/scitech/mysteries/tvdinner.html
SIRI SUGGESTED WEBSITES
loc.gov
answers.com
imdb.com
WEBSITES
independent.co.uk
en.wikipedia.org
sciencefocus.com
DEVELOPER
abf0dc68-42cb-492a-8ea1-1b9bf...
bf014ef4-b296-4dad-a4eb-e824f...
8ec302bc-e95c-4ff9-949e-b547e...
35267398-1e0e-4502-88fa-450a...
eae1158b-03ea-47f3-822b-822f7...
LIBRARY OF CONGRESS
loc.gov

Who "invented" the TV dinner? Everyday Mysteries:Fun Science F...

Who "invented" the TV dinner? Everyday Mysteries:Fun Science Facts from the Library of Congress) The Library of Congress >> Research Centers >> Science Reference Services << HOME << See More Everyday Mysteries>> << Ask a Question >> Find in Everyday Mysteries Pages Science Reference Service Pages All Research Center Pages All Library of Congress Pages Who “invented” the TV dinner? The invention of the TV dinner has been attributed to at least three different sources, primarily Gerry Thomas, the Swanson Brothers, and Maxson Food Systems, Inc. Maxson Food Systems, Inc. manufactured the earliest complete frozen meal in 1945. The Food Timeline - Lynne Olver, a

(a) Web search functionality integrated within the image editing software Adobe Photoshop.

(b) Web search functionality available within the Microsoft Office application suite.

(c) Web search results merged within the Spotlight search functionality available on macOS systems.

Figure 1.3: Several examples illustrating how web search is expanding beyond its traditional browser-bound environment and becoming embedded within various other areas of human-computer interaction.

To summarise and answer our question, on one hand, web search has extremely far reaching impact on our world and as such, any improvement to the technology that drives it generates enormous value; on the other, web search, in its *query to results-list* interaction style, is developing into a foundational layer upon which other technology interactions are constructed, and which underlies many aspects of our technology-mediated access to information. Both views indicate the importance of web search and motivate its further study.

What else besides Google?

This question is concerned with two critical issues that face any researcher involved in the study of web search: ***(a)*** Given its adoption and overall success, why is Google not good enough as it is? ***(b)*** Why is the study of web search a matter for the public domain[12], and not only specific to commercial search engines? Wilson et al. (2010) tackle the first issue head on in their monograph on future search user interfaces for the web:

> “ The elegant way in which search results are returned has been well researched and is usually remarkably effective. [...] Google is really good. For what is does. [...] Search as embodied by the text box and keyword has framed our understanding of what the web is (schraefel, 2009). It turns out that this elegant paradigm is especially effective for 1-minute search [...]. But many users have come to the web for substantive research that takes hours or weeks – find all songs written about poverty in the 1930s, prove that there are no patents that cover my emerging innovation, or locate the best legal precedents for my brief.
>
> A second motivator for [research on] new search strategies is that the next generation web will offer fresh possibilities that go well beyond finding documents by way of keyword search. Hall and O’Hara (2009) stress that what we know as the web today, is the Document web, and not the web of Linked data that is imminently upon us. ”

There are many ways of searching the web and commercial web search engines today are optimised for a certain type of search that involves a query box, very few query terms (on average) and a results list — something they are, indeed, extremely good at. It is with respect to different, and perhaps less studied, types of search interactions (e.g., slow search (Teevan et al., 2013), exploratory search (Marchionini, 2006; White and Roth, 2009), serendipitous search (André et al., 2009; Bordino et al., 2013; Rahman and Wilson, 2015)) that searchers may need something that most modern search engines simply do not offer yet.[13] Indeed, ima-

[12] Sponsored through public funds and conducted at public institutions.

[13] Unlike the previous section, where new areas of web search advancement refers to the application of keyword-driven search in new contexts, here novel search interactions refers to fundamentally different approaches to extracting and interacting with information from the web.

gining a future for these novel types of web search requires research effort that is unconstrained by commercial interest.

Secondly, the main reason for studying web search in the public domain, even though it is almost entirely defined by commercial efforts, is that "web giants [Google, Facebook and others] [...] are the closest thing we have to information utilities" (O'Neil, 2016, p. 211). As mentioned previously, web search underlies many, if not most, aspects of our technology-mediated interactions with information. In light of recent discussions on algorithmic fairness (Corbett-Davies et al., 2017; Feldman et al., 2015; O'Neil, 2016) and excessive personalisation (Pariser, 2011) it has become apparent that biases — intentional or accidental — encoded within algorithms can lead to consequences that are harmful to both individuals and communities (the fake news and filter bubble phenomena are some examples of these consequences). It is, then, a matter of public virtue, if not outright responsibility[14], to scrutinise and understand algorithms, especially those that organise and retrieve documents from the web.

Relatively little effort has been made so far in quantifying algorithmic biases in web search (Mehrotra et al. (2017) is one such example) even though their effects have been discussed widely (O'Neil, 2016; Pariser, 2011), as has the ethics of web search (Introna and Nissenbaum, 2000; Mager, 2018). This is not surprising given that web search, today, is entirely opaque in order to preserve competitive advantage. For example, Google has been known to have "prohibited researchers [interested in investigating] the biases of the search engine" (O'Neil, 2016, *op. cit.*) Together, the issues of algorithmic fairness and mixed commercial incentives in web search suggest that the science and technology that drive these "information utilities" should become more and more a matter for the public domain, in a way that can be scrutinised by those it affects. In the words of the very founders of Google: "(...) the issue of advertising causes enough mixed incentives that it is ***crucial*** to have a competitive search engine that is transparent and in the academic realm" (Brin and Page, 1998*a*, emphasis added).

Returning to our research topic, why should, then, the study of result composition be a matter for the public domain? Assembling results — and, in particular, placing items on the first page of results returned by a search engine — is becoming as much about merging heterogeneous items (or groups of items) within a unified space, as it is about ranking and retrieval from homogeneous collections. Web results, aggregated search blocks, entity cards (Bota et al., 2016; Hasibi et al., 2017; Navalpakkam et al., 2013), rich format ads (Lagun et al., 2016) or in-

[14] Such matters are becoming national policy issues in the European Union and the United States.

line answers (Chilton and Teevan, 2011) are all competing for visual real-estate in what is shown to users. Often results pages contain enough information, assembled from various sources, that searchers have no need to explore results in detail and can access all the information they need directly on the first page of results. Merging items into a singular page is just as critical as ranking, and all layers of search, from document representation to retrieval, ranking and aggregation, carry influence on what information users interact with; Ford and Graham (2016) provide an interesting discussion on the social implications of complex information objects embedded within the search engine results page.

The work presented in this thesis is concerned with the aggregation of results which, much like the entire machinery of web search, should be public. In addition, our work looks explicitly at the effects of displaying composite results within a traditional ranked-list setting on search behaviour, and as such, can be informative to all those who use modern web search.

Why study result composition?

As mentioned in the previous section, web search engines now deploy a variety of "non-traditional" items on the results page. Assembled around the long-established ranked list of results are aggregated search blocks, rich format ads, in-line answers or entity cards. Indeed, the results page is often an assembly of diverse items extracted from various sources.

The composition of heterogeneous results into unified items displayed on the search page — e.g., in-line answers or entity cards — has become an indispensable feature of modern web search interfaces. Chilton and Teevan (2011) show that more than one hundred different types of in-line answers are already being returned to users of Bing (although the authors report that, at the time, such results were shown in less than 1% of sampled searches). Bota et al. (2016) indicate that entity cards are returned for entity oriented queries 67% of the time, which shows how prevalent these objects have already become. In a less formal setting, Enge (2017) suggest that rich answers (i.e., in-line answers or entity cards) are now shown on more than 50% of the result pages returned by Google. Overall, presenting diverse results on a unified page is important to users, and due to the limitations of merging blocks of results from different sources within a ranked list of web documents, search engines are increasingly making use of complex result aggregates to expose and help users navigate diverse results. As such, understanding how users interact with and consume heterogeneous content, how complex result aggregates (e.g., entity cards or other types of composite results) can be created effectively, and how the properties of these aggregates influence user search behaviour is a major aspect of modern web search.

The content and properties of complex aggregates (e.g., entity cards, in-line answers) are even more important considering that users frequently satisfy their information need directly on the results page, rather than through the exploration of any result in particular — a phenomenon called *good abandonment* (Chuklin and Serdyukov, 2012) in the research literature. Williams et al. (2016*a*) indicate the presence of complex aggregates (entity cards and in-line answers) on the results page as one of the factors that drive good abandonment in web search[15].

Overall, the study of composite objects is important because, on one hand, these types of objects are being deployed more and more by current web search engines, and on the other, because their presence on results pages leads to atypical user behaviours — e.g., atypical eye gaze patterns (Navalpakkam et al., 2013) or good abandonment — which are insufficiently understood yet critical in advancing modern web search. What information do users expect to find within these objects? How can these objects be described in terms of their properties? How do their properties influence user behaviour? How can results be aggregated within such objects? These are relatively understudied questions that lie at the core of constructing complex result aggregates, and therefore central to modern web search, that our work aims to answer.

Motivating scenario for result composition

To discuss in more detail the practical benefits of result composition, we revisit the example shown in figure 1.2, illustrating Google's response to the query *"taylor swift"*. Based on our definition, in this example, the entity card displayed on the right of the results listing is a type of composite object. The entity card assists searchers by providing a summary of relevant and diverse content directly on the results page, but also by highlighting potentially interesting facts regarding its topic of focus, and by providing easy access to various other topics (i.e., related entities) or related sources of additional information (in this example, social media profiles). As with aggregated search blocks, entity cards provide an overview of diversity in the information space. However, unlike aggregated search blocks, entity cards provide access to diverse content in a structured manner, with diverse and complementary items being organised around a shared topic, and the overall object addressing a common informational goal.

In our example, imagine a searcher familiar with one of the more popular

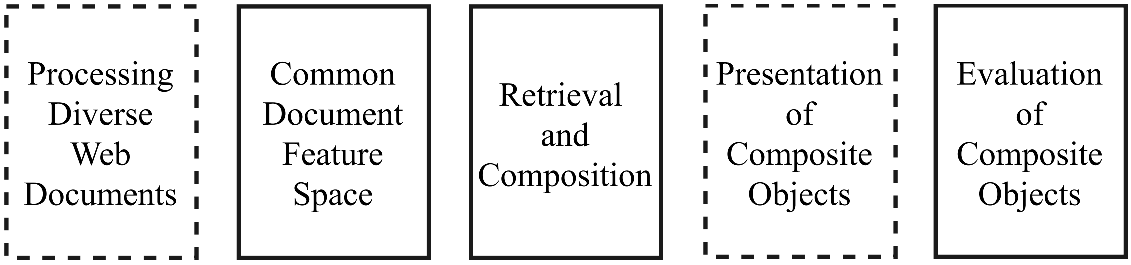
[15] The study referenced here focuses exclusively on mobile web search.

songs by Taylor Swift, but unaware of her biography, other songs or albums, or other artists that create similar music. By issuing an ambiguous query to a search engine, the searcher is perhaps expressing an exploratory information need (Marchionini, 2006), characterised by a desire to learn and discover more about their expressed topic. In such a scenario, integrating an entity card on the results page can help searchers not only resolve their exploratory information needs without having to click on specific results, by providing a summary of most relevant images and biographical facts, as well as a list of popular songs in a visually salient manner directly on the results page, but can also help them make sense of highly diverse (both topically and vertically) results. In addition, in exploratory search conditions, where searchers are not sure how to initiate a search or parse results returned by a search engine, it is perhaps reasonable to assume that entity cards organise and simplify results in such a way that aids searchers in refining their queries and comprehending the entirety of the results page — i.e., they enhance users' search literacy (Wilson et al., 2016). Even though aggregated search blocks, like composite objects, expose diverse content on a unified results page, in contrast to entity cards, they do not organise this information. For instance, in our example, both the news and the social media aggregated search blocks are merged within the page. However, our searcher might find it difficult to understand how individual news items and tweets, for example, are related to each other on such a page or how they provide complementary information on their search topic, given that aggregated search results are merged within the ranking at block level, without highlighting the relationships or complementarity between the items they contain, across blocks or in relation to the result ranking.

In addition to entity cards, other types of composite objects could be useful in this scenario as well. In response to our user's query (e.g., *"taylor"*), it is not difficult to imagine a search system returning results that include both song videos and song lyrics, for example, or results that include both Taylor Swifts' Instagram[16] photos and her most recent tweets, perhaps displayed in a style similar to how in-line answers (Chilton and Teevan, 2011) are used in modern web search interfaces today. These types of results are, in our definition, composite objects and, as such, the work discussed throughout this thesis could be informative for the design and evaluation of such search systems. Indeed, given the heterogeneous nature of modern web documents and the overall success of entity cards, it is likely that web search engines will make use of various types of composite results more and more in the future.

[16] www.instagram.com

1.5 Contributions


Processing Diverse Web Documents
Common Document Feature Space
Retrieval and Composition
Presentation of Composite Objects
Evaluation of Composite Objects

The figure above shows a high-level, approximate representation of the result composition process, from a system-centric perspective. The components of this process that are explicitly addressed by our work are displayed with a solid contour. Within this representation, the main contributions of this thesis are:

1. An exploration of user-constructed composite objects. We approach the result composition problem from a user-centric perspective, by attempting to understand how users interact with heterogeneous content in a search scenario, and how they manually construct composite objects that satisfy their information needs. Our work represents a first study of users' perspectives on result composition for web search. We show that composite objects constructed manually by users tend to contain a central document or set of documents which are perceived as more relevant and reflect the topic of the composite object. We also show that item diversity within composite objects is important to users, but that a hierarchy of object properties is not obvious. Our contribution addresses the evaluation component of the result composition process.

2. An analysis of composite object influence on user search behaviour. Composite objects are already being used extensively by modern web search engines (e.g., entity cards). We conduct a large-scale study to investigate how different manipulations of entity card properties influence users' behaviour in a traditional web search environment. Our work represents a first study investigating the interplay between composite object properties, search behaviour and user perceived workload. We report that composite objects influence both searcher behaviour and perceived workload, and analyse the way composite object properties mediate this effect. Our analysis provides a complementary perspective on the evaluation component of the result composition process.

3. An algorithmic framework for constructing composite objects under constraints. We adapt a general constraint clustering framework to the task of constructing composite objects from heterogeneous web search results. Our evaluation results demonstrate performance improvement over aggregated search or federated search when structuring the results page within ranked composite objects. Our contribution is an algorithmic framework that can be used to construct composite objects, from highly heterogeneous results, while maintaining constraints on composite object properties. Our contribution addresses the retrieval and composition components of the process described above.

4. A method for representing heterogeneous documents within a unified feature space. We investigate the application of graph-learnt document representations to the retrieval and aggregation of heterogeneous content. Our study provides an in-depth analysis of click-graph structures with respect to the distribution of heterogeneous content within the graph, and proposes methods of manipulating click graph structure in order to limit biases introduced by the skewed distribution of heterogeneous content across sub-graphs. Bridging the cross-vertical gap is a difficult problem, and our work is a first analysis of how graph-learnt representations can be used to effectively bridge this gap and provide a unified representation space for the construction of composite objects.

1.6 Publications

Most of the material presented here has appeared in several conference papers published during the course of this programme:

- Bota, Zhou, Jose and Lalmas (2014), Composite Retrieval of Heterogeneous Web Search, in Proceedings of the International Conference on World Wide Web, ACM, New York, NY, USA, pp. 119 – 130. This paper is at the core of chapter 6.

- Bota, Zhou and Jose (2015), Exploring Composite Retrieval from the Users' Perspective, in A. Hanbury, G. Kazai, A. Rauber and N. Fuhr, eds, Advances in Information Retrieval, Springer International Publishing, Cham, Switzerland, pp. 13 – 24. This paper is at the core of chapter 4. ***Best paper award at BCS ECIR 2015***.

- Bota (2015), Heterogeneous Information Access Through Result Composition, in Proceedings of the Symposium on Future Directions in Information

Access, BCS Learning & Development, Swindon, UK, pp. 20 – 24.

- Bota, Zhou and Jose (2016), Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload, in Proceedings of ACM CHIIR, ACM, New York, NY, USA, pp. 131 – 140. This paper is at the core of chapter 5. ***Best paper award at ACM CHIIR 2016***.

- Bota (2016), Nonlinear Composite Search Results, in Proceedings of ACM CHIIR, ACM, New York, NY, USA, pp. 345 – 347.

1.7 Thesis Structure

This thesis is structured into eight chapters:

- Chapters 1, 2 and 3 introduce the subject of our research, and review the theoretical (chapter 2) and applied (chapter 3) contexts in which our work is placed.

- Chapters 4 and 5 are conceptually related, and discuss our user-centric contributions to the study of result composition.

- Moving from a user-centric perspective towards a system-centric one, chapters 6 and 7 discuss our algorithmic contributions with respect to result composition on the web.

- Chapter 8 provides an in-depth discussion of our research outcomes as well as presents directions for future work in the space of result composition.

- Finally, chapter 9 ends this thesis by summarising the key contributions and conclusions derived from of our work.

Chapter 2

Background

In this chapter, we discuss the background to many of the concepts used throughout this thesis, focusing primarily on the theoretical context, rather than application context, within which our work is based. Chapter 3 complements this chapter by providing a detailed background on the application context (i.e., web search) of our work. Prior research has defined much of the conceptual framework and vocabulary used when discussing *search*. As such, we begin this chapter by attempting to explain what *search* is, from a user-centric perspective, and how it connects to related concepts, such as *information retrieval* or *information objects*.

2.1 User Models of Search

Search is an ambiguous term: searching for information on Fidel Castro for a school project, searching for an address on the street, searching for nutritional information on a product label or for a better price in the supermarket, searching for words to express an idea — all these are instances of searching for information, that may or may not make use of search engine technology. In a holistic perspective on searching for information, Kekäläinen and Järvelin (2002) introduce a model of search activity, later extended in Järvelin and Ingwersen (2004) and examined in Ingwersen and Järvelin (2005), that considers the multiple contexts in which searching for information occurs, taking into account the various personal, organisational and even cultural constraints that affect search behaviour.

In their model, searching for information can be viewed as a hierarchy of tasks, goals and processes, with their associated contexts, where each level of the hierarchy defines the context of its immediate subordinate. Figure 2.1 makes this hierarchy of tasks and goals explicit. At the top level of their *search* hierarchy, the socio-organisational and cultural context defines broad requirements

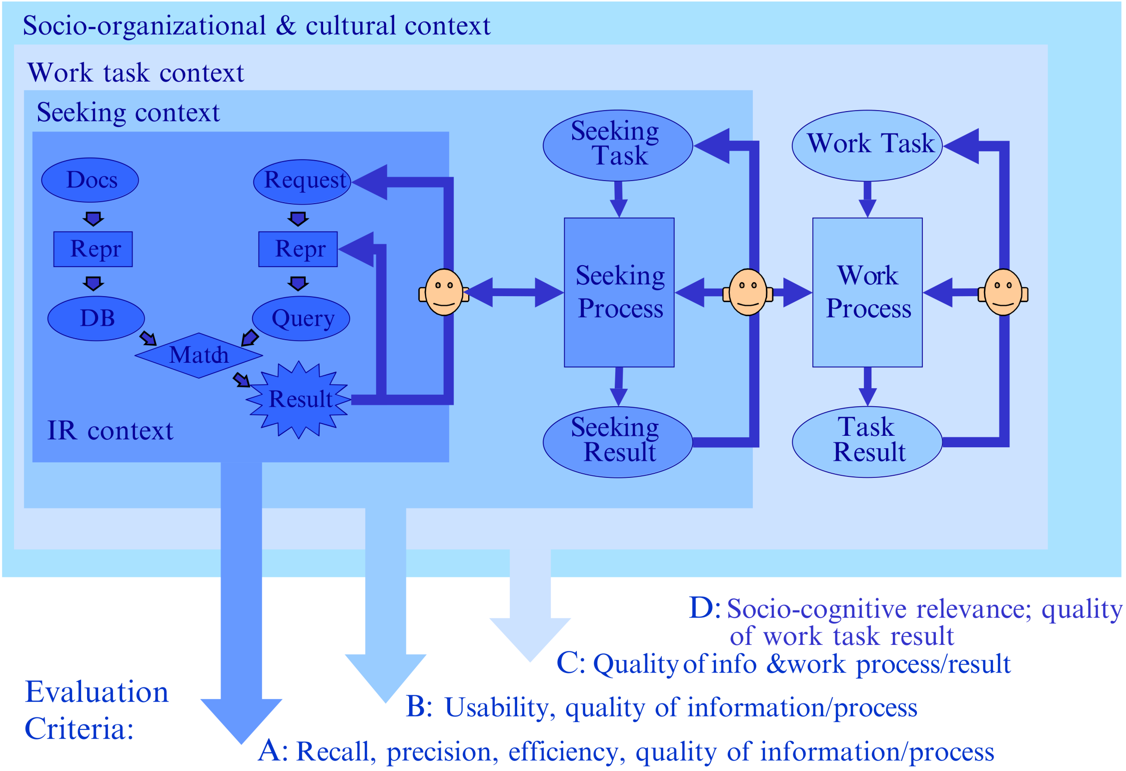

Socio-organizational & cultural context
Work task context
Seeking context
Docs
Request
Repr
Repr
DB
Query
Match
Result
IR context
Seeking Task
Seeking Process
Seeking Result
Work Task
Work Process
Task Result
Evaluation Criteria:
D: Socio-cognitive relevance; quality of work task result
C: Quality of info &work process/result
B: Usability, quality of information/process
A: Recall, precision, efficiency, quality of information/process

Figure 2.1: Nested contexts and evaluation criteria for information seeking and information retrieval, extended from Kekäläinen and Järvelin (2002) in Järvelin and Ingwersen (2004). Reproduced from Ingwersen and Järvelin (2005, p. 322).

and constraints on searching for information. For example, accessing and interacting with information in a safety critical environment, such as a hospital, compared to a library, has different (broad) implications on how people search for information. Nested within the socio-organisational context, there is a work context that defines more specific aspects related to searching for information (e.g., finding information about a general research topic, under certain long-term constraints such as time or budget). Within the work context, search goals are addressed through information seeking strategies, such as analysing and comparing different pieces of information (e.g., finding and comparing ways to structure a background chapter). Finally, as the smallest component of the search hierarchy, information retrieval is most often embodied by keyword search, the activity through which searchers are trying to access a very specific, and often known[1], piece of information (e.g., how to change reference style in LaTeX). The work presented in this thesis relates to the *information retrieval* and *information seeking* aspects of searching for information, and as such, we review these aspects in more detail next.

[1] What information is needed is known, rather than the actual piece of information.

2.1.1 Information Retrieval

Ingwersen and Järvelin (2005, p. 21) define the information retrieval component of the search model as "[t]he processes involved in the representation, storage, searching, finding, filtering and presentation of potential information perceived relevant to a requirement of information desired by a human user in context". Wilson et al. (2010, p. 16) clarify:

> "Within the information retrieval (IR) context, the searcher's goal is focused on finding documents, document sub-elements, summaries, or surrogates that are relevant to a query. This may be an iterative process, with human feedback, but it usually is limited to a single session. Typical IR tasks involve finding documents with terms that match terms presented by the searcher, or finding relevant facts or resources related to a query. Typically, within each IR task, the searcher formulates queries, examines results, and selects individual documents to view. As a result of examining search results and viewing documents, searchers gather information to help satisfy their immediate information-seeking problem and eventually the higher-level information need. The common element of all IR tasks as defined [...] is the query-result-evaluation cycle conducted over a collection with the "unit" of information being the document, a sub-element of the document, a summary, or a document surrogate."

In the user search model describe above, searching for web pages that contain certain terms (e.g., issuing the query *"brexit"* on Google) is an example information retrieval task. Modern web search engines are very good at supporting these types of search tasks — and are slowly becoming better at supporting other types of tasks through in-line answers (Chilton and Teevan, 2011) or widgets embedded directly in the results page (i.e., search with transactional intent (Broder, 2002; Rose and Levinson, 2004), supported through various results page widgets, such as conversion tools, metronome, timer and other types of search assistance).

2.1.2 Information Seeking

Within the search model introduced above, information seeking is the task of satisfying a perceived *"information need"* (Shneiderman, 1997). This task typically assumes the initial recognition and specification of an information need, followed by the examination of search results and iteration through the query-result-evaluation cycle until a satisfactory outcome is achieved (i.e., a result is found or the information need is satisfied) (Marchionini and White, 2007; Shneiderman, 1997; White, 2016). Although enhanced by modern search technology, which extensively support information retrieval tasks, information seeking is common even outside technology-mediated information access, and is typically viewed a fundamental human activity (Marchionini, 1995) that is a component of planned behaviour (e.g., finding directions), decision making (e.g., finding the best car insurance), and the creation of new information (e.g., writing a thesis).

At the information seeking level, searchers define and employ strategies about where, how and whether to find information that addresses their information need (Wilson et al., 2010). They can choose to make use of an information retrieval system (e.g., Google), browse webpages or follow a trail of information (White and Huang, 2010; White and Ruthven, 2006). Even more, they can consult printed materials, ask friends or call information services. Information accumulated through one or multiple information retrieval tasks, as part of the overall information seeking task, will then be examined and synthesised into a solution to the information need (Wilson et al., 2010). Search as the interplay between two main components, information retrieval and browsing on one hand, and the analysis and synthesis of results on the other, is often referred to as *sensemaking* in the research literature (Pirolli and Card, 2005; Russell et al., 1993).

Web search engines, as provided by popular commercial entities, make available much of the infrastructure needed to perform extensive information retrieval tasks, but, historically, have offered limited support for broader information seeking efforts. However, modern web search interfaces have begun to integrate various types of search assistance tools that are aimed at providing information seeking support. For example, entity cards or in-line answers, as discussed in the previous chapter, extract and summarise relevant content from various sources in an attempt to address information needs directly on the results page, effectively reducing the query-result-evaluation cycle (or the number of information retrieval tasks in an information seeking strategy). In addition, the commercial web search engine Bing has recently started providing users with side-by-side summaries of controversial or highly debated topics (e.g., is coffee good for you?) directly on the results page[2], in an effort to support decision making directly. Other web search systems that aim to support information seeking – either by exposing search history, by enabling users to tag related documents or by allowing users to aggregate information across results – exist, yet are not widely popular[3].

[2] Accessed March 2018: https://blogs.bing.com/search-quality-insights/february-2018/Toward-a-More-Intelligent-Search-Bing-Multi-Perspective-Answers.

[3] Yippy.com is one such example.

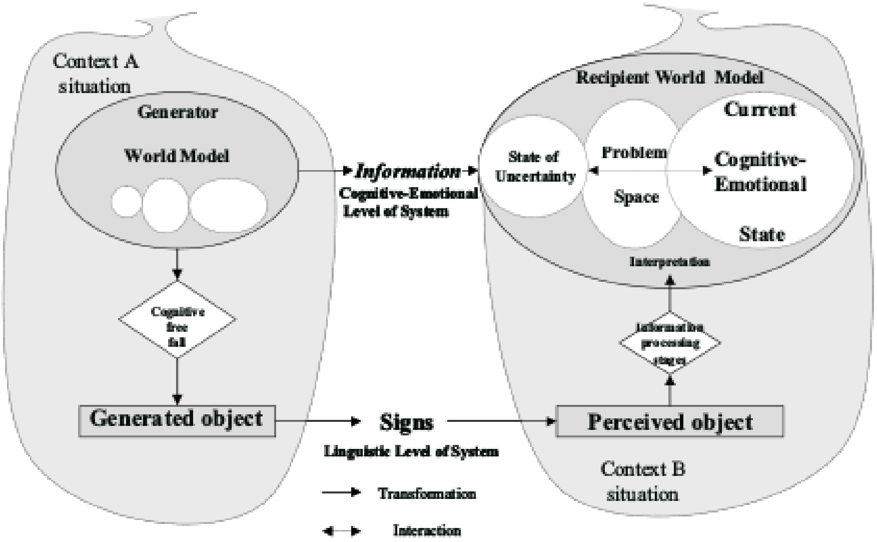


Figure 2.2: The cognitive communication system for information seeking, displaying the exchange of information objects from Generator (information object author) to Recipient (information object seeker). Reproduced from Ingwersen and Järvelin (2005, p. 33).

Although we do not explicitly investigate interactions between result composition and information seeking in our work, result composition can be viewed as an approach to providing information seeking support. By extracting content from various sources of information and assembling it around common topics in relation to a searcher's query, result composition can help users assess and synthesise the diversity of the information space, as well as reduce the number of information retrieval tasks in their overall information seeking approach.

2.1.3 Information Objects

Ingwersen and Järvelin (2005, p. 19) define information objects as physical or digital entities "in a variety of media [,] that belong to the information space of IR [information retrieval] systems, providing potential information" — they use the terms documents and information objects interchangeably, as do other studies in the area of Information Seeking and Retrieval (Belkin, 1978; Belkin and Cool, 2002; Belkin and Robertson, 1976, and related). In Ingwersen and Järvelin (2005), information objects are *authored* by *cognitive actors*. Actors, when generating information objects, are influenced ***(i)*** by the hierarchy of contexts in which they are placed, and ***(ii)*** also by their own understanding (i.e., *world model*), defined through their prior *information seeking* or information interaction behaviour.

The term *information object* is, therefore, deliberately general and recursive (i.e., information objects can contain multiple information objects), being used to refer to results, regardless of presentation style, or documents, regardless of type. This thesis claims to investigate the aggregation of web search results retrieved from various document sources (e.g., images, tweets, Wiki pages) within *information objects*. From a user-centric perspective on search modelling, our use of the term *information object* follows previous usage and definitions, with composite objects being "digital entities, in a variety of media, that belong to the information space of IR systems, providing potential information". Composite objects are also authored by cognitive authors, however through the mediation of algorithms for result composition (i.e., in the case of composite objects, the author is the algorithm designer and developer). As such, in addition to author context and their model of the world, algorithmic biases can also influence information objects derived from result composition.

2.1.4 Relevance versus Usefulness

Relevance is a complex topic in Information Science and Information Retrieval, and has been discussed extensively for more than half a century (Cleverdon, 1967). It is "a, if not the, key notion in information science in general and information retrieval in particular" (Saracevic, 1975) yet it is used in widely dissonant ways in the research literature. It has been systematically reviewed and examined in numerous studies (Borlund, 2003*a*; Mizzaro, 1997; Saracevic, 2007*a*,*b*) concluding that ***(i)*** relevance is a multidimensional cognitive concept, largely dependent on searchers' perception of information and their own information need situation (Schamber et al., 1990), and that ***(ii)*** relevance is a dynamic concept, which can take any meaning (e.g., topical relatedness, user satisfaction), and changes over time, or with users' cognitive state. In some cases, relevance can even be irrelevant (Millan-Cifuentes et al., 2014).

We do not attempt to clarify the complexities of relevance here, but distinguish between the two "types" of relevance used in this text: *topical relevance* and *user relevance* (what we describe as *usefulness*). In our work, an information object is *topically relevant* to a query if it is on the same topic, as assessed by a person (Saracevic, 1996) (i.e., a person determining topical relevance in a formal setting; a relevance assessor). For example, a web page containing a biography of Fidel Castro would be topically relevant to the query "fidel castro", and would also be topically relevant to queries "cuban history" and "cuban politics".

User relevance or *usefulness* takes into account other factors that go into a searcher's judgement of relevance, and has been proposed as a measure of search quality in numerous prior studies (Belkin et al., 2008; Cole et al., 2009; Mao et al., 2016; Yilmaz et al., 2014). Usefulness is typically discussed in contrast with topical relevance and refers to the "highly situational[4], subjective, user-perceived usefulness" (Mao et al., 2016) of an information object. For instance, Yilmaz et al. (2014) show that effort (e.g., the effort of scanning a document to locate relevant pieces of information) plays an important role in user satisfaction and that effort needs to be considered together with topical relevance when the quality of a search systems is evaluated. This further supports prior work that conceptually distinguishes relevance from the utility (or usefulness) of a document to an actual user, suggesting evaluation metrics should be utilitarian in nature (Saracevic, 1975, 2007*a*,*b*). Cole et al. (2009) propose usefulness as a modern criterion for the evaluation of search systems, and clarify that:

> "Usefulness is specifically distinguishable from relevance in several dimensions. Most strikingly, a usefulness judgment can be explicitly related to the perceived contribution of the judged object or process to progress towards satisfying the leading [search-related] goal or a goal on the way. In contrast to relevance, a judgment of usefulness can be made of a result or a process, rather than only to the content of an information object. It also applies to all scales of an interaction. Usefulness can be applied to a specific result, to interaction sub-sequences, and to the session as a whole. Usefulness, then, is more general than relevance, and well-suited to the object of providing a measurement appropriate to the concept of task goal realization."

This thesis investigates the *usefulness* of composite objects. In particular, we explore how composite object usefulness is constrained by documents contained within the object and by the object's overall properties (e.g., object *relevance*, *diversity* or *coherence*). As in Cole et al. (2009), we consider usefulness to be the perceived contribution of an information object (i.e., composite result) towards achieving a search-mediated goal. In our work, this contribution can be topical (e.g., a composite object that is topically relevant or contains items that are topically relevant to a given query), where topical relevance (or topicality) refers to the degree of aboutness between an information object and a user request, as assessed by a person (Saracevic, 1996), or situational (e.g., a composite object that reduces the workload users perceive during their search-related task). In chapter 4, we explore situational assessments of usefulness with respect to composite object structure and content, whereas in chapter 5 we explore how different manipulations of composite object properties affect their usefulness, as reflected in impli-

[4] Usefulness is also referred to as *situational relevance* in Ingwersen and Järvelin (2005).

cit (e.g., number of clicks, mouse hovers) and explicit (e.g., perceived workload) measures of user search engagement. Chapters 6 and 7 investigate algorithmic aspects of result composition and, as such, focus solely on the topical dimension of usefulness.

2.1.5 Other Models of Search

Numerous theoretical models of search have been proposed over the past decades, in addition to the model by Järvelin and Ingwersen (2004) discussed in the previous sections. Ingwersen and Järvelin (2005, c. 5), Hearst (2009, c. 3), and White (2016, c. 4) all provide comprehensive reviews of user search models, from various perspectives. We briefly discuss some of the models here.

Many attempts at modelling the process of searching for information assume an interactive query-result-evaluation cycle. Marchionini and White (2007) describe the search process as consisting of several steps: recognizing a need for information, accepting the challenge to take action to fulfil the need, formulating the problem, expressing the information need in a search system, examining results, reformulating the problem and using results. Sutcliffe and Ennis (1998) describe a similar cycle, consisting of four main activities: problem identification, articulation of needs, query formulation and evaluation of results. These models are typically based on observations of people engaging in search activities.

The model proposed in Järvelin and Ingwersen (2004) can be described as a cognitive model, as it is informed by a broader model of general task performance (Norman, 1988), and considers an inclusive perspective of how people operate in the world, how their context influences them and how mental models are engaged in the search process. Many different cognitive models of search have been discussed in the literature — Ingwersen and Järvelin (2005, c. 5) review cognitive models of search in extensive detail.

Some of the models of searching for information discussed above make the assumption that searchers' information need is static. Informed by observational studies showing that people's information needs change as they interact with search systems, Bates (1989) introduced the *berry-picking* model of search, which suggests that searchers learn throughout the search process, and, as such, their information need and their queries evolve; and that searcher's information needs are not satisfied by a final set of relevant results but by a series of pieces of information found during their search (in contrast to other models which suggest as the main goal of the search process locating a unique set of relevant documents that perfectly match the searcher's information need). After examining search

	Active	*Passive*
Directed	**Searching**: Active searching directed to particular sources to answer specific questions	**Monitoring**: Passive alertness, primed by interest, that enables an individual to notice information of interest.
Undirected	**Browsing**: Active exploration/search without a clear goal, or for only loosely defined objectives.	**Awareness**: Passive, undirected absorption of experiences and learning.

Table 2.1: Types of information searching, as seen in White (2016, p. 98), adapted from Wilson (1997).

behaviour over extended periods of time, Kuhlthau (1991) and Vakkari (2000) identify several stages of the search process (initiation, selection, exploration, formulation, collection and presentation), which do not always proceed linearly, and have not only practical implications on immediate search behaviour, but can also emotionally influence searchers. The intersection between economic theory and information searching has also been explored in a number of recent studies looking at how microeconomic theory can be applied to interactive information retrieval (Azzopardi, 2011*a*, 2017; Azzopardi and Zuccon, 2016). These modelling approaches attempt to both explain and predict fine-grained search behaviour. Finally, Wilson (2017) attempt to clarify the relationships between models and theories related to the information seeking process.

Many of the models we briefly described in this section have played a fundamental role in the development of Information Seeking and Information Retrieval. These models not only formalise concepts related to searching for information, but also allow researchers to communicate more clearly and create novel hypotheses about human behaviour around search (Wilson, 2017). However, it is not always clear how these models can be translated into actionable insights that inform system design (White, 2016, c. 4). Wilson et al. (2010) attempts to connect theory and practice by proposing a framework for the evaluation of search interfaces informed by prior efforts in modelling user search behaviour.

2.1.6 Other Types of Search

In addition to various models of search behaviour, different *types* of search have been discussed in the literature (e.g., *exploratory*, *serendipitous* or *slow* search).

Exploratory search is a type of information exploration which refers to the activity of searching for information in a context where the searcher ***(i)*** is unfamiliar with the domain of their search-related goal, or ***(ii)*** uncertain about the ways to achieve their goal (i.e., uncertain about the technology to use or the process to employ), or ***(iii)*** uncertain about their goals. White (2016) describe two perspectives on exploratory search: on one hand, exploratory search can be used to describe an information-seeking goal that is open-ended, persistent and multifaceted, and on the other hand, exploratory search is an information-seeking process that is opportunistic, iterative and multi-tactical. Exploratory search is commonly employed in scientific, learning and decision-making contexts — where other exploratory information interactions, such as exploratory data analysis, are typically employed as well. White (2016) also point out that almost all searches are, to a certain extent, exploratory. However, exploratory search includes complex cognitive activities associated with knowledge acquisition and the development of cognitive skills (White and Roth, 2009).

Serendipitous search — the act of encountering information unexpectedly as part of task-focused information seeking — has been explored in a number of recent studies (André et al., 2009; Bordino et al., 2013; Rahman and Wilson, 2015). The motivation behind serendipitous search is derived from the fact that ever-improving, personalised search experiences, which display exactly what the searchers are looking for, limits serendipitous encounters with interesting (and relevant) information. As such, understanding serendipitous information encounter in traditional search contexts (such as web search) can be informative for the design of novel search systems. Within the context of entity search, Bordino et al. (2013) explore the use of entities extracted from different sources of content (i.e., Wikipedia and Yahoo! Answers) in promoting serendipitous search on the web. Rahman and Wilson (2015) deploy a live search system which makes use of searchers' social media data to promote micro-serendipitous information access, and conclude that work related queries drive serendipitous encounters more than leisure search.

Various other types of search have been described in the literature, including search as a leisure activity (Azzopardi, 2011*b*; Elsweiler et al., 2010, 2011; Harvey et al., 2014) or slow search (Teevan, 2015; Teevan et al., 2013).

Although associated by many, if not most, with the service provided by Google, searching for information is a complex, multi-dimensional human activity, that is influenced by a wide range of factors. The aim of this section was to introduce the conceptual framework and vocabulary used to communicate about search, from a user-centric perspective, and also highlight the complexity of search and the breadth of prior research that addresses this topic. In the following section, we attempt to review developments in the area of information retrieval models, moving from a user-centric towards a system-centric perspective.

2.2 System Models of Search

Salton (1968) defines Information Retrieval as "the field concerned with the structure, analysis, organization, storage, searching, and retrieval of information"[5] — a definition that is adopted or re-worded by many modern textbooks (Baeza-Yates and Ribeiro-Neto, 1999; Croft et al., 2009).

This section introduces some of the theoretical developments that underlie the process of structuring and retrieving information (i.e., matching documents to queries in what is known as *keyword search*). The main focus of this section is on theoretical models of structuring and retrieving information, and techniques for their evaluation, rather than specifically on algorithms or search engines — even though search engines are, ultimately, the main product of information retrieval modelling and research. The following chapter reviews the specific application of retrieval models to their most popular domain of use: web search.

2.2.1 Information Retrieval Models

A retrieval model is a formal representation of the process of matching a query and a document (Croft et al., 2009). It is typically used as the basis of ranking algorithms deployed in search systems, for the goal of producing a ranked list of results. In other words, the goal of retrieval models is to assign higher importance (or probability of relevance) to documents (within a collection) that are topically relevant to a given query, and rank documents according to their importance. Retrieval models most often encode statistical properties of text rather than linguistic structure of documents (Croft et al., 2009). This means that the frequency of words within a document is modelled, rather than relationships between words or their grammatical properties.

Various retrieval models have been discussed over the years. In chapters 6 and 7, we refer to several of these models (e.g., *BM25*) and related concepts (e.g., *cosine similarity*), as such, we describe some of the most widely known (and used) information retrieval models in this section.

[5] Which, as Zobel (2018) point out, is very similar to standard definitions of information science, which is typically seen as a much broader field.

2.2.1.1 Boolean Retrieval Models

The Boolean model is a basic retrieval model based on set theory and Boolean algebra (Baeza-Yates and Ribeiro-Neto, 1999, c. 2) — it is sometimes referred to as *exact-match* retrieval, because documents are ranked by the model if they match the searcher query exactly. The name of the model is derived from the fact that ***(i)*** there are only two possible outcomes for query evaluation (i.e., the document is relevant or not, as relevance is assumed to be binary rather than graded), and ***(ii)*** the query is typically specified using Boolean operators (e.g., NOT, AND, OR). Regular expressions, proximity operators and wild-card characters are other types of operators sometimes used in building queries for Boolean retrieval.

The Boolean model is simple and easy to grasp by searchers. It selects as relevant all documents that contain the exact query terms. Because the query can contain any document feature, not only terms, it is typically used where specifying document meta-data, such as document creation date or document type, in the query is useful (e.g., email search). The drawbacks of the Boolean model are that ***(i)*** it is not always easy to translate information needs to Boolean expressions, ***(ii)*** search effectiveness depends entirely on the searcher's ability to construct queries[6], and ***(iii)*** it makes no distinction between documents that contain query terms, yet are more or less relevant to the query.

Although extended in various ways over its history (Salton et al., 1983), the Boolean model is now superseded by retrieval models that encode term weighting, recognising that certain terms within documents are more discriminative with respect to topical focus than others. One such model is the vector space model, discussed next.

2.2.1.2 Vector Space Model

The vector space model (Salton and Lesk, 1968) has been used in information retrieval research for over 50 years, and continues to be used frequently in research publications to this day. Indeed, in chapters 4 and 7, we make use of some of the concepts related to the vector space model of information retrieval.

The model assumes documents and queries to be part of a p-dimensional vector space, where p is the number of unique terms observed (in the collection of documents) or used in implementing the model. Document D_i (of a collection containing N documents, $1 \leqslant i \leqslant N$) and query Q are represented by the follow-

[6] To such an extent that in certain domains, such as legal search or patent search, specialists are employed to translate human information needs to Boolean expressions and queries.

ing term vectors:

$$D_i = \{d_{i1}, d_{i2}, ..., d_{ip}\}$$
$$Q = \{q_1, q_2, ..., q_p\}$$

Given this representation of both queries and documents, documents in a collection can be ranked by computing the distance between their corresponding vectors and the query vector. One of the most widely used methods of computing this distance in information retrieval is *cosine similarity*. Chapter 7 makes use of cosine similarity as described here, to compute the distance between vectors of heterogeneous documents learnt through a graph-based representation learning approach. The cosine similarity measures the angle between document vectors and query vector, such that, when the vectors are normalised, the angle between identical vectors is 1, and between non-overlapping vectors (i.e., vectors that do not share any non-zero weighted elements) the angle is 0 (Croft et al., 2009).

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{p} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{p} {d_{ij}}^2 \cdot \sum_{j=1}^{p} {q_j}^2}}$$

The vector space model ranks documents according to their degree of similarity to the query, and as such, a document can be ranked highly even if it only partially matches the query (in contrast with the Boolean model, where partial matching is a consequence of query "design"). There is no theoretical motivation for preferring cosine similarity over other similarity measures, but it performs better in evaluations of search quality (Croft et al., 2009).

Term weighting is another concept derived from research relate to the vector space model (although it has been applied to a wide range of retrieval models). Term weighting refers to scaling the weights of individual terms, within the vector representation of both queries and documents, according to their relative importance in *discriminating* document or query topic. Many weighting schemes have been explored, but perhaps the most widely known is *tf.idf* term weighting (i.e., term frequency - inverse document frequency weighting). The term frequency component (*tf*) reflects the importance of a term within a document, whereas the the inverse document frequency (*idf*) reflects the importance of the term in the collection of documents. Typically, the two elements are computed:

$$tf_{ik} = \frac{f_{ik}}{\sum_{j=1}^{p} f_{ij}}$$

$$idf_k = \log \frac{N}{n_k}$$

where tf_{ik} is the *term frequency* weight of term k in document D_i, f_{ik} is the number of occurrences of term k in document D_i, and the denominator is a normalising factor. Similarly, idf_k is the *inverse document frequency* weight for term k, N is the total number of documents in the collection, and n_k is the number of documents in the collection in which term k occurs. This weighting scheme has been developed empirically over the past decades (Croft et al., 2009; Spärck Jones, 1972, 1973), rather than theoretically, although an argument for a principled derivation of *tf.idf* has been made (Robertson et al., 2004). We make use of this weighting scheme in chapter 4, as an aid in computing the similarity of composite object titles, as assigned by participants in one of our studies, and in chapter 6, where we use it to rank entities associated with documents in an effort to compute the similarity of heterogeneous documents.

The vector space model makes the implicit assumption that relevance (topical or situational) is related to the similarity between query and document vectors (i.e., documents that are more similar to the query, as determined by cosine similarity, are more likely to be relevant). Various extensions of the model have been considered (Harman, 1992; Lundquist et al., 1997; Rocchio, 1971), primarily concerned with incorporating *relevance feedback* in the process of generating vector representations for queries or documents. Despite its simplicity, the vector space model is still considered a resilient ranking strategy for general collections (Baeza-Yates and Ribeiro-Neto, 1999).

2.2.1.3 Probabilistic Models

Probabilistic models attempt to frame the problem of matching queries to documents within a probabilistic framework. Given a query Q and a document D_i (from a collection containing N documents), probabilistic models try to estimate the probability that the searcher issuing query Q finds document D_i relevant. Typically, the assumption that relevance depends on the properties of the query and the documents only (i.e., topical relevance) is made in probabilistic models as well. As stated in the *Probability Ranking Principle* by Robertson (1997):

> “If a [search engine’s] response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.”

Croft et al. (2009) note that, under certain assumptions, such as the relevance of a document to a query being independent of other documents, it is possible to prove that the *Probability Ranking Principle* statement is true — that is, that ranking documents by their probability of relevance maximises the precision (i.e., the proportion of relevant documents, at any given rank) of a search system. However, there are different ways to estimate or define the probability of relevance and, as such, many instantiations of probabilistic retrieval have been developed, with different approaches proposing different methods of estimating the probability of document relevance. Fuhr (1992) and Spärck Jones et al. (2000) provide comprehensive reviews of probabilistic models for information retrieval. In our work, we make use of the popular *BM25* probabilistic retrieval algorithm and, as such, we review the model from which it is derived in more detail next[7].

The Binary Independence Retrieval Model. This retrieval model assumes that within any collection there are two non-overlapping sets of documents: those that are relevant to a query and those that are not. Given this assumption, the task of the retrieval model is to decide which set a document belongs to (i.e., classify documents using two labels: relevant or non-relevant). Framing this within probabilistic context, the task of the model is to assess whether:

$$P(R|D) > P(NR|D)$$

where $P(R|D)$ is the conditional probability of document relevance, given document features and $P(NR|D) = 1 - P(R|D)$ (this inequality is commonly known as the *Bayes Decision Rule* when used in classification problems).

In practice, estimating $P(R|D)$ directly is less straightforward than first estimating $P(D|R)$, and then using *Bayes’ Rule* (Bayes, 1763) to determine the probabil-

[7] Our review of The Binary Independence Retrieval Model and *BM*25 follows the derivations in Croft et al. (2009, c. 7).

ity of document relevance given document features:

$$P(R|D) \propto P(D|R) \cdot P(R)$$

where $P(R)$ is the *a priori* probability of relevance. The decision rule above can be rewritten as $P(D|R) \cdot P(R) > P(D|NR) \cdot P(NR)$, which is commonly expressed as:

$$\frac{P(D|R)}{P(D|NR)} > \frac{P(NR)}{P(R)}$$

where the left hand side of the inequality is known as the *likelihood* or *odds* ratio. In a search setting, documents are ranked based on their likelihood ratio.

To calculate individual document likelihood ratios, $P(D|R)$ can be computed in various ways. The approach used by this model is to make certain simplifying assumptions about the representation of documents: ***(i)*** each document D is represented by a term vector $\{t_1, ..., t_t\}$, where each term has a weigh of 1 if present in the document, or 0 if it is not present (i.e., *binary* weights); and ***(ii)*** terms within documents are independent of each other. These two assumptions give the model its name. Given the term independence assumption, the likelihood ratio can then be computed using the following:

$$\frac{P(D|R)}{P(D|NR)} = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i}$$

where p_i is the probability of a term occurring in the relevant document set, and s_i is the probability of a term occurring in the non-relevant set. This can be manipulated into:

$$\frac{P(D|R)}{P(D|NR)} = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot (\prod_{i:d_i=1} \frac{1-p_i}{1-s_i} \cdot \prod_{i:d_i=1} \frac{1-s_i}{1-p_i}) \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i}$$

$$\frac{P(D|R)}{P(D|NR)} = \prod_{i:d_i=1} \frac{p_i\,(1-s_i)}{s_i\,(1-p_i)} \cdot \prod_i \frac{1-p_i}{1-s_i}$$

where the second factor of the product is the same over all documents and can be ignored. For practical reasons, the log of this ratio is typically used to score and rank documents:

$$\frac{P(D|R)}{P(D|NR)} = \sum_{i:d_i=1} \log \frac{p_i\,(1-s_i)}{s_i\,(1-p_i)}$$

It is common that the query provides the only information about which documents are in the relevant set – assuming terms that are not in the query are equally likely in both the relevant and the non-relevant set (i.e., $p_i = s_i$). In that case, the summation above is only over terms that are in the query and the document and, thus, given a query, a document's score is the sum of weights for terms that occur in both the query and the document.

If there is no information about which documents are in the relevant set, the additional assumption can be made that p_i is constant, and s_i can be approximated using term frequencies in the collection as a whole (i.e., because the number of non-relevant documents in a collection is typically much larger than the number of relevant documents). Assuming a value of $p_i = 0.5$, in the scoring function described above, each term i has a weight of:

$$w_i = \log \frac{0.5\,(1 - \frac{n_i}{N})}{\frac{n_i}{N}\,(1 - 0.5)} = \log \frac{N - n_i}{n_i}$$

where n_i is the number of documents in a collection of N documents that contain term i. This is approximately equivalent to the inverse document frequency weight discussed above, in the context of the vector space model (here, the term frequency component is absent because document terms have binary weight).

If additional information about the relevant set is available, such as r_i, the number of relevant documents in which term i occurs, and R, the number of relevant documents for a given query, the probability of a term appearing in the relevant set can be estimated as $p_i = r_i/R$ — similarly $s_i = (n_i - r_i)/(N - R)$. The scoring function above can then be expressed as:

$$\sum_{i:d_i=q_i=1} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}$$

where 0.5 is added to various terms to prevent division by zero or an undefined logarithm. Note that if no relevance information is available, setting r and R to 0 would give a p_i value of 0.5, as discussed above. Overall, the absence of a term frequency component in this scoring approach makes the *Binary Independence Model* not very effective, compared to $tf.idf$ ranking. However, this scoring function is the basis of one of the most widely used probabilistic retrieval algorithms, which we discuss next.

The BM25[8] Retrieval Algorithm. *BM25* extends the Binary Independence Model scoring function by including term and document weights explicitly. It is not a formal model of retrieval, but rather an experimentally derived scoring function. The most common form of the *BM25* scoring function is:

$$\sum_{i \in Q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)\, f_i}{K + f_i} \cdot \frac{(k_2 + 1)\, q\, f_i}{k_2 + q f_i}$$

where the summation is over all terms of query Q, f_i is the frequency of term i in the document, qf_i is the frequency of term i in the query and k_1, k_2 and K are empirically-derived parameters. The parameters k_1 and k_2 determine how the term-frequency component (in document or query, respectively) affects overall document weight, whereas K is a parameter that normalises the term-frequency component by document length, and is typically expressed as:

$$K = k_1(1 - b + b \cdot \frac{dl}{avdl})$$

where dl is document length and $avdl$ is average document length across the collection, measured in either number of characters or number of terms, and the parameter b controls the impact of the length normalisation (most commonly set to the empirically-derived 0.75).

Overall, *BM25* is an effective model, used widely in practice and in the research literature as a common baseline. In both chapters 6 and 7, we make use of *BM25* scoring to retrieve documents from heterogeneous collections and construct composite objects from these initial rankings. In addition to the retrieval models discussed so far, many other types of retrieval models have been discussed in the research literature over the past decades. We briefly review some of them in the following section.

2.2.1.4 Other Models of Information Retrieval

Much like user models of search, many different information retrieval models have been studied over the past decades. Even the models discussed in the previous sections have been extended into many different variants. Croft et al. (2009, c. 9) and Manning et al. (2008, c. 11–15) provide thorough discussions of retrieval models from an applied perspective, whereas Baeza-Yates and Ribeiro-Neto (1999, c. 2) provide a comprehensive review of research developments in

[8] *BM* stands for *Best Matching*, whereas 25 is just an artefact of a numbering scheme used by the developers of the algorithm.

the area of information retrieval modelling. Although the models we use in our work have been discussed in the previous sections, we briefly review here other popular information retrieval models that have played a fundamental role in the development of Information Retrieval.

One of the most widely known family of probabilistic models, in addition to *BM25*, is the *Divergence-from-Randomness (DFR)* set of models (Amati and Van Rijsbergen, 2002). The paradigm proposed by *DFR* encodes the idea that the divergence of the within-document term frequency from the within-collection term frequency is an accurate estimate of the information carried by a term t in document D. Specifically, term weights are computed by measuring the divergence between the actual term distribution and the term distribution produced by a random process. Because there are many random processes to choose from, different alternatives are proposed in Amati and Van Rijsbergen (2002), with each option defining a basic *DFR* model.

Another probabilistic approach to information retrieval is found in the application of language modelling to the task of matching queries to documents. Language modelling approaches to information retrieval typically encode the idea that a document is a good match to a query if the document language model (i.e., the probability distribution over terms within the document) is likely to generate the query, something that is more probable if the document contains the query terms often. Unlike traditional probabilistic approaches, which aim to model the probability of document D being relevant to query q, language modelling approaches build a probabilistic model from each document D_i within a collection of documents, and rank documents based on the likelihood of their corresponding language models to generate query q. Manning et al. (2008, c. 12) provide an expert review of language modelling-based information retrieval.

There is considerable overlap between the fields of Information Retrieval and Machine Learning. Many of the developments in Information Retrieval (e.g., incorporating relevance feedback into ranking) can be described as machine learning algorithms, and many machine learning algorithms can be directly applied to the task of matching documents to queries (e.g., text classification). Over the past decades, machine learning techniques applied to information retrieval have been described under the label of *Learning-to-Rank (LTR)*. *LTR* techniques are seen as a subset of supervised learning, which means that these techniques typically require training data to *"learn"* optimal ways of combining any types of features extracted from query-document pairs (e.g., term frequencies, user clicks as observed in a search engine query log, or even the output of other ranking models like *BM25*) such that the retrieval model can accurately output the probability of a document being relevant, given its training and structure. Although usually not as principled as some of the previous models discussed here (in that, in general, *LTR* techniques do not necessarily make underlying assumptions about the nature of relevance, but rather optimise certain measures of success, as defined by a loss function), *LTR* technique have proved very effective in large scale search applications, such as web search, mostly because commercial search companies incorporate hundreds of behavioural and content features in their retrieval models. Liu (2009) provides a review of modern *LTR* models.

In the most popular application domain of information retrieval, web search, the network structure of the environment allows for modelling approaches that incorporate link information. *HITS* (Kleinberg, 1999) and *PageRank* (Brin and Page, 1998*b*) are, perhaps, some of the most widely known web search algorithms that are based on models of linked structures. In particular, *PageRank* simulates a searcher navigating on the web who jumps to a random page with probability p, or follows a random hyperlink from their current page with probability $1 - p$. This process is modelled using a Markov Chain (Norris, 1997), from where the stationary probability of being in each page is computed and is then used in the ranking mechanism (Baeza-Yates and Ribeiro-Neto, 1999).

The aim of this section was to introduce the models used throughout our work, as well as review some of the models that have played a fundamental role in the development of Information Retrieval. Advances in information retrieval can only be achieved by determining whether novel approaches to retrieval modelling outperform existing ones. In this process, evaluation plays a key role, and we review different aspects related to information retrieval evaluation next.

2.2.2 Evaluation of Retrieval Models

Evaluation is a key component of developing information retrieval systems. From an algorithmic perspective, the evaluation of search systems is typically concerned with two aspects: *effectiveness* and *efficiency* of retrieval. Effectiveness is concerned with measuring the ability of a search system to place relevant documents closer to the top of the results ranking, and efficiency is concerned with how fast and how many resources are used in generating a response to searcher requests issued to the system. Other types of evaluation, which consider situational relevance, are discussed in the following chapter. In this section, we review some measures of search system effectiveness, as these types of measures are used later, in chapters 6 and 7.

2.2.2.1 Effectiveness Measures

The effectiveness of search systems is typically considered in a laboratory setting, where the task of submitting a query to the search system is simulated in batch mode. In broad terms, the laboratory setting implies that ***(i)*** a collection of documents is available to researchers, ***(ii)*** a set of queries pertinent to the collection is selected by researchers and ***(iii)*** for each selected query, a candidate set of documents (i.e., potentially relevant documents) is manually judged by a group of employed assessors as being relevant to the query or not (using either binary or multi-level relevance gradings). The candidate set of documents may contain the entire collection, or may be a sample of the collection assembled using filtering methods. In the evaluation procedure, each of the selected queries is automatically issued to the retrieval algorithm under evaluation, which returns a ranked list of documents from the underlying collection, ordered by their estimated relevance to the query. The quality of the rankings returned by the retrieval algorithm is then assessed, using measurements not unlike the ones described later in this section, which make use of assessor relevance labels. Typically, average performance metrics are computed (i.e., averages of quality measurements over individual queries) and interpreted as measures of overall algorithm or system effectiveness. Multiple algorithms can be evaluated and compared using this methodology — yet decisions about which algorithm performs better in an application setting usually include aspects regarding *efficiency* or other constraints (e.g., if a retrieval algorithm performs twice as better as another one, but takes ten times as long to respond to queries, it might not necessarily be useful in a practical setting). Baeza-Yates and Ribeiro-Neto (1999) mention repeatability and scalability as the two main advantages of laboratory-based evaluations of search effectiveness over real life experiments. Indeed, we evaluate the effectiveness of our result composition in a laboratory setting in chapter 6, and the retrieval (and aggregation) effectiveness of using unified representations for heterogeneous documents in chapter 7. We now review some of the popular measures used to evaluate ranking quality in laboratory settings.

Precision and Recall Two of the most popular measures of search quality are *precision* and *recall* — introduced in Cleverdon (1970). Recall measures how well the search system is at finding all the relevant documents in the collection, for a given query, whereas precision measures how well the system is at rejecting non-relevant documents (Croft et al., 2009). Specifically, if A is the set of relevant documents (that exist in the collection) for a given query, B is the set of documents retrieved by a system under examination, and $|.|$ gives the cardinality of a set, precision and recall are defined as:

$$precision = \frac{|A \cap B|}{|B|}$$

$$recall = \frac{|A \cap B|}{|A|}$$

where *precision* is interpreted as the proportion of documents that are retrieved and are relevant (based on labels given by assessors), and *recall* is interpreted as the proportion of relevant documents from the collection that are retrieved.

Precision and recall as defined above, used in evaluating search, make the assumption that all documents returned by a search system are examined by searchers. However, searchers typically examine a ranked list of results, ordered by estimated relevance, with the result most likely to be relevant (as estimated by the system) placed at the top of the list. Given that users tend to inspect the top ranked results only, precision and recall measures are usually computed at specific (top) ranks of the results list. For example, if set B_k contains the top k results returned by the search system, precision or recall at rank k are defined as:

$$precision@k = \frac{|A \cap B_k|}{|B_k|}$$

$$recall@k = \frac{|A \cap B_k|}{|A|}$$

Common values of k are $5, 10, 20$. It is important to note that the implicit search task assessed by rank-limited measures is that of placing relevant items in the top k section of the ranking, rather than finding as many relevant documents as possible (i.e., beyond k, ranking quality is not considered). In addition, such rank-limited measures do not distinguish between differences in rankings at positions 1 to $k-1$, which is usually an important aspect of ranking quality (e.g., precision@10 has the same value for two systems, one that returns relevant documents at positions 1 and 2, and the other that returns relevant documents at positions 9 and 10). We make use of *precision at rank* in chapters 6 and 7.

Another common method of measuring system quality is by computing the average of precision values at ranks where relevant documents are placed (i.e., where ranking recall increases). For example[9], consider a ranking of documents,

[9] Example taken from Croft et al. (2009, c. 8).

with each document d_i subscripted by its ranking position, their associated relevance labels (i.e., a value of 0 at position i meaning that the document at position i is not relevant) and the set of corresponding *precision@k* values:

$$\begin{aligned}
Ranking &= \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}\} \\
Relevance &= \{1, 0, 1, 1, 1, 1, 0, 0, 0, 1\} \\
precision@k &= \{1.0, 0.5, 0.67, 0.75, 0.8, 0.83, 0.71, 0.63, 0.56, 0.6\} \\
average\ precision &= (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78
\end{aligned}$$

Average precision (AP) has the advantage of being a single measure, derived from multiple rank-limited precision values, but depends heavily on relevant documents being placed at the top of the results list. As such, this measure is appropriate for evaluating both the task of finding as many relevant documents as possible, and the task of placing these documents at the top of the ranking.

The measures of quality discussed so far are used in single instances of retrieval (i.e., for a single query and its associated ranking). However, the effectiveness of retrieval algorithms is typically assessed with respect to many different queries. As such, methods that summarise search quality for an entire set of evaluation queries have been proposed. One of the more popular summary measures of search quality is the average of *average precision* values across queries, more commonly known as *mean average precision (MAP)*. *MAP* provides a compact summary of the effectiveness of retrieval algorithms across many different queries, although, much like other instances of averaging, information can be lost in creating such summaries. Various modifications to *MAP*, such as *GMAP*, the geometric mean of average precisions (Robertson, 2006), that partially address this loss of information, have been proposed. Fuhr (2018) discusses some of the limitations of *MAP*, suggesting it is based on a superficial user model and can be misleading. We make use of *MAP*, in addition to other measures of search quality, in chapter 7, where we compare system effectiveness in retrieving (and aggregating) diverse documents represented in a unified feature space.

As mentioned earlier, searchers typically inspect ranked lists of results in a top-down approach. It seems, then, that placing relevant items at the top of a ranking, rather than lower down in the ranking, can be useful, as it reduces the effort required to locate relevant information. We now discuss measures of search quality that explicitly encode this assumption.

Gain One of the popular measures used in assessing retrieval algorithms is *discounted cumulative gain (DCG)*, proposed by Järvelin and Kekäläinen (2002). *DCG* encodes two assumptions: ***(i)*** that documents are less useful to searchers if they are ranked lower in the list, and that ***(ii)*** marginally relevant documents are less useful than highly relevant documents (which can be operationalised only in cases where multiple levels or relevance are assessed). *DCG* measures the (assumed) usefulness or *gain* of examining a document at various ranks. The most gain is derived from examining a highly relevant document at the top of the ranking, and the gain associated with relevant documents lower in the ranking is reduced or *discounted*. The total gain obtained from a given ranking, at a particular rank *k*, is measured through *DCG@k* or DCG_k, formalised as:

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{log_2 i}$$

where rel_i is the relevance level of document d retrieved at rank i, although binary relevance judgements can also be used. The $log_2 i$ used in the denominator provides the degree with which gain is reduced at lower rank — the log is used to provide a gradual, rather than linear, reduction of gain.

Similar to *precision@k*, specific values of k can be used in evaluating search quality, and these values can be averaged across a set of queries to compute and overall estimate of (rank biased) search quality. One of the problems with this approach is that queries can have different numbers of associated relevant documents (in a given collection), and, as such, *DCG* values can vary widely across queries, making an average *DCG* value difficult to interpret. To allow for comparisons across queries, Järvelin and Kekäläinen (2002) propose a normalised version of *DCG (nDCG)*. Normalising occurs by comparing the *DCG@k* values to the ideal *DCG (IDCG)* at position *k*. *IDCG* is computed by sorting all the documents in the collection that are relevant to a given query by their relative relevance, and computing *DCG@k* over this sorted list; formally:

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

$$IDCG_k = \sum_{i=1}^{|A_k|} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Some of the limitations of *nDCG* include the fact that it does not penalise non-relevant documents placed high in the ranking, and that it is unsuitable for queries that often have many equally relevant results (e.g., *"things to do in Glasgow"*). We make use of *nDCG* in chapters 6 and 7 to evaluate different types of (rank biased) system effectiveness.

Many other measures of search quality have been proposed in the (relatively) long history of Information Retrieval. With respect to system effectiveness, measures such as the harmonic mean of precision and recall (*F-score*), mean reciprocal rank (*MRR*) or binary preference-based measures (Sakai and Kando, 2008) are commonly used in the research literature. In a more user-oriented approach to evaluation, Borlund (2003*b*) discuss the use of simulated work tasks together with alternative performance measures, such as relative relevance or ranked half-life, as a more realistic alternative to system evaluation.

The design and interpretation of retrieval evaluation measures is typically explored in test environments, where information about documents and their relevance to queries is available to researchers. We briefly discuss environments that enable laboratory-based information retrieval research and model evaluation in the following section.

2.2.2.2 Test Collections

To facilitate the reproducible study of novel retrieval algorithms, test collections specialised for this purpose have been developed over the past decades — an early review of test collections for information retrieval research is provided in Spärck Jones and van Rijsbergen (1976). Many of these publicly available collections are assembled by the community around the Text Retrieval Conference[10] (Voorhees and Harman, 1999), where researchers and practitioners of search typically compete in a series of challenges aimed at advancing the field of Information Retrieval in various ways.

Many collections have been developed over the past decades, tailored towards different search-related tasks, such as microblog retrieval (Sequiera and Lin, 2017), medical record search (Voorhees, 2013) or search over specialised information structures, such a genome data (Hersh, 2002) or large-scale chemical data (Lupu et al., 2009). The characteristics of such collections have been discussed briefly in the previous section. Typically, these collections contain a set of documents, a set of queries that might be issued to a search system tasked with locating relevant content within the collection of documents, and a set of binary or multi-level document relevance judgements for each query (usually collected

[10] Commonly known as TREC.

by employing people to assess relevance).

In our work, we make use of the collection assembled by Nguyen et al. (2012). This publicly available collection contains results from 108 real web search engines, ranging from large general web search engines such as Google or Bing, to small domain-specific engines (such as image search services). For a set of 50 test queries, Nguyen et al. (2012) collected relevance judgements for the top 10 results returned by *each* of the search engines they considered. In chapter 6, we make use of this collection to investigate our approach to result composition in a heterogeneous web environment.

Test collections and their corresponding test environments allow research in the field of Information Retrieval to advance, by enabling easily controllable and reproducible studies. The efforts made to develop theses publicly available datasets are significant, and we acknowledge them here, as they also enable our work.

2.2.3 Retrieval Across Collections

The retrieval models discussed in the previous sections make the assumption that documents within a collection have the same underlying (textual) structure, or that the same model applies equally as well to different collections. In practice, however, different models, algorithms or algorithm parameters are used in conjunction with different collections to optimise retrieval effectiveness. In certain situations, it is desirable to merge the outputs of several retrieval systems, each aimed at a different collection, within one single ranking, in order to provide users with unified search access to a wide range of collections. This raises interesting questions about how to select collections that contain relevant information, when multiple collections are available, how to assess collection relevance to a given query, or how to merge rankings output from different retrieval models (and different underlying collections) within a unified ranking. The following sections reviews prior research efforts focused on modelling information retrieval across collections of documents.

2.2.3.1 Federated Search

Federated search — also known as federated information retrieval or distributed information retrieval — is a technique for searching multiple text collections simultaneously (Shokouhi and Si, 2011). In federated search, queries are submitted to search systems connected to different document collections. Results are returned by selected systems (i.e., systems that are estimated as relevant), and are then merged within a single ranked list. Federated search is preferred over centralised search alternatives in many environments. Shokouhi and Si (2011) mention the example of the hidden web, where specialised search systems are able to access hidden collections of documents, that search engines such as Google cannot easily index (examples of such systems are *PubMed*[11], *the US census bureau*[12] or various patent offices that provide search over private information). Instead of attempting to index documents in the underlying *"hidden"* collections of such systems (some of which cannot be accessed at all), federated search techniques pass queries to the interface of the specialised search systems and merge their results in a unified ranking. The final output is a ranking of documents, originating from various collections, ranked by their estimated relevance to the query.

The major challenges in federated search are related to ***(i)*** determining which specialised search systems to issue the query to (known as the collection *selection* problem), ***(ii)*** creating representations of search systems for the purpose of determining which systems and their associated collections are relevant to the query (known as the collection *representation* problem), and ***(iii)*** merging results from various systems within a single ranking by computing a unified relevance estimate across specialised rankings (known as the result *merging* problem).

The collection representation problem is typically addressed by issuing a wide range of queries (e.g., a representative sample of queries observed previously) to each specialised search system and collecting the results returned. Each result sample (associated with a specific search system) is then assumed to be representative of the specialised service and its underlying collection (sometimes referred to as the *representation set* of a search system). What make a sample of documents representative for a collection is one of the problems that is addressed by prior research in this space. Shokouhi and Si (2011, c. 2) provide an expert review of modern approaches to collection representation.

In federated search, after a query is issued by a user, the system needs to determine which collections contain potentially relevant information. In other words, collections need to be ranked according to their estimated relevance to the query. Various strategies for selecting relevant collections have been explored. Shokouhi and Si (2011) defines the following categories of techniques: ***(i)*** lexicon-based strategies, which treat collections as *bags of words*[13] and rank them according to their (i.e., their representation set) lexical similarity to the query; ***(ii)*** document surrogate strategies, which incorporate information from individual document rankings in addition to lexical similarity; or ***(iii)*** classification-based

[11] https://www.ncbi.nlm.nih.gov/pubmed/ – Accessed March 2018.

[12] http://www.census.gov – Accessed March 2018.

[13] A set of unique terms observed in the collection.

strategies, which attempt to determine relevant collections by assessing the similarity of a query to other queries for which relevant collections are known — other specialised strategies are reviewed in Shokouhi and Si (2011, c. 3).

In our work, we make use of the relevant document distribution estimation (*ReDDE*) collection selection strategy (Si and Callan, 2003). *ReDDE* attempts to select a small number of collections, with the largest number of relevant documents, by explicitly estimating the distribution of relevant documents across all the collections. It then ranks collections based on the likelihood of containing relevant information. Formally, the number of documents relevant to a query q in a collection c is estimated as follows:

$$R(c,q) \approx \sum_{d \in S_c} P(R|d) \frac{|c|}{|S_c|}$$

where $P(R|d)$ is the estimated probability of document d being relevant (e.g., by counting how many query terms exist in the document), $|c|$ is the size of the collection, which is typically known or can be computed *a priori*, $|S_c|$ is the size of the collection's representation set, and $|c|/|S_c|$ is an estimate of prior probability of collection relevance (e.g., if there is one relevant document in the representation set S_c, $|c|/|S_c|$ relevant documents are expected in the overall collection). Given that the probability of document relevance to a query can be estimated in various ways, *ReDDE* approximates this probability by merging all representation sets into a single collection (*CSI*) and ranking all documents in *CSI* by their similarity to the query. A document's estimated relevance is then assumed to be proportional to its position in this unified ranking. Typically, the probability of relevance is assumed constant for the top ranked documents of the unified ranking of documents; formally:

$$p(R|d) = \begin{cases} \alpha, & \text{if } r(d) < \beta \sum_i |c_i| \\ \\ 0, & \text{otherwise} \end{cases}$$

where $|c_i|$ denotes the number of documents in collection c_i, β is a percentage threshold (Si and Callan (2003) suggest the value $\beta = 0.003$ achieves robust performance across collections) which selects how many documents at top rank to consider relevant, and $r(d)$ is the rank of a document in the unified ranking.

Many different variants of *ReDDE* have been proposed in the literature, for various specific search tasks, primarily investigating alternative methods of estimating relevance (Si and Callan, 2004; Thomas and Shokouhi, 2009). We make use of *ReDDE* in chapter 6 to estimate collection relevance in a heterogeneous web environment in order to apply our result composition approach.

The last step of federated search involves merging (or *interleaving*) results[14]. Each search system returns a list of documents that is sorted by document relevance to the query. The federated search system must, then, estimate a unified relevance measure associated with individual documents, across rankings, and rank all documents according to this estimate. Most merging techniques assume that the degree of overlap among individual rankings is negligible (Shokouhi and Si, 2011) — although de-duplication has been explored in the context of federated search as well (Bernstein et al., 2006; Shokouhi et al., 2007).

One of simplest (and most popular) approaches to result merging is the *CORI*[15] formula (Callan et al., 1995). *CORI* combines relevance estimate scores for both collections and documents into a unified document score. Formally, given a collection score C_i, its normalised relevance score C'_i is computed by:

$$C'_i = \frac{C_i - C_{min}}{C_{max} - C_{min}}$$

where C_{min} is the score of the collection estimated as least relevant, and C_{max} is the score of the collections estimated as most relevant. Collection score is then combined with individual document scores. Given a document D_i, its unified relevance score D'_i is estimated by:

$$D'_i = \frac{D_i + \alpha \cdot D_i \cdot C'_i}{\beta}$$

where α and β are heuristic-derived weighting parameters (Callan et al. (1995) suggest $\alpha = 0.4$ and $\beta = 1.4$). We make use of the *CORI* re-weighting technique in chapter 6, where we explore merging results, retrieved from various heterogeneous collections, within *composite objects*.

Other types of result merging technique have been explored, from using regression to determine optimal collection and document weighting strategies (Si and Callan, 2002, 2003), to merging multi-lingual documents in a unified ranking (Si and Callan, 2006; Si et al., 2008). Shokouhi and Si (2011) provide an excel-

[14] Different types of result merging are discussed in the literature: *federated search merging*, *data fusion* and *metasearch*. Data fusion typically refers to merging the outputs of various search systems that return results from the *same* collection, whereas *metasearch* is used interchangeably with federated search merging.

[15] From *collection retrieval inference network*.

lent review of modern research efforts and practical developments in the space of federated information retrieval.

2.2.3.2 Aggregated Search

Much like federated search, *Aggregate Search* is the task of selecting collections (i.e., verticals) and merging results into a unified ranking. Unlike federated search, aggregated search explores collections of documents that are *heterogeneous*, meaning that the they contain different types of media (e.g., *image*, *video* or *tweets*) or focus on different types of search tasks (e.g., search for local business, products for sale, scientific articles) — indeed, aggregated search is sometimes referred to as federated search over heterogeneous environments.

As with federated search, the main tasks in aggregated search consist of vertical *selection*, vertical *representation* and results *merging*. In contrast to federated search, where most approaches assume all collections contain textual documents (and therefore can be processed in the same way), in aggregated search, most collections contain very different types of items that can not be indexed and searched in the same way (Arguello, 2017). This heterogeneity of content makes estimating relevance across verticals challenging — unlike federated search, where the same scoring function can be applied to every available collection. Thus, in aggregated search, vertical selection and merging approaches need to make use of vertical-specific features. Finally, most federated search approaches assume that results retrieved from different sources can be merged in an unconstrained fashion (Arguello, 2017). However, in most aggregated search approaches, results are typically merged at block-level, because document surrogates vary widely across collections and also because estimating unified relevance scores across heterogeneous vertical is challenging. As such, result presentation approaches in aggregated search need to consider where and how to display vertical results within a unified ranking. To summarise, the additional challenges in aggregate search compared to federated search are: ***(i)*** documents across collections have widely varying underlying representations, and as such a uniform approach to estimating collection relevance is difficult; ***(ii)*** document relevance across collections is not directly comparable; ***(iii)*** merging results in unified ranking needs to consider vertical-specific features, in addition to document relevance estimates. Many of these challenges transfer to result composition as well, some of which we address in chapter 7, where we study the problem of representing heterogeneous documents in a unified feature space.

Modern approaches to aggregated search make use of machine learning to combine a wide range of features into vertical selection and vertical presentation models. These features can be derived from query or vertical properties (e.g., for a query containing the term *"images"*, it is likely that the image vertical is relevant). However, given differences between collections in an aggregated search context, not all features may be available across verticals — Arguello (2017) give the example of the weather vertical, which is typically not directly searchable, and therefore does not have a vertical-specific search log or features derived from such a log. Another challenge is that features across verticals might indicate relevance in different ways. For example, items retrieved from the news vertical can receive more clicks than items from the weather vertical (i.e., items from the weather vertical can satisfy a searcher's information need directly on the unified results page), therefore incorporating such features into vertical selection or vertical presentation models is a challenge. Arguello (2017) provide an expert review of modern approaches to aggregated search, covering both user-centric and system-centric perspectives.

Like aggregated search, result composition is concerned with selecting heterogeneous documents from different verticals and merging these documents within a unified results page. Unlike aggregated search, result composition aims to construct complex result aggregates, that contain information from multiple verticals, which are specific to individual aspects of a searcher's query, and are not integrated within the results page as a (vertical or horizontal) ranking. In addition, result composition also explores *which* results to return to users from individual verticals, rather than simply returning the top-k results, as in the case of most modern approaches to aggregated search.

Aggregated search has been studied widely in the context of web search, as such, we review its application in more detail in the following chapter, where we discuss interfaces and interactions in web search.

2.2.4 Clustering and Composition

Organising documents around common properties or features has been widely explored in Information Retrieval. Ranking, for example, is a type of organising documents around their common properties in relation to a user's query, whereas clustering and composition typically relate to organising documents around their common properties in relation to each other. Most of the basis for research around clustering and composition is provided by the cluster hypothesis, originally formulated by van Rijsbergen (1979):

> "Closely associated documents tend to be relevant to the same request."

In this section we review clustering as applied to information retrieval, as well as *composite retrieval* and its defining characteristics in contrast to clustering.

2.2.4.1 Clustering and Retrieval

Clustering algorithms for information retrieval group documents within a collection into closely-related subsets (i.e., clusters). The goal of clustering algorithms is to identify documents that are similar to each other, and assign them to a common clusters in such a way as to maximise the cluster's internal cohesion, while at the same time minimising similarities between clusters.

Different types of clustering algorithms have been developed and applied to information retrieval. Manning et al. (2008) distinguishes between *flat* and *hierarchical* clustering algorithms, where flat clustering generates a set of clusters without explicit structure that can relate clusters to each other, whereas hierarchical clustering outputs a hierarchy of clusters that have explicit connections. Flat clustering approaches have the benefit of being conceptually simple, but require defining a specific number of clusters K *a priori* and are non-deterministic. Hierarchical clustering algorithms on the other hand output a more informative hierarchy of clusters which can be helpful in revealing relationships between documents, do not require pre-specified parameters and are (mostly) deterministic.

In some of the earlier work on clustering for information retrieval, Voorhees (1985) investigated whether the cluster hypothesis — as mentioned above — characterises different types of document collections, and found that even though clusters of documents are more readily identified within some collections than others, cluster-based retrieval (e.g., ranking results from a relevant cluster higher) is not affected by how well clustering characterises the underlying collection.

Even though there is evidence suggesting the cluster hypothesis holds across collections, employing clustering in information retrieval can take many different forms. Typically, clustering can be applied to a collection as a whole, after which clusters can be ranked and returned in response to user queries, rather than individual documents, or clustering can be applied after retrieval, on a subset of documents extracted from the collection in response to a given query (i.e., *query-specific clustering*). Clustering can also be used to improve ranking algorithms by providing additional information regarding document content, relative to cluster or collection content. Specifically, Kurland (2008) or Liu and Croft (2004, 2008) use clustering as a way to estimate a distribution of terms relevant to a given query, by ranking clusters in response to the query and using the distribution of terms within highly ranked clusters as a basis for assessing the likelihood of documents (within or across clusters) being relevant.

Clustering algorithms are general techniques, widely discussed in the field of Machine Learning, and are not specific to information retrieval modelling nor do they provide a theoretical perspective on the relationship between queries and documents. As such, we not review clustering algorithms in more detail here. However, clustering has very practical applications in search, and we review the application of clustering in the following chapter, where we discuss interfaces used in searching the web.

2.2.4.2 Composite Retrieval

Composite retrieval has been proposed in a number of recent studies (Amer-Yahia et al., 2013, 2014; Basu Roy et al., 2010; Leroy et al., 2015) as a method of assembling information objects that adhere to certain constraints, using items retrieved from various sources. Composite information objects — called *bundles* in Basu Roy et al. (2010), or Amer-Yahia et al. (2013), *composite items* in Leroy et al. (2015), *composite objects* or *composite results* in this thesis — address the task of finding complementary items that together achieve a common goal (Amer-Yahia et al., 2014). For example, many of the studies referenced above mention finding restaurants while visiting an unfamiliar city as a plausible scenario in which composite retrieval can be applied. In such a scenario, a searcher might want to find several restaurants to try throughout their visit. A search system would typically respond with a ranking of restaurants in response to a user's query. However, a searcher in this context might have budget or time constraints, and so the restaurants retrieved by the system would need to satisfy these constraints, as well as the underlying information need. Beyond resource constraints, the searcher might have subjective preferences for different cuisines or allergy considerations. Even more, the searcher might prefer visiting different types of restaurants on successive days, and as such, returning composite objects, each containing multiple restaurant, for different days might be appropriate. These composite objects would need to be compatible (e.g., not contain the same restaurants, be within a certain geographical distance) not only meet searcher's resource constraints. In such scenarios, composite retrieval can provide searchers with results assembled within *composite objects* that are *complementary* and *representative* for searchers' information needs. Other scenarios that are typically mentioned in prior work include finding a set of compatible gadgets (e.g., finding a mobile phone and its related accessories) or finding a set of landmarks that are close to each other.

More formally, composite retrieval is defined as the task of retrieving a set of composite objects $\mathcal{S} = \{S_1, ..., S_k\}$, where a composite objects S_i is a set of items that satisfy constraints of *complementarity* and *budget*. The set of all possible items is defined as $\mathcal{I}$, where each item is uniquely identified and has a number of associated attributes (i.e., features). Typically, given two items $u, v \in \mathcal{I}$, composite retrieval assumes the existence of a similarity function $s(u, v) : \mathcal{I} \times \mathcal{I} \rightarrow [0, 1]$ (e.g., cosine similarity between feature vectors). The validity of a composite object is formally defined in Amer-Yahia et al. (2014) as:

> ***Complementarity.*** Given a property α of the items in $\mathcal{I}$, no two items in $S_i \in \mathcal{S}$ exhibit the same value for that property: $\forall u, v \in S_i, u.\alpha \neq v.\alpha$.
>
> ***Budget.*** Given a non-negative and monotone function $f : 2^{\mathcal{I}} \rightarrow \mathbb{R}$, and a budget threshold β, $\forall S_i \in \mathcal{S}$, $f(S_i) \leqslant \beta$. An example budget limit is the number of items within a composite object.

Continuing the example above, Amer-Yahia et al. (2014) suggest property α as cuisine type (e.g., Korean, Italian) or neighbourhood (e.g., Ballard or Capitol Hill in Seattle). Amer-Yahia et al. (2014) then define composite retrieval as:

> ***Composite retrieval.*** Given a set of items $\mathcal{I} = \{i_1, ..., i_n\}$, a pairwise similarity function $s(u, v)$ for each $(u, v) \in \mathcal{I} \times \mathcal{I}$, a complementarity attribute α, a budget function $f : 2^{\mathcal{I}} \rightarrow \mathbb{R}$, a budget threshold β, an integer k, and an empirically defined scaling parameter γ, find a set $\mathcal{S} = \{S_1, ..., S_k\}$ of valid composite objects that maximises:

$$\sum_{1<i<k} \sum_{u,v \in S_i} \gamma \cdot s(u, v) \;+\; \sum_{1 \leqslant i<j \leqslant k} (1 - \gamma) \cdot (1 - \max_{u \in S_i,\, v \in S_j} s(u, v))$$

Given this formal definition, it is apparent that composite retrieval has an objective function that is very similar to traditional clustering, where the quality of clustering is defined as the weighted combination of the quality of single clusters (i.e., intra-cluster cohesion), and the distance between clusters (i.e., inter-cluster separation). Unlike traditional clustering, composite retrieval does not aim for a total partitioning of the input space within clusters, but rather seeks to retrieve a (potentially small) set of k objects, that meet certain constraints. Composite retrieval can be seen as a type of *constrained clustering* (Wagstaff and Cardie, 2000), where the requirement for items within a composite object to have different values for property α is a type of *cannot-link* constraint (Wagstaff and Cardie, 2000; Wagstaff et al., 2001) (i.e., given the presence of an item with value x for property α within a composite object, no other item with value x for property α can be added to the composite object). Constraint clustering is an emerging area of Machine Learning, which, like clustering, is typically concerned with a complete partitioning of the input space (Davidson and Ravi, 2007; Davidson et al., 2007). In contrast, composite retrieval is concerned with selecting and ranking (in response to a query) a set of k *representative* composite objects.

Beyond providing a definition, Amer-Yahia et al. (2014) also formally prove that the task of composite retrieval is *NP-hard*, by reducing from the *Maximum Edge Sub-graph* problem. They propose several two-step greedy approximation algorithms to address the complexity of the task by first producing and then selecting an approximately optimal set of composite objects, and evaluate their algorithms. Extending their work, Leroy et al. (2015) adapts a constrained clustering algorithm (i.e., *Fuzzy C-Means* (Bezdek et al., 1984)) to the task of composite retrieval, by incorporating measures of *validity*, *cohesion* and *representativeness* in the algorithm objective function. They evaluate their work on three different datasets and suggest that an integrated approach, as proposed in their work, outperforms two-stage ones, as proposed in Amer-Yahia et al. (2014), leading to composite objects that are more representative of the input space.

Similar approaches to composite retrieval have been employed in recommendation tasks. The *Composite Alternative Recommendation Development (CARD)* framework proposed by Brodsky et al. (2008) investigates an approach to recommendation that returns groups of compatible items (e.g., a phone and its related accessories), that also meet certain user-defined constraints, rather than individual items in a ranked list. In a similar context, Xie et al. (2010) explore recommending variable size composite objects, and provide approximate solutions to composite recommendation; De Choudhury et al. (2010) and Basu Roy et al. (2010) also explore different approaches to composite recommendation.

In chapter 6, we study the application of composite retrieval, in a similar definition to that provided by Amer-Yahia et al. (2014), to heterogeneous web search. In our work, we explicitly look at ranking composite objects, in addition to creating and selecting representative objects. Given that in web search, topical relevance is highly correlated with a satisfactory user experience (Sanderson et al., 2010), in contrast to prior studies of composite retrieval, we use topical relevance as our main criteria of optimisation. In addition to relevance, we explicitly encode *coherence*, *topical diversity* and *vertical diversity* within our approach to object selection. Given the heterogeneous nature of our experimental collection, we also explicitly address the problem of estimating item similarity, something that is typically assumed in prior studies of composite retrieval. We then evaluate our approach against other methods of aggregating heterogeneous content on the web that are widely used in practice (e.g., aggregated search), rather than only compare between different methods of constructing composite objects, as in most prior work in this space.

In general, our work on result composition differentiates itself from prior efforts on composite retrieval in two directions: firstly, our work provides a novel user-centric perspective on result composition, something that has not been investigated in previous studies; secondly, our system-centric work explores the application of result composition to the context of *heterogeneous web search*, and addresses problems specific to this context (e.g., bridging the cross-vertical gap), something that has not been explored in prior studies.

2.3 Chapter Summary

In this chapter, we review some of the more theoretical aspects of Information Retrieval that provide a foundation for this thesis. We begin the chapter with an attempt at reviewing prior efforts at conceptualising search, from a user-centric perspective, discussing user models of search, as well as reviewing some of the vocabulary used in the literature and throughout this thesis to communicate about search (e.g., information objects, topical relevance, usefulness). We then review concepts from the system-centric perspective of information retrieval, discussing general retrieval models, evaluation metrics, as well as retrieval across collections. Finally, we review previous work on *composite retrieval*, an area of research within which our work is contextually placed.

The study of Information Retrieval is motivated by the application of theoretical models to real-world settings. For many people, the only real-world setting in which they interact with Information Retrieval is web search. In the following chapter, we review previous work on web search, from an applied perspective, discussing web search interfaces and interactions, highlighting how our work on result composition complements prior efforts.

Chapter 3

Searching the Web

The main goal of theoretical research in Information Retrieval is improving real-world search systems. By far the most popular search systems today are those that help people find information on the web. In this chapter, we review prior research related to the way people engage with web search systems, focusing on search interfaces and user search interactions, rather than algorithms or formal retrieval models. In addition, we place emphasis on heterogeneous information access, as this aspect of web search is directly related to our work. We begin the chapter by reviewing prior research on users' interactions with web search systems, and the use of interaction records in improving search experience.

3.1 Search Interactions

The primary way people interact with web search systems is by issuing queries and selecting results. All commercial search providers record these interactions, on one hand to improve the underlying search machinery in various ways (e.g., train algorithms, evaluate system performance), and on the other to enable commercial profit through targeted advertising. Research efforts directed at web search also make use of search interactions, typically recorded in laboratory-based user studies. Given that we make extensive use of search behaviour records throughout this thesis, in this section, we briefly discuss different types of search interactions, how they are recorded in both real-world and research settings, and how they have been previously used to better understand and improve search experience on the web.

3.1.1 Recording Search Interactions

To conduct research on users' behaviour when engaging in web search, various data are recorded by search systems (real or experimental) regarding both search activity (e.g., clicks) and search context (e.g., mobile or desktop). It is common that these data contain the type of event being logged, a timestamp, and other information associated with the interaction event (White, 2016, c. 2). Numerous approaches to recording search interactions can be employed, depending on whether the search system under examination is used in a laboratory or natural setting. We briefly review these approaches next.

3.1.1.1 Laboratory Settings

Laboratory-based studies of web search are studies conducted by researchers in controlled environments (i.e., by using an experimentally controlled search system) with the goal of understanding how people engage in web search. Such studies allow researchers to capture a broad range of user interactions with search systems, and also enable them to isolate the effects of extraneous variables (e.g., screen size or network latency) on search behaviour. Capturing richer interaction records is also possible in laboratory settings, such as records containing information about users' keystrokes, application switching events or eye movements, which are typically not possible in real-world settings (White, 2016, c. 2). The laboratory-based studies discussed in this section can also be described as *user studies*, which, as Kelly (2009, c. 7) point out, has become a term generally used to describe any study that involves human participants.

Different types of laboratory-based user studies have been used to investigate users' web search behaviour. Kelly (2009, c. 4) distinguish between exploratory, descriptive and explanatory studies. Exploratory studies are typically conducted when a particular phenomenon is not understood, descriptive studies usually document and describe a particular phenomenon, whereas explanatory studies examine the relationship between multiple variables with the goal of predicting or explaining a certain phenomenon (Kelly, 2009). Chapter 4 presents the outcomes of an exploratory, laboratory-based user study we conducted aimed at investigating searchers' perspective on result composition, whereas chapter 5 presents the outcomes of an explanatory user study investigating interactions between entity card properties, searchers' behaviour and their perceived task effort when engaging in web search.

In most laboratory-based information seeking studies, search scenarios and information needs are simulated and assigned by researchers to study participants, who then use an experimental search system to address their assigned information need. However, participants need not be physically co-located with the search system under investigation (i.e., in the same lab as the researchers). Indeed, in information retrieval research, it has become common to conduct user studies with participants recruited from crowdsouring platforms, such as Amazon Mechnical Turk[1] or CrowdFlower[2], who engage with an experimental search system remotely — Arguello and Capra (2014, 2016) and Maxwell et al. (2017) are some examples of such studies. The main benefits of crowdsourcing, over co-located study participants, are the ease with which participants can be recruited and filtered based on a wide range of criteria, and the low cost in conducting such studies with large numbers of participants. In contrast, lower reliability of experimental outcome measures in crowdsourcing experiments is often mentioned as a disadvantage (Alonso and Lease, 2011; Lease and Yilmaz, 2013) (as well as the more obvious disadvantages of not limiting the effects of extraneous variables or not being able to observe searchers' context in detail). In chapter 5, we conduct an experiment investigating the effect entity cards have on search behaviour, and recruit participants to engage with our experimental search system via crowdsourcing platforms. We address the outcome reliability issue from several angles, as discussed further in chapter 5.

In addition to rich interaction records, complementary methods of recording users' search experience are typically employed in laboratory-based user studies, such as interviews or questionnaires, which can help researchers develop a broader understanding of search experience, and which cannot be easily deployed in natural settings. One of the most widely used questionnaires to gauge users' perceived task effort in relation to various experimental manipulations is the *NASA Task Load Index (TLX)* (Hart and Staveland, 1988). The *TLX* is used to assess perceived workload, and measures various types of demands imposed on participants during their task, as well as self-assessed effort, frustration and performance. The *TLX* has been applied in various studies related to information seeking (Brennan et al., 2014; Speier and Morris, 2003; Stasi et al., 2011), and in chapter 5, we make use of the *TLX* to estimate the effect entity card manipulations have on users' perceived effort when searching the web.

The most common criticisms of laboratory-based studies are that they are artificial (i.e., they use simulated search tasks and information needs), and do not generalise to the real-world (Kelly, 2009, c. 4). To address these issues, collecting and analysing search interaction data generated in *natural* settings has also been

[1] https://www.mturk.com/

[2] Recently renamed to Figure Eight: https://www.figure-eight.com/

explored in information retrieval research, as briefly discussed next.

3.1.1.2 Natural Settings

Different methods of recording search interactions in a natural settings have been explored. Perhaps the most widely used approach by commercial search engines is collecting interaction events from each individual interacting with the system, recorded in what is commonly known as a *query log*. Query logs are used extensively by search providers, as discussed in the following section. Various query logs from commercial search engines have also been made available for public research[3]. In chapter 7, we make use of a large-scale query log to derive representations for heterogeneous documents within a unified feature space.

Query logs are records of entirely natural search behaviours, but typically focus on few interaction events, primarily due to practical aspects (e.g., network bandwidth or data storage costs). To address some of the limitations of query logs, researchers have also explored using specialised software that searchers can install on their devices (e.g., browser toolbars) which can track search activity in more detail (e.g., record what browser tabs are open during search) or prompt users with questionnaires when executing different types of search interactions (e.g., Fourney et al. (2017) use a browser extension to detect when searchers are looking for word definitions on the web, and insert a questionnaire in the results page for them to fill in, in an effort to enhance query log data with additional information on searchers' motivations). Methods such as longitudinal studies (e.g., Vakkari (2000)) or case studies (e.g., Ford and Graham (2016)) have also been explored, but are less common, in information retrieval research.

3.1.2 Types of Interactions

White (2016, c. 2) distinguish between two types of search interactions: ***(i)*** *atomic interaction events*, such as individual queries or clicks, and ***(ii)*** *sequences of interaction events over time*. This distinction is motivated by the different ways in which interactions can be used to improve search systems (e.g., clicks can be used to improve the ranking of results, whereas sequences of clicks can be used to personalise search over time). In our work, we interpret and report on atomic interaction events and, as such, we discuss these events and how they have previously been used to improve web search next.

[3] AOL, MSN or Yandex query logs.

3.1.2.1 Queries

In web search, queries are the only method of expressing information needs. Web search engines typically record what queries are issued by searchers to enable a range of applications, such as spelling correction or query recommendation. For example, Cucerzan and Brill (2004) show how recording user queries at scale can be used to generate corrected versions of misspelled queries. Query corrections have become common elements of modern web search interfaces. Query auto-completion is another application area of search interaction records. Query auto-completion is typically enabled on the search interface as the user is typing a query, and is intended to support more rapid expression of information needs. In web search, query auto-completion is based on query — or query prefix (Chaudhuri and Kaushik, 2009) — frequency across the searcher population, but can also be personalised to consider individuals' search history (Shokouhi, 2013). Moreover, queries issued by a population of searchers and recorded by search engines can also be used to provide query recommendations (i.e., queries that may be useful as follow-up queries to a searcher's current query).

In addition to query-related applications, such as spelling correction or auto-completion, query interaction records can be used to determine the similarity of documents (e.g., documents retrieved and clicked for a given query are likely related). In chapter 7, we make use of query interaction records, collected from a real-world web search engine, to assess document similarity across verticals.

3.1.2.2 Clicks

Clicks on results are used widely by popular web search systems as indicators of result relevance. Aggregated across the searcher population, result clicks, as collected in query logs, are some of the most important features used in improving ranking quality in commercial web search (Jiang et al., 2016). Even though clicks are not equivalent to explicit relevance judgements (assigned by professionals employed to determine a document's relevance to a given query), there is extensive evidence in the literature suggesting that clicks, aggregated across searcher populations, are useful in improving the direct ranking of documents (Agichtein, Brill and Dumais, 2006; Agichtein, Brill, Dumais and Ragno, 2006). Besides the number of result clicks, *click-through rate* (i.e., the proportion of times a result is clicked on by searchers, relative to how often it is displayed on different result pages) is a measure derived from direct user interactions commonly used to improve and evaluate ranking quality.

In addition to clicks on individual results, the click distribution at page level is also typically recorded and used to improve search quality. Searchers have been shown to scan the results page, from top to bottom, regardless of query or results relevance (Craswell et al., 2008). Click *inversions* occur when more clicks are issued, across searchers, on a given page, on results lower in the ranking (Clarke et al., 2007). Click inversions are typically interpreted as signals of poor ranking quality or result bias (e.g., White and Horvitz (2013) suggest that, in web searches related to medical conditions, results associated with more serious illnesses receive more clicks, regardless of ranking position).

In chapter 7 we make use of result clicks collected in a large-scale search interaction log to generate a unified representation space for documents across verticals. In chapter 5, we conduct an experiment where we manipulate entity card properties and measure the effect of our manipulations on search engagement and perceived workload. As discussed in this section, measuring clicks is a common and effective proxy for user search satisfaction, and as such, in the work presented in chapter 5, we report the number of clicks searchers issue on web results, in different experimental conditions, as evidence for our claims that entity card properties influence overall user search experience.

3.1.2.3 Cursor Movements

In addition to click interactions, mouse cursor movements have been used reliably in estimating which results are considered by searchers (Huang et al., 2011). Cursor movements can then be used to develop models of search attention, which in turn can be used to evaluate and optimise result presentation. However, recording cursor movements can have impact on interface responsiveness and page loading times (Huang et al., 2011), and can also generate large amounts of data that are difficult to transfer and store. Another approach to using cursor movement signals for improving search quality is by monitoring engagement with *areas of interest (AOI)* on the results page (e.g., a cursor being placed over an entity card for a long period of time). Edmonds et al. (2007) demonstrate the benefit of recording rich representations of cursor movements (i.e., thorough the monitoring of *AOIs*) over cursor position tracking, with respect to measuring task success, quantifying error conditions and assessing feature usage versus feature discovery. We use a similar approach in chapter 5, where we monitor cursor movements over areas of interest (e.g., entity cards and document surrogates) to assess different levels of user engagement with search interfaces across the experimental conditions we explored.

3.1.2.4 Other Types of Interactions

Other types of interactions, in addition to mouse clicks or cursor movements, have been used in previous efforts to model and improve search experience. One of the most widely used sources of evidence indicating result relevance is dwell time (i.e., the amount of time that searchers spend examining a particular document, after accessing it from the results page) (White, 2016, c. 2). Related to dwell time, scroll depth has also been shown as a reliable indicator of document relevance. Guo and Agichtein (2012) use dwell time, cursor movement, and scroll depth to accurately predict whether searchers are engaging with relevant or non-relevant documents.

Eye-tracking has also been used to monitor search interactions. Unlike cursor movements, eye-tracking requires the use of special hardware, and is therefore mostly used in laboratory studies of search behaviour. Eye-tracking can be used to determine which elements of the interface attract searchers' attention even in the absence of mouse movements or clicks (e.g., on devices that do not make use of cursors) (Buscher et al., 2010), or to measure searchers' pupillary response, as an indicator of relevance (Gwizdka and Zhang, 2015). Differences between searchers with respect to visual attention patterns on the results page have also been observed (Dumais et al., 2010), suggesting that eye-tracking can be useful for personalising search results as well.

All types of search engagement recorded in user studies or real-world systems are generated by users interacting with a system interface. Web search interfaces have been studied extensively over the past decades and, as such, we briefly review prior research related to search interfaces next.

3.2 Search Interfaces

Hearst (2009) begin their seminal review of search interfaces by clarifying that the task of a search user interface is to *"aid users in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and in keeping track of the progress of their information seeking efforts"*. In contrast to their ambitious goal, search user interfaces, as seen in web search systems today, are relatively simple, enabling users to ***(a)*** type keywords in a box to express what they want to find, and ***(b)*** scan a vertical list of results. In addition to simple, web search interfaces have not changed much in their (somewhat short) history: Hearst (2009, p. 1) remark on web search interfaces remaining *"nearly identical"* in recent years, whereas White (2016, p. 24) highlight limited innovation in the way results are presented in web search, over the past few *decades*. Perhaps the most significant change to web search interfaces, over the past few years, has been the integration of diverse content within a unified results space, through aggregated search or entity cards (as briefly discussed in chapter 1 and detailed later in this chapter).

Why have web search interfaces remained so consistently simple[4] over the past decades? Hearst (2009) suggest as reasons for this simplicity that ***(i)*** search is a means towards a goal, and as such any distractions from the goal should be minimised; ***(ii)*** search is a mentally intensive task, which involves reading and thinking about information, thus fewer distractions lead to a more usable interface; and ***(iii)*** a wide range of people use web search, therefore the interface needs to be navigable by people with a variety of skills and abilities.

However, even apparently simple interfaces, as those used in web search today, contain a range of complex elements that influence users' interaction with the underlying system. Relatively simple display properties of search results — such as the colour of links, amount of white space around results, or the number of results on the page — have been shown to influence the way searchers interact with the system. As such, we review prior efforts investigating results presentation strategies in more detail next.

3.2.1 Search Results Presentation

Most modern web search systems display results in a vertical list, typically containing ten items, on a page that is known as the search engine results page or *SERP*. Each result is associated with a document from the web, and usually displays the document title (i.e., a blue link), a short summary of the document (also known as a *snippet*), and potentially other metadata about the document, such as author, date, URL or rating. The representation of a document on the results page is commonly called a *document surrogate*.

Surrogates shown on the results page play an important role in searchers' assessment of document relevance. If the quality of a surrogate is poor (e.g., misleading title or uninformative summary), it is unlikely that its corresponding document will be visited by searchers, regardless of document relevance. Many studies have been conducted over the past decades exploring the properties of document surrogates and their role in making search interfaces more

[4] Simple in contrast with older search system — for example, those used by librarians, patent officers or legal professionals — which commonly required specific querying syntax (e.g., Boolean syntax or command language) and returned highly specialised results that only a small group of trained professionals might interact with.

usable. Clarke et al. (2007) compared 10000 pairs of results in an effort to understand why certain surrogates receive more clicks than others (in the case of results returned for the same query and shown on the same page, in proximity to each other). Their findings suggest that ***(i)*** highlighting query terms in the surrogate summary is beneficial, that ***(ii)*** when the document title contains query terms, they do not need to appear in the summary as well, and that ***(iii)*** the length of URLs displayed in surrogates should be minimised, and their relationship with the query emphasised. Overall, their findings suggest that highlighting query terms in the various components of document surrogates can lead to more clicks.

Similar findings have been reported in other studies looking at query-biased document summaries[5] (i.e., document summaries that contain query terms or fragments of text related to query terms). For example, Tombros and Sanderson (1998) and White et al. (2003) report higher performance in terms of precision, recall and total time taken to find relevant information (i.e., searchers find more relevant documents faster) when query-biased summaries are used; Varadarajan and Hristidis (2006) report that users assign higher ratings to query-biased summaries than those produced by commercial search engines at the time (which were less biased towards the query). Query-biased summaries have become a default element of document surrogates in modern web search (Hearst, 2009, c. 5).

Highlighting query terms in result summaries has also been explored in a number of studies, over the past decades. Highlighting query terms refers to modifying the display properties of text in order to make it more noticeable (e.g., boldface, different colour text). Landauer et al. (1993) report that highlighting terms in document surrogates increases the number of results scanned by searchers, as well as reduces the number of non-relevant results examined. More recently, Yue et al. (2010) found that searchers click more on results that had query-terms in the surrogate summary displayed in boldface. Baudisch et al. (2004) suggest that highlighting query-terms in more detailed surrogates, that present a document overview rather than just a document summary and its title, is also preferred by searchers.

The length of surrogate summaries has been studied in detail as well. Cutrell and Guan (2007) show that longer snippets (6-7 lines) improve performance for information tasks, but degrade performance for navigational tasks, a finding further supported by Kaisser et al. (2008). Maxwell et al. (2017) reiterate, suggesting that searchers broadly prefer longer summaries, and perceive them as more informative, but perform equally well in terms of identifying relevant documents

[5] Also known as *query-oriented* or *keyword-in-context* summaries

regardless of summary length (as manipulated in their experiment).

Other ways of enhancing surrogates, by incorporating images or non-textual elements, have been studied also. Teevan et al. (2009) investigate different sized surrogates that contain visual elements (i.e., images and logos) extracted from their associated documents. Their findings suggest visual summaries of documents support browsing behaviour while being significantly smaller than text-only surrogates (which can be important in contexts such as mobile search), and that visual summaries are particularly valuable in re-finding relevant information. Capra et al. (2013) augment surrogates using images extracted from their associated documents, and conduct a large-scale user study to examine the effects image-augmented surrogates have on effectiveness (as determined, for example, by searchers' accuracy in finding relevant content) and efficiency (e.g., task duration). In their study, they look at document surrogates in isolation, as well as document surrogates in context (i.e., on a results page). In addition, they investigate the *"goodness"* of images with respect to underlying document content, and its effect on experimental outcome measures. Their findings suggest that at individual level, augmenting surrogates with images provides very small benefits in term of relevance judgement accuracy and duration, whereas at page level, augmenting surrogates with images has no effect on measures of effectiveness or efficiency, compared to text-only surrogates. Their findings also suggest that, on result pages that are explicitly diversified (e.g., result pages returned for ambiguous queries), searchers are more precise at identifying relevant documents when augmented surrogates are displayed. Overall, their study highlights different tradeoffs in the use of image-augmented surrogates in different situations. Similar findings have been reported in studies prior to Capra et al. (2013), which generally report mixed results regarding image-augmented surrogates, with some studies suggesting benefits of using images with or as document surrogates (Jiao et al., 2010; Li et al., 2008), whereas other studies finding less clear benefits over traditional text-based surrogates (Aula, Khan, Guan, Fontes and Hong, 2010).

Augmented surrogates have also been explored in domain specific search contexts. BioText (Hearst, Divoli, Guturu, Ksikes, Nakov, Wooldridge and Ye, 2007) is a search engine that provides a novel way for researchers to access bioscience literature. In addition to text summaries and titles, BioText enhances document surrogates with figures extracted from the underlying document. Extensive research using the BioText search interface (Divoli et al., 2010; Hearst, Divoli, Ye and Wooldridge, 2007) found that searchers have a strong preference for image-augmented surrogates, when searching for research literature.

In contrast to image-augmented surrogates, which can be viewed as a merging of heterogeneous results, composite results contain multiple surrogates, of various types, each associated with a different document and aggregated around a common topic, rather than a common document. Although our work is not directly concerned with the display and presentation of composite results, it it possible to conclude from the studies reviewed in this section that presentation aspects are crucial for the usability of search systems, and that studying the presentation of complex result aggregates is necessary for search systems that make use of such aggregates. We indicate indicate presentation aspects as salient matters for the future study of result composition in chapter 9.

In addition to the display properties of document surrogates, numerous aspects of the search engine results page have been studied, such as the visual display of relevance estimates associated with documents (White et al., 2007), or the integration of interface elements that support *browsing* as part of the search process. We review prior studies investigating browsing support in web search in the following section.

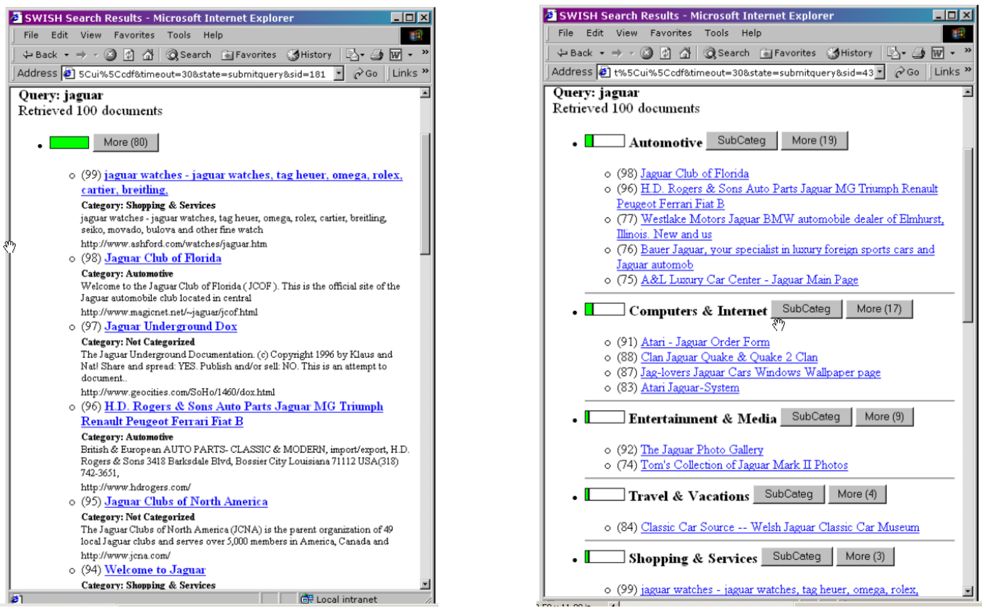
3.2.2 Assistance in Browsing Results

Browsing and searching are two components of the information seeking process. Hearst (2009, c. 8) distinguish browsing from searching by noting that searching tends to produce new, previously unseen listings of information objects (i.e., rankings of results) that have not necessarily been retrieved together before, whereas browsing is a sequence of scan-and-select operations which restrict the information space to pre-defined groups of information objects (e.g., clicking on a result restricts the information space from a ranking of possibly relevant documents to the content of one document).

In web search, support for browsing behaviour has been offered (at interface level) through grouping search results into pre-defined categories (i.e., result categorisation) or into arbitrary groups of inter-related documents (i.e., result clustering). Given that result composition is a type of result grouping, we review some of the more widely known approaches to categorisation and clustering next.

3.2.2.1 Categorisation

Categorisation is a system of applying one or multiple meaningful labels to information objects in a way that reflects their topical focus. Typically, the set of possible labels is limited (and relatively small, in contrast to the number of information objects) and pre-defined in order to provide structure to the inform-


SWISH Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address 5Cui%5Ccdf&timeout=30&state=submitquery&sid=181
Go Links
More (80)
Local intranet

Query: jaguar
Retrieved 100 documents

- ○ (99) jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling,
  Category: Shopping & Services
  jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling, seiko, movado, bulova and other fine watch
  http://www.ashford.com/watches/jaguar.htm
- ○ (98) Jaguar Club of Florida
  Category: Automotive
  Welcome to the Jaguar Club of Florida (JCOF). This is the official site of the Jaguar automobile club located in central
  http://www.magicnet.net/~jaguar/jcof.html
- ○ (97) Jaguar Underground Dox
  Category: Not Categorized
  The Jaguar Underground Documentation. (c) Copyright 1996 by Klaus and Nat! Share and spread: YES. Publish and/or sell: NO. This is an attempt to document..
  http://www.geocities.com/SoHo/1460/dox.html
- ○ (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
  Category: Automotive
  British & European AUTO PARTS- CLASSIC & MODERN, import/export, H.D. Rogers & Sons 3418 Barksdale Blvd, Bossier City Louisiana 71112 USA(318) 742-3651,
  http://www.hdrogers.com/
- ○ (95) Jaguar Clubs of North America
  Category: Not Categorized
  The Jaguar Clubs of North America (JCNA) is the parent organization of 49 local Jaguar clubs and serves over 5,000 members in America, Canada and
  http://www.jcna.com/
- ○ (94) Welcome to Jaguar
  Category: Shopping & Services

(a) Experimental condition in Dumais et al. (2001): results displayed in a linear view with category labels shown, not grouped into individual categories.

SWISH Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address t%5Cui%5Ccdf&timeout=30&state=submitquery&sid=43
Go Links
SubCateg
More (19)
More (17)
More (9)
More (4)
More (3)

Query: jaguar
Retrieved 100 documents

- **Automotive**
  - ○ (98) Jaguar Club of Florida
  - ○ (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
  - ○ (77) Westlake Motors Jaguar BMW automobile dealer of Elmhurst, Illinois. New and us
  - ○ (76) Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob
  - ○ (75) A&L Luxury Car Center - Jaguar Main Page
- **Computers & Internet**
  - ○ (91) Atari - Jaguar Order Form
  - ○ (88) Clan Jaguar Quake & Quake 2 Clan
  - ○ (87) Jag-lovers Jaguar Cars Windows Wallpaper page
  - ○ (83) Atari Jaguar-System
- **Entertainment & Media**
  - ○ (92) The Jaguar Photo Gallery
  - ○ (74) Tom's Collection of Jaguar Mark II Photos
- **Travel & Vacations**
  - ○ (84) Classic Car Source -- Welsh Jaguar Classic Car Museum
- **Shopping & Services**
  - ○ (99) jaguar watches - jaguar watches, tag heuer, omega, rolex,

(b) Experimental condition in Dumais et al. (2001): results displayed in a linear view but grouped into individual categories, with category labels shown.

Figure 3.1: Interface used in Dumais et al. (2001) to explore result categorisation. Findings suggest that grouping results within categories can lead to less time taken by searchers to locate relevant information, and that searchers have a strong subjective preference for category grouping, over linear displays, as long as category labels are shown on the page. From Dumais et al. (2001), as seen in Hearst (2009, p. 179).

ation space. Hearst (2009, c. 8) differentiates between three types of category systems available on various web search interfaces: *flat*, *hierarchical* or *faceted*.

Flat categorisation Figure 3.1 shows an experimental interface used in Dumais et al. (2001) to study (flat) result categorisation in a web search setting. In their work, web results are assigned one of ten top-level category labels (e.g., Computer & Internet, Travel & Vacations, Shopping & Services). Their findings suggest that grouping results within categories can lead to less time taken by searchers to locate relevant information, and that searchers have a strong subjective preference for category grouping of results, as long as category labels are shown on the page. More detailed analysis revealed that time-gains were achieved especially for queries where relevant results were displayed below rank 20 in the linear results page, in which case categorisation typically moved relevant documents upward in the ranking.

In a similar study, Käki and Aula (2005) design a text categorisation algorithm that extracts the most frequent words and phrases from a list of results, and then use this list of labels in a search interface, displayed next to a results ranking, as a flat category structure. Selecting a label in the interface filters the ranking to show results containing specific words or phrases. They compare their interface to a ranking only interface, and suggest that allowing searchers to filter using categories not only leads to faster and more accurate use of web search, but also leads to more positive attitudes towards the search system. In a follow-up longitudinal study, Käki (2005) follow 16 searchers over two months and show that categories are successfully used as part of users' search habits — especially when more results are needed to satisfy an information need, as in the case of exploratory or undirected search. Kules and Shneiderman (2008) conduct a similar study and find that searchers consider the categorised search interface as more appealing than a linear interface.

In the context of heterogeneous web search, the different vertical tabs, available on most modern web search interfaces, can be viewed as a type of flat categorisation assistance on the basis of document type, rather than content.

Hierarchical categorisation Another approach to structuring information is by defining a hierarchy of concepts related to the information space (e.g., the table of contents in a thesis). Hierarchical categorisation has been explored in the context of web search, to support user browsing. Much like flat categorisation, search interfaces making use of hierarchical categories typically displayed a tree of categories in addition to search results, on the same page. As an example, although not in the context of web search, Landauer et al. (1993) explore the use of hierarchical categorisation in a novel search interface, which they suggest can increase search accuracy and speed.

Faceted categorisation Assigning documents a single categories is problematic, given that most documents discuss multiple topics. Faceted categorisation addresses this problem, by assigning documents a *set* of category labels (e.g., Computer & Internet, Web Search) or attribute labels (e.g., date created) and means for manipulating these labels within the interface. Even though consistently found as helpful, across studies, it still remains *"an open question whether these [categorisation approaches] will eventually be widely and regularly used on the open-domain Web"* (Hearst, 2006). However, categorisation approaches are common in specialised search interfaces, such as e-commerce or other types of vertical search — for instance, both the Google and Bing *image* search interfaces make use of categorisation tools to help users browse and easily filter results.

3.2.2.2 Clustering

Clustering refers to the automatic grouping of documents or results according to some measure of similarity (Hearst, 2009, c. 8). Unlike categorisation, where document categories are assigned by hand or by an algorithm, clustering is driven by similarities between document features (e.g., the terms they contain). The assignment of documents to categories also typically requires a relatively limited set of categories (Hearst, 2009, c. 8) (which tend to become very large on large-scale collections of documents, such as the web), whereas clustering can be applied effectively, regardless of collection size.

Clustering, as discussed in the previous chapter, can be performed at collection level (in which case, search interfaces can enable the exploration of collection clusters) or at ranking level (in which case, documents retrieved for a given query are grouped, in the search interface, by their similar features). Although not directed at the web, one of the most widely known search systems enabling the exploration of collection-level document clusters is Scatter/Gather (Cutting et al., 1992). The goal of Scatter/Gather is to group documents (in a collection) within coherent groups (i.e., clusters) and present document groups (and their automatically generated textual summaries) to searchers. Searchers can then select clusters (i.e., gather) that appear to contain relevant content, based on their textual summaries, and re-cluster selected documents (i.e., scatter) based on their feature similarity. Through this iterative clustering process, the contents of a collection can be explored at incremental levels of detail. Pirolli et al. (1996) show that using Scatter/Gather on a large collection of documents induces a more coherent view of a text collection, leads to searchers expressing their information needs through a richer vocabulary, and also communicates the distribution of relevant documents across collection clusters.

At ranking level, and in the context of web search, a number of search engines have made use of result clustering over the past decades. Hearst (2009, c. 8) mention *Clusty.com*[6] and *iBoogie.com*[7] as examples. Browsing assistance through clustering is, however, not widely adopted in modern web search, as far as general web results are considered, but is widely adopted with respect to heterogeneous results (e.g., images or videos). Indeed, most web search interfaces today assemble diverse content, displayed in what can be described as clusters of heterogeneous documents, on a unified results page. This paradigm has become known as aggregated search. Given that result composition is also a type of clus-

[6] Now renamed to Yippy.com: https://yippy.com/

[7] No longer available.

tering heterogeneous results, we review prior studies on the use of aggregated search systems in the following section.

3.2.3 Aggregated Search Results

Aggregated search refers to the task of merging results from different collections (i.e., verticals) within a unified results page. Arguello (2017) provide a comprehensive review of research efforts in the field of aggregated search — with a particular focus on user search behaviour and interaction with aggregated search systems in Arguello (2017, c. 5).[8] Given that aggregated search is the *de facto* method of integrating diverse content within a unified results page in modern web search, and that our work is also concerned with merging results from different sources within the same results space, we review prior efforts investigating aggregated search interfaces in more detail next.

3.2.3.1 Presentation of Aggregated Search Results

Most aggregated (web) search interfaces are defined by three design aspects:

(i) They integrate top-k results from verticals estimated as relevant within the ranking of web results, with the goal of *"showcasing"* verticals that might be useful to searchers. An alternative to this approach is allowing searchers to access diverse content in a tabbed interface.

(ii) They display results from the same vertical in a block (with results stacked vertically or horizontally), clearly delimited from the list of web results, as opposed to interleaving results, regardless of type, as in the case of merging federated search rankings.

(iii) They display only relevant verticals, and the position at which vertical blocks are merged within the unified ranking depends on their relevance to the query (i.e., the page is *dynamically* assembled in response to a query). An alternative would be to show results from all verticals, or show vertical results at fixed positions.

Many of these design decisions are supported by prior studies exploring what searchers expect from aggregated search systems. Multiple studies have found that searchers are more likely to click and mark vertical results as useful when they are blended within a unified results listing (Arguello et al., 2012; Sushmita

[8] We follow their structure in presenting our review of prior work in this section.

et al., 2009; Turpin et al., 2016). Sushmita, Joho, Lalmas and Villa (2010) use query-log analysis to show that searchers often click on diverse results for non-navigational search, when diverse content is shown in a blended ranking of results. In addition, Bron et al. (2013) investigate a search system that allows users to switch between an aggregated view of results (i.e., results merged within a unified ranking) and a source-specific view (i.e., tabbed view), and find that the blended view increases awareness about information available across sources. These findings suggest that merging results from different sources in a unified ranking, rather than exposing diverse content solely through a tabbed interface, can be beneficial to search user experience.

Results from different sources can be interleaved directly, not at block-level, like in the case of federated search. However, Arguello (2017, p. 86) point out, to date, no single study has directly compared block-level versus item-level interleaving of results and user preference on this matter. Nevertheless, the argument for block-level merging can be made by referring to principles of pattern recognition, as studied in psychology (Koffka, 2013; Palmer, 1992), which suggest that items displayed together (i.e., in the same block, delimited from other items) are perceived as a group and are more easily interpreted. In search, group displays are assumed to help users more quickly identify vertical results that are relevant and more easily disregard vertical results which are not relevant. Grouping of results based on their type can also be seen as a form of categorisation or clustering (as discussed previously in this chapter). Given the benefits of categorisation on searcher effectiveness and their overall perceived satisfaction with search systems — e.g., as shown in Dumais et al. (2001) — it is reasonable to assume that block-level merging is preferable to item-level merging of heterogeneous results.

A number of studies experimented with aggregated search systems that always presented results from different verticals, regardless of query, and found that searchers rated systems poorly compared to other systems than only provided tabbed access to diverse content (Arguello and Capra, 2012; Turpin et al., 2016). Chen et al. (2015) found that searchers report greater levels of satisfaction when using search interfaces that display content extracted from relevant verticals only, the effect being stronger for visually salient verticals (e.g., images).

3.2.3.2 Factors Affecting Aggregated Search Use

Various factors related to ***(i)*** vertical block (e.g., relevance, position in the ranking), ***(ii)*** search task (e.g., task complexity) or ***(iii)*** the searcher (e.g., perceptual speed) have been shown to affect searchers' decision to engage with aggregated search results merged within a unified results page.

Vertical relevance has been shown to increase engagement with vertical results in a number of studies (Arguello and Capra, 2012; Chen et al., 2015; Turpin et al., 2016) as well as the number of eye fixations on vertical results (Liu et al., 2015). In addition, Chen et al. (2015) suggest that searchers perceive higher levels of satisfaction when relevant vertical results are displayed higher in the ranking (e.g., rank 1 versus rank 3). With respect to vertical block presentation, Sushmita, Joho, Lalmas and Villa (2010) and Sushmita, Piwowarski and Lalmas (2010) uncover click biases in favour of visually salient vertical results (e.g., images or videos), whereas Liu et al. (2015) report greater levels of visual attention given to salient verticals, regardless of vertical relevance. Arguello (2017) conclude that there is complex interaction between vertical relevance, position and presentation factors, with each factor being influential on searchers' decision to engage with the results page, either through clicks or eye fixations.

Other factors related to search task or user characteristics have been shown to affect engagement with aggregated search results. Specifically, Arguello et al. (2012) found that searchers who are assigned more complex tasks engaged more with vertical results, but only when vertical results were merged within a unified results page, rather than exposed through a tabbed interface. With respect to user characteristics, Turpin et al. (2016) found that users with low perceptual speed (i.e., *"speed in comparing figures and symbols [...] or carrying out other simple tasks involving visual perception"* (Ekstrom et al., 1979)) took longer to complete their tasks when using a blended aggregated search interface, rather than a tabbed one. Overall, these findings suggest that a *"one size fits all"* approach to aggregating results might not be optimal across searchers (Arguello, 2017).

3.2.3.3 Spillover Effects in Aggregated Search

Many studies in the space of aggregated search have explored how results from one source affect user engagement with results from other sources — an effect known in the literature as *"spillover"*. Studies investigating the spillover effect typically focus on *ambiguous* queries (i.e., that have multiple senses — for example, the query *"tesla"* which can refer to the scientist Nikola Tesla, or Tesla the electric car). In response to ambiguous queries, search systems can either diversify results, returning items relevant to all possible meanings of an ambiguous query, or predict which meaning of the query to return results for. Arguello and Capra (2016) found that top web results are typically diversified in response to ambiguous queries, whereas top vertical results are skewed towards a particular meaning of an ambiguous query. Therefore, it is possible in aggregated search that vertical results address a different meaning of a searcher's ambiguous query than their intended meaning, which can then influence their decision to interact with other web or vertical results present on the page.

Several studies report that the spillover effect is stronger for visually salient verticals, such as images (Arguello, 2015; Arguello and Capra, 2012, 2014; Arguello et al., 2013). These studies suggest that searchers are more likely to notice visually salient verticals which, if reflecting a different query-meaning than the searcher's intended one, can lead to searchers assuming the entire page contains results that are not relevant. In addition, the spillover effect is stronger when vertical results are shown higher in the ranking rather than the middle (Arguello and Capra, 2014) or the right side of the ranking (Arguello and Capra, 2016). Arguello and Capra (2014) also suggest that enclosing vertical results by a visual border has a subtle moderating effect on spillover. Arguello and Capra (2016) complement this finding, suggesting no spillover effects when images are enclosed by a border and displayed to the right of the results ranking.

Overall, findings in the context of aggregated search further support our study of result composition. On one hand, numerous studies investigating aggregated search show that searchers subjectively prefer and engage more with heterogeneous results when results are merged within a unified page, therefore suggesting that diversity on the results page is an important aspect of web search. Our work on result composition attempts to enhance result page diversity in a more structure approach. On the other hand, delimiting diverse results from the general ranking is shown to have positive effects in certain contexts, which suggests that investigating novel ways of aggregating and presenting diverse results (such as result composition) can be beneficial for overall search experience. Such novel methods of merging complex aggregates on the results page have been attempted previously in web search, in the form of entity cards. Therefore, we review formal studies related to the use of entity cards in web search next.

3.2.4 Entity Cards

In addition to merging different types of results within a unified page, most modern web search engines also display entity cards[9] in response to entity oriented

[9] Entity cards are also known in the research literature as *knowledge graph results* or *knowledge graph panels*. We use the former term to refer to such objects on one hand because the term *knowledge graph* is also the name of the knowledge base used by Google to assemble these information objects, and therefore is more of a proprietary brand name rather than a generic term, and on the other hand, the term *entity card* is useful in distinguishing between the actual search page

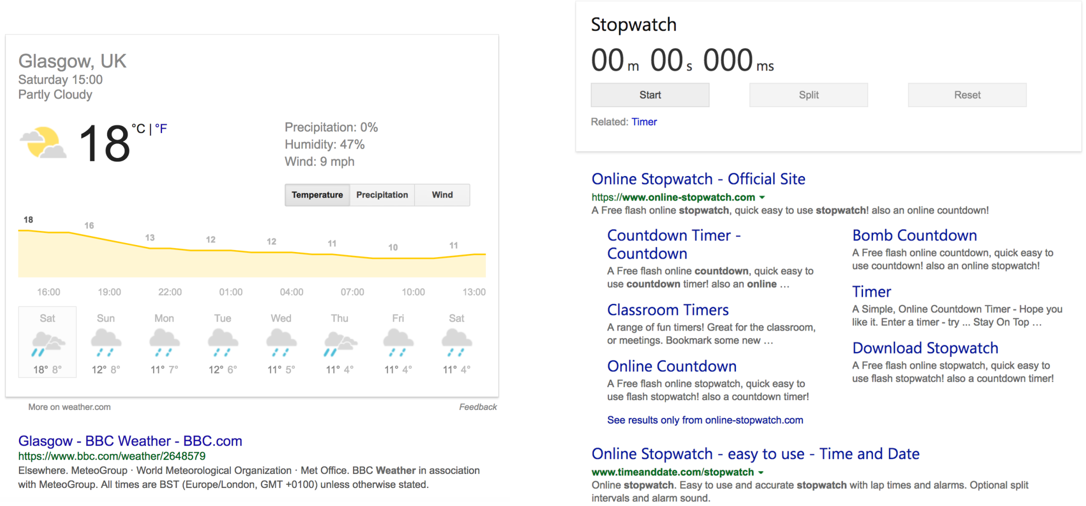
queries. Entity cards are intended to enhance search experience in several ways: ***(i)*** they help searchers navigate diversified results, ***(ii)*** provide a summary of relevant content directly on the results page and ***(iii)*** support exploratory search by highlighting relevant entities associated with a given user query. Entity cards are, in our definition, a type of composite result.

Entity cards have become very popular. Bota et al. (2016) report that entity cards are displayed in more than two thirds of searches triggered by ambiguous queries. In a less formal setting, Enge (2017) suggest that rich answers (i.e., entity cards and in-line answers) are shown on the majority of Google's results pages. Although popular, entity cards have received relatively little attention in the research community[10], compared to aggregated search. In perhaps the first effort exploring the use of entity cards in web search, Navalpakkam et al. (2013) conduct a laboratory study aimed at understanding the effect of *rich informational panels* (i.e., entity cards) on eye fixations and mouse movements. They report that searchers' flow of attention is different when an entity card is present on the results page, compared to the widely believed top-down linear examination of search results. Specifically, they show that searchers' attention is captured by entity cards (as reflected in both eye fixation and mouse hovers over the card) and that, when the entity card is relevant, this leads to searchers terminating their tasks faster. In addition, they report that a majority of searchers focus their attention on the top ranked result first, but are equally likely to scan further down the ranked list or to shift their attention to the entity card immediately after. Overall, their study concludes that as search pages become increasingly complex, with the addition of non-linear interface elements, searchers' behaviour and examination patterns change significantly as well.

In the context of mobile web search, Lagun et al. (2014) study the effect answer results (e.g., entity cards and in-line answers (Chilton and Teevan, 2011)) have on searchers' satisfaction by varying their presence and relevance. They monitor eye fixations and viewport (i.e., the part of a search page in focus on a mobile device), and record explicit user satisfaction assessments. Their findings suggest scrolling past the entity card or increased time spent viewing the page below the answer result as a clear, measurable signal of user dissatisfaction with the results page. Also in a mobile setting, Williams et al. (2016*a*) find that relevant entity cards contribute to *good abandonment* (Chuklin and Serdyukov, 2012) in web search.

interface element and the underlying data representation model used to construct these objects (i.e., a knowledge graph or knowledge base).

[10] Entity cards as interface elements have been relatively understudied in the research community, whereas their underlying data structures (i.e., knowledge graph) have been studied extensively.



Glasgow - BBC Weather - BBC.com
https://www.bbc.com/weather/2648579
Elsewhere. MeteoGroup · World Meteorological Organization · Met Office. BBC **Weather** in association with MeteoGroup. All times are BST (Europe/London, GMT +0100) unless otherwise stated.

Online Stopwatch - Official Site
https://**www.online-stopwatch.com** ▾
A Free flash online **stopwatch**, quick easy to use **stopwatch**! also an online countdown!

Countdown Timer - Countdown
A Free flash online **countdown**, quick easy to use **countdown** timer! also an **online** …

Classroom Timers
A range of fun timers! Great for the classroom, or meetings. Bookmark some new …

Online Countdown
A Free flash online stopwatch, quick easy to use flash stopwatch! also a countdown timer!

Bomb Countdown
A Free flash online countdown, quick easy to use countdown! also an online stopwatch!

Timer
A Simple, Online Countdown Timer - Hope you like it. Enter a timer - try ... Stay On Top …

Download Stopwatch
A Free flash online stopwatch, quick easy to use flash stopwatch! also a countdown timer!

See results only from online-stopwatch.com

Online Stopwatch - easy to use - Time and Date
www.timeanddate.com/stopwatch ▾
Online **stopwatch**. Easy to use and accurate **stopwatch** with lap times and alarms. Optional split intervals and alarm sound.

(a) Weather answer embedded in the results page assembled by Google in response to the query *"weather"*.

(b) Stopwatch answer embedded in the results page assembled by Bing in response to the query *"stopwatch"*.

Figure 3.2: Example interface elements used by modern web search engines to address information needs directly on the results page. Accessed April 2018.

More recently, Hasibi et al. (2017) investigate dynamic entity summarisation for entity cards (i.e., generating query-dependent factual summaries of entities) and show that searchers prefer dynamic summaries over static ones.

In chapter 5, we report on our study investigating the use of entity cards in web search. Like previous studies, we manipulate card relevance and report on measures of search engagement. Unlike previous studies, in our work we manipulate card vertical diversity and card coherence, in addition to relevance, and also report on searcher-assessed measures of perceived task workload.

3.2.5 Other Search Interface Elements

In addition to entity cards, modern web search engines make use of a variety of interface elements to address searchers' information needs directly on the results page — known as in-line answers or answers (Chilton and Teevan, 2011). Figure 3.2 shows examples of in-line answers displayed on Google and Bing.

Much like entity cards, in-line answers have not been studied widely. Chilton and Teevan (2011) conduct a first study explicitly investigating the use of in-line answers at scale, and report that in-line answers "cannibalize" clicks from the web results ranking. In addition, they suggest repeat search behaviour (i.e., using the same query multiple times without issuing a click) as an indication of in-line answer usefulness. Their findings support previous reports by Li et al. (2009) which found that in a majority of searches that are abandoned (i.e., searches in which no click is issued on any of the ranked results) different types of in-line answers are shown on the results page.

In connection with search abandonment, but in a mobile context, Williams et al. (2016*a*) show that *good abandonment* is driven by various search interface elements, including in-line answers. In a more detailed follow-up study, again in a mobile search context, Williams et al. (2016*b*) investigate the effect different types of in-line answers have on searchers' satisfaction. They report that satisfaction rates vary across in-line answers that have similar abandonment rates on their associated pages, showing that answer-related abandonment is not always a type of *good abandonment*. They hypothesise that answers' ability to fully address searchers' queries (e.g., factoid queries are more easily addressed by in-line answers than broader informational needs) is a factor in determining answer-related user satisfaction. Fourney et al. (2017) also briefly discuss the use of in-line answers with respect to addressing linguistic information needs.

3.3 Chapter Summary

In this chapter, we review prior studies related to users' search interactions and how these interactions can be used to improve web search experience. We begin the chapter by reviewing different methods of recording users' search behaviour, and how these interaction records are used in different research or commercial contexts to address the many challenges associated with web search. We then review prior studies of web search interfaces, highlighting several findings that suggest grouping results on the search interface, using various methods, can support searchers' browsing behaviour. With respect to heterogeneous information access on the web, we discuss prior efforts investigating aggregated search which suggest that merging diverse content within a unified results page can improve overall user experience. In addition, we highlight how limited prior studies of entity cards are in contrast to their wide adoption by modern web search engines, which provides additional motivation for our work.

Given this overview of web search interactions and interfaces, we present our first contribution to understanding heterogeneous information access through web result composition in the following chapter.

Chapter 4

Exploring Result Composition from the Users' Perspective

Aggregating results from heterogeneous sources and presenting them in a unified interface (i.e., aggregated search) has become standard practice for most commercial web search engines. In addition, *composite objects* containing results originating from different sources have started being integrated in result pages (e.g., entity cards or in-line answers). In this chapter, we report on our study of *result composition* from the users' perspective. We conducted an exploratory user study where 40 participants were required to manually generate *composite objects* that satisfy various information needs, using pre-retrieved heterogeneous results. Our main objective was to analyse the contents and characteristics of user-generated composite objects. The outcomes reported in this chapter show that users generate composite objects on common *subtopics*, centred around *pivot* documents, and that a clear hierarchy of object properties, with respect to user preference, is not easily determined. Research presented in this chapter is based on previously published work available in Bota et al. (2015).

4.1 Introduction

The past three decades have seen an explosion of information on the web, in terms of both quantity and diversity of content. Modern web search engines aggregate results from heterogeneous information sources (i.e., *verticals*) in order to satisfy diverse user information needs (Zhou et al., 2012). Different approaches to aggregating information on the web have been proposed and studied, such as *federated search* or *aggregated search* (Diaz et al., 2010). In general, these approaches focus on merging results from multiple homogeneous text collections into one ranked list or inserting blocks of results from different heterogeneous information sources within a standard search engine results page (SERP). As the web is becoming more diverse, it is important to return to users more structured information objects, containing information extracted from different sources. Consider the following user information need: *"travelling to Austria"*. Finding all the information that satisfies this need typically involves submitting several queries, each focusing on different aspects of travelling, such as directions, accommodation or points of interest. Composition of web results aims to address the limitations of merging homogeneous blocks within heterogeneous rankings, as in the case of aggregated search, and return to users heterogeneous results organised within composite objects, each object containing results from multiple verticals and satisfying different aspects of their information need.

Prior research on composite retrieval has primarily focused on either analysing the algorithmic aspects of generating composite objects or formalising the desirable properties of composite objects (Amer-Yahia et al., 2013, 2014). Limited effort has been dedicated to understanding result composition from the user perspective. For example, how do users manually aggregate information? What are the most important criteria for users when assessing the quality of information aggregates? Answering these questions can align future developments of heterogeneous web search systems to users' expectations. Therefore, we pursue this line of research and conduct an exploratory user study which allows us to investigate the contents, characteristics and topical focus of user-generated composite objects. Broadly, in our study, participants were shown search results originating from various heterogeneous sources and were required to manually generate composite objects that satisfy their information needs. After building objects, different assessments of object characteristics were collected from users, in order to understand their preferences regarding composite object properties. The experiment presented in this chapter aims to answer the following questions:

(***RQ1***) Do users agree with each other with respect to the *subtopics* they form composite objects on?

(***RQ2***) How do users aggregate information to construct composite objects? How vertically diverse are the composite objects generated by users?

(***RQ3***) Which composite object characteristics are most important to users? What are the interactions between these characteristics?

Although our study did not consist of traditional search interactions (i.e., formulating queries followed by scanning a ranked list of results), but rather *composi-*

tion interactions, the analysis of user-built composite objects and accompanying assessments offers insight into user expectations from such objects, which can inform the development of modern web search systems. The main contributions of the study presented in this chapter consist of: ***(i)*** a first exploration aimed at understanding how users manually construct composite objects; and ***(ii)*** insight into user preference with respect to composite object content and characteristics.

4.2 Prior Work

This work builds on two broad areas of prior research that we briefly review here: composite retrieval and user behaviour in heterogeneous information access.

4.2.1 Composite Retrieval

Responding to information retrieval queries by presenting composite items has been proposed and investigated in a number of recent papers (Amer-Yahia et al., 2013; Angel et al., 2009; Deng et al., 2012; Guo and Ishikawa, 2011). Many of the above papers have provided contributions on the theoretical side, studying the complexity of evaluating queries with constraints, and proposing different algorithmic formulations. Amer-Yahia et al. (2014) studied the complexity of generating composite objects with constraints (such as budget), and proposed different algorithmic formulations to solve the problem of object generation. In many ways, composite retrieval on the web is similar to both result clustering and result categorisation, as discussed in chapters 2 and 3, which have been shown to improve both searchers' effectiveness in identifying relevant content, as well as their subject assessments of search experience.

4.2.2 User Behaviour in Information Access

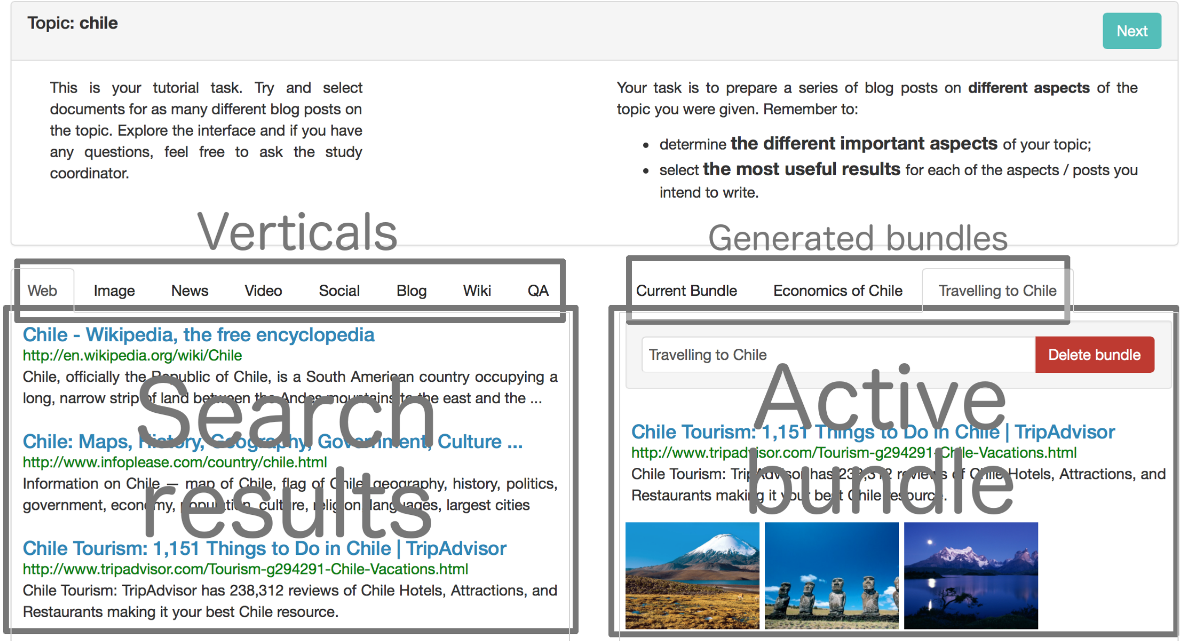
Prior work has looked at user search behaviour in detail, mainly focusing on behaviour in traditional search environments — some examples include Jansen and Pooch (2001); Rose and Levinson (2004); Spink et al. (2002). Our work aims to go beyond traditional search scenarios and investigates user behaviour in a result composition setting. User behaviour in exploratory collaborative web search has been studied in work related to ours, specifically focused on modelling user search processes (Yue et al., 2014).

Significant effort has been made to understand user behaviour in an aggregated search setting (Arguello and Capra, 2012, 2016; Bron et al., 2013; Zhou et al., 2012, 2013). In particular, Arguello and Capra (2012) — later extended in Arguello and Capra (2016) — investigate different aspects related to results page coherence that influence search behaviour in an aggregated search scenario. User preference of result aggregation methods is investigated by Bron et al. (2013), where it is shown that users prefer heterogeneous blocks blended into traditional lists over tabbed displays when trying to obtain an overview of the available information space. This indicates that aggregation of results within a unified results page can be beneficial, and motivates our investigation of more elaborate aggregation techniques in the form of result composition.

Although search behaviour in aggregated search contexts has been investigated extensively over the past few year, as discussed in detail in section 3.2.3, user engagement with web results in a composite search setting has not been studied at all. As such, in this study we aim to investigate user behaviour in a result composition scenario and analyse manually generated composite object to gain insight into user expectations regarding the structure, contents and characteristics of composite objects.

4.3 Experimental Methodology

Our objective was to determine the contents and characteristics of user-generated composite objects. In light of this objective, we ran a laboratory-based user study where participants constructed composite objects using search results originating from different verticals and assessed their own objects in terms of several criteria. For the study, we employed 40 participants (17 female, 23 male) with an average age of 24 ($\mu = 24.75$, $\sigma = 5.42$). Each participant was compensated with £10 for their help. Half of the participants were undergraduate students at the time of the study, 17 were postgraduates, and 3 were in active employment. In terms of background, 60% of them had obtained, or were interested in obtaining, a technical degree. Participants were given 4 different *composition* tasks and were asked to construct composite objects, as described below, using results cached from several existing search engines. Each task was completed in approximately 15 minutes ($\mu = 15.55$, $\sigma = 8.80$). We used 40 different topics, collected from public aggregated search collections (Demeester et al., 2013; Zhou et al., 2012). Topics were assigned randomly to participants, the only constraint being that each topic needed to be assigned to exactly 4 different users. Overall, each of the 40 participants performed 4 separate tasks, for a total of 160 result composition tasks.



Topic: chile

This is your tutorial task. Try and select documents for as many different blog posts on the topic. Explore the interface and if you have any questions, feel free to ask the study coordinator.

Your task is to prepare a series of blog posts on **different aspects** of the topic you were given. Remember to:

- determine **the different important aspects** of your topic;
- select **the most useful results** for each of the aspects / posts you intend to write.

Chile - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Chile
Chile, officially the Republic of Chile, is a South American country occupying a long, narrow strip of land between the Andes mountains to the east and the ...

Chile: Maps, History, Geography, Government, Culture ...
http://www.infoplease.com/country/chile.html
Information on Chile — map of Chile, flag of Chile, geography, history, politics, government, economy, population, culture, religion, languages, largest cities

Chile Tourism: 1,151 Things to Do in Chile | TripAdvisor
http://www.tripadvisor.com/Tourism-g294291-Chile-Vacations.html
Chile Tourism: TripAdvisor has 238,312 reviews of Chile Hotels, Attractions, and Restaurants making it your best Chile resource.

Travelling to Chile

Chile Tourism: 1,151 Things to Do in Chile | TripAdvisor
http://www.tripadvisor.com/Tourism-g294291-Chile-Vacations.html
Chile Tourism: TripAdvisor has 238,312 reviews of Chile Hotels, Attractions, and Restaurants making it your best Chile resource.

Figure 4.1: Web based interface used by our participants to build composite objects. In our experiment, we used the term *"bundle"* in what we assumed to be a more user-friendly synonym for the term *composite object*.

4.3.1 Task Design

To reflect complex information needs suited for result composition, participants were asked to imagine that they are bloggers, preparing a series of blog posts on different aspects related to a given topic (e.g., living in India). Their choice of aspects (or *subtopics*) to focus on was unrestricted, however the subtopics were required to be distinct. For each subtopic, they were instructed to select the *most useful search results* — that they considered to be the most helpful for writing the blog post — and place them in a *composite object* containing multiple heterogeneous search results. Although they were required to title the objects they created, they were not required to write an actual blog post, only to pre-select search results that might be useful for writing it.

During the study, participants were first shown a description of their general task, that of constructing composite objects, and were guided through the system interface. The interface allowed participants to explore eight different verticals (*General Web* or *GW*, *Image*, *News*, *Video*, *Social*, *Blog*,*Wiki*, *Q&A* — shown in the *Verticals* box in figure 4.1), each containing 50 pre-retrieved results, for the topic they were assigned. All text-based results were presented using a standard web search engine style, namely a highlighted title above a short document summary. Hovering over any search result displayed a tooltip window that contained additional information about the result. For example, hovering over *Video* results played a 10 second extract from the actual video result (without sound).

Figure 4.1 shows the system interface used in our experiment. The verticals were presented as part of a tabbed section which occupied the left half of the interface. Search results were displayed in a traditional search engine layout — text-based documents were displayed in a ranked list of results, whereas *Images* and *Video* were displayed in a grid of thumbnails. The right section of the interface was occupied by the *composition* area, where participants could create objects by adding documents from any of the verticals, and assign titles to the objects. There were no restrictions imposed on the number or size of composite objects participants were required to construct. After the *composition* phase, participants were required to assess each of their objects in terms of the five criteria described below. They were also required to assign relevance labels (*non-relevant*, *relevant*, *highly relevant*, *key* and *navigational*) to each of the documents contained by the objects they constructed.

Finally, participants were presented with pairs of their own composite objects in a side-by-side view and asked to make a preference judgement between the two objects. When indicating preference, they were also required to indicate the motive behind their preference in both free-form text and by indicating one of the five object-level criteria as being most influential on their choice (options *None* and *Overall* were also available). Pairwise object-level preference assessments allow us to determine which composite object characteristics are the most frequent indicators of user preference, and also determine the degree to which our participants effectively assess object characteristics independently.

4.3.2 Object-level Characteristics

After generating composite objects, participants were required to rate them on five different criteria, using a five point scale (*very*, *fairly*, *somewhat*, *slightly*, *not at all*). Our choice of evaluation criteria was inspired by previous work on evaluating search results in context (Bailey et al., 2010*a*,*b*; Golbus et al., 2014), where it has been shown that certain aspects of search relevance are difficult or impossible to judge in isolation. In line with previous work, we focus our evaluation of composite object characteristics on the five criteria described in table 4.1.

4.3.3 Limitations

We chose to present results in a traditional layout to maintain user interface familiarity. Results added to composite objects were presented using the same type

Relevance	–	Are the documents in your object relevant to the topic?
Diversity	–	Does the object contain a diverse set of documents?
Coherence	–	Are the documents in your object related and about one specific aspect of the topic?
Freshness	–	Is the object interesting and current?
Overall	–	How satisfied are you with your object?

Table 4.1: Criteria used for the evaluation of object-level characteristics. Participants were asked to rate each of the objects they constructed with respect to these criteria, using a five point scale (*very*, *fairly*, *somewhat*, *slightly*, *not at all*).

of layout because there is limited understanding of how this type of objects can be presented effectively on a search results page, without confusing searchers. Because we are interested in the contents and characteristics of composite objects, not in the actual search interaction, we believe the presentation of composite objects only minimally influenced their contents.

Verticals were presented in a fixed tabbed interface, in a predefined order (i.e., *General Web* occupied the first tab, followed by *Image*, *Video* and other types of results). Even though the ordering of verticals may have had a biasing effect on document selection, we believe it was minimal: on one hand because participants were explicitly encouraged to explore all verticals before engaging in the result composition task; on the other hand, our interaction logs show that study participants explored an average of 7 ($\mu = 7.27$, $\sigma = 1.58$) verticals per tasks, suggesting they were at least acquainted with the top results in the majority of verticals displayed on the experimental interface.

One of the limitations of our study is the fact that participants were unable to explore actual documents, but were constrained to generating composite objects using search results. Because we wanted to minimise the cognitive load on our participants, as well as keep task duration manageable, we chose not to give participants access to actual documents. Even so, we consider result snippets to be highly representative of actual documents, and partially mitigated this limitation by allowing users to view highlighted document snippets, and for *Video* results, short excerpts (10 seconds) from the actual material.

4.3.4 Experimental System

Search results for all topics were cached on our server. The *General Web*, *Image* and *News* results were retrieved using the Bing Web Search API; the *Video* vertical was populated using the YouTube API; the *Social* vertical was populated using the Twitter API; all other verticals were populated using the Google Custom

Search API over specific websites[1] that matched a certain vertical profile, sourced from Demeester et al. (2013).

4.4 Experimental Results

Our aim was to examine result composition from two different perspectives: on one hand, we intended to analyse the contents and structure of composite objects by looking at the types of documents they contain; on the other hand, we were interested in determining how users assess composite objects in terms of the five criteria we outlined in table 4.1. In particular, we were interested in determining which criteria are most important to users. Broadly, the main questions we aim to answer through this study are:

- What documents do user-generated composite objects contain? What *roles* do these documents play within the composite object?
- What properties of composite objects do users consider most important?

The next sections of this chapter describe our findings. Section 4.4.1 provides a brief analysis of object subtopic agreement among users, followed by an analysis of composite object contents and of potential *roles* documents perform within composite objects. Section 4.4.2 presents our results regarding user assessed object-level characteristics.

4.4.1 Composite Object Contents

In this section, we present our analysis of composite object contents. In total, our 40 participants constructed 519 composite objects, using 2982 unique documents sourced from all verticals used in our study.

4.4.1.1 Subtopics

Participants were asked to construct objects using pre-retrieved results, focusing each of their objects on a specific aspect, or subtopic, of the topic they were given. The choice of subtopic was unrestricted as long as it was pertinent to the general topic. They were also required to assign a free-form text title to each of the objects they constructed. Therefore, we define the *subtopic* of a composite object as *the facet of a specific topic around which an object is focused, as reflected by its title*. We

[1] For example, the *Blog* vertical was populated using a Google custom search engine over the following domains: *wordpress.com*, *medium.com*, *tumblr.com* and *blogspot.com*.

employ this specific definition of object subtopic because we intend to determine objects that are similar, in terms of their topical focus, and analyse their common contents and common properties. We use assigned object titles as proxies for evaluating the topical similarity of composite objects, across participants.

To determine the semantic similarity of titles (and hence, that of their corresponding composite objects), we used a directional similarity metric — i.e., similarity of title t_i with respect to title t_j — inspired from Corley and Mihalcea (2005). Object titles are tokenised and part-of-speech tagged, and because they are relatively short (mean length of $\mu = 2.55$ words, $\sigma = 1.61$) we annotate each non-stopword in a title with a subset of its most likely synonyms, as determined by WordNet (Miller, 1995). Starting with one title, for each word in its word class set (e.g., *noun* or *adjective*), we determine the most similar word – using the Jiang and Conrath (1997) similarity – from the corresponding class set in the other title. We use the word similarity scores, weighed by the idf[2] scores of corresponding words and normalised by the idf scores of starting words, to compute the directional similarity of two titles, as elaborated in more detail in Corley and Mihalcea (2005). We use the directionality of the metric to determine whether two composite objects are mutually about the same subtopic. Given two composite objects $c_i, c_j \in C$, with their respective titles t_i, t_j, where C is the set of all user-generated objects on a given topic, we assume that two composite objects focus on the same subtopic if their titles mutually have the highest semantic similarity score:

$$max(\{\ \forall c_k \in C,\ i \neq k \mid sim(\mathbf{t_i}, \mathbf{t_k})\ \}) = sim(\mathbf{t_i}, \mathbf{t_j})$$
$$max(\{\ \forall c_k \in C,\ j \neq k \mid sim(\mathbf{t_j}, \mathbf{t_k})\ \}) = sim(\mathbf{t_j}, \mathbf{t_i})$$

This measure of title similarity is used to determine participant agreement on object subtopic: we want to determine whether participants building composite objects on a given topic choose to focus their objects on similar subtopics, as this is interesting in itself, but also helps us identify similar objects across users, which further allows us to describe more general patterns in their common contents and common properties.

On average, our study participants built 3 composite objects of search results for every topic they were assigned ($\mu = 3.21$, $\sigma = 0.90$, averaged across all users and all study tasks) – with each object being focused on a distinct subtopic, as per our instructions. Because we want to determine different levels of subtopic

[2] The British National Corpus was used to derive document frequency counts. www.natcorp.ox.ac.uk

		Proportion of participants per topic involved in determining subtopic agreement		
		100%	***75%***	***50%***
Proportion of objects about same subtopic		12%	14%	16%
Proportion of topics with	at least ***1*** common subtopic	32%	75%	90%
	at least **2** common subtopics	0%	32%	85%
	at least **3** common subtopics	0%	5%	60%

Table 4.3: Subtopic agreement based on semantic similarity of titles assigned to composite objects by study participants.

agreement among users — e.g., 2 out of 4 (50%) participants generating objects on the same topic agree on at least one common subtopic — we can restrict set C to include only objects generated by a subset of users. The results in table 4.3 show that there is a general tendency for user agreement on at least one common subtopic for a given topic — e.g., half the users built composite objects on at least two common subtopics, for 85% of the topics used in our study. As an example, for the topic *"living in India"*, the following objects, constructed by different users, were determined to be similar based on their titles: *"Cost of living in India"*, *"average prices in india"* and *"Employment, Cost and Standard of Living"*.

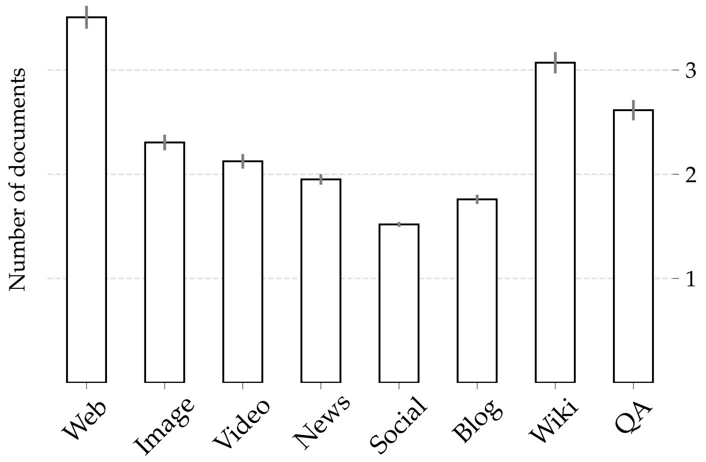


Figure 4.2: Mean number of documents of different types, averaged over all composite objects created by participants. Vertical error bars represent standard error of the mean.

4.4.1.2 Vertical composition

One of our main research objectives was to analyse the types of documents user-generated composite objects contain. Figure 4.2 shows the average vertical structure of a composite object. Averaged across all composite objects created by participants, the mean number of documents contained by an object is 7 ($\mu = 7.82$, $\sigma = 5.57$), with, on average, 3 ($\mu = 3.02$, $\sigma = 1.53$) unique verticals being represented within each object. On average, *General Web (GW)* is the most represented vertical in user-generated objects, which is not unexpected given its intended, indeed highly optimised, purpose of satisfying wide ranges of information needs. The multimedia verticals (i.e., *Image* and *Video*) are also well represented, with roughly two documents, on average, in each object. This is a reflection of the vertical orientation of topics used in the study and a potential click bias towards this type of media (Sushmita, Joho, Lalmas and Villa, 2010), but also suggests the importance of vertical diversity for users when assembling composite objects. Even more, user inclination towards vertical diversity is suggested by the number of verticals represented in each object, given that more than 80% of composite objects contain more than two verticals, as shown in figure 4.3). Note that we instructed participants to select for composition only the results they found most useful to their task, and did not explicitly encourage vertical diversity in our instructions. A more detailed view on the distribution of verticals within user-generated composite objects is shown in figures 4.4 and 4.5.

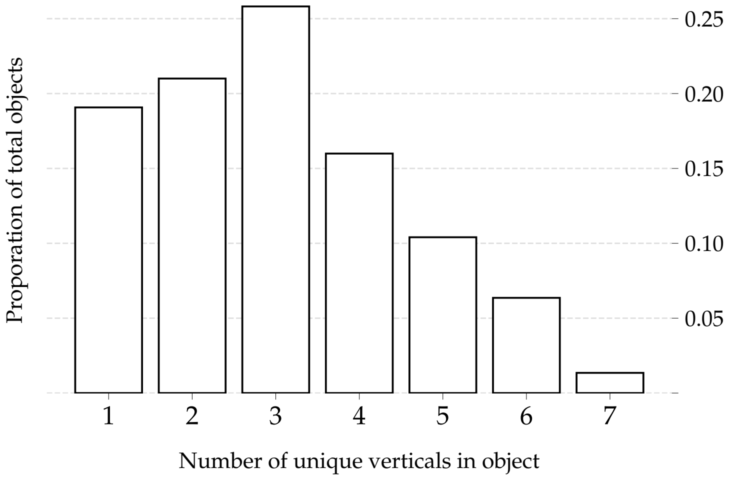


Figure 4.3: Number of unique verticals in all objects created by participants.

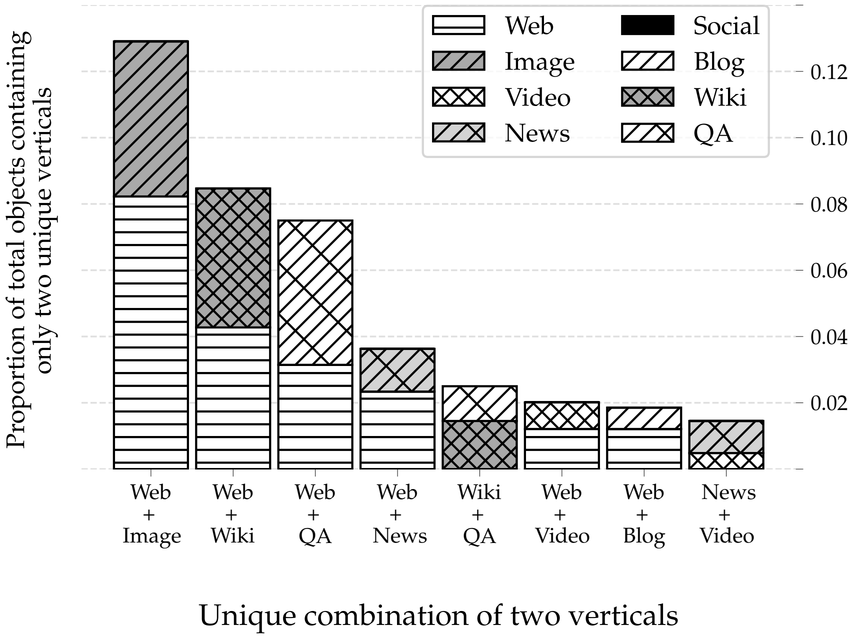


Figure 4.4: Frequency of vertical combinations in the subset of composite objects containing documents from exactly ***two*** distinct verticals. Each bar shows the frequency of vertical combinations, whereas the patches are proportional to individual verticals within the subgroup.

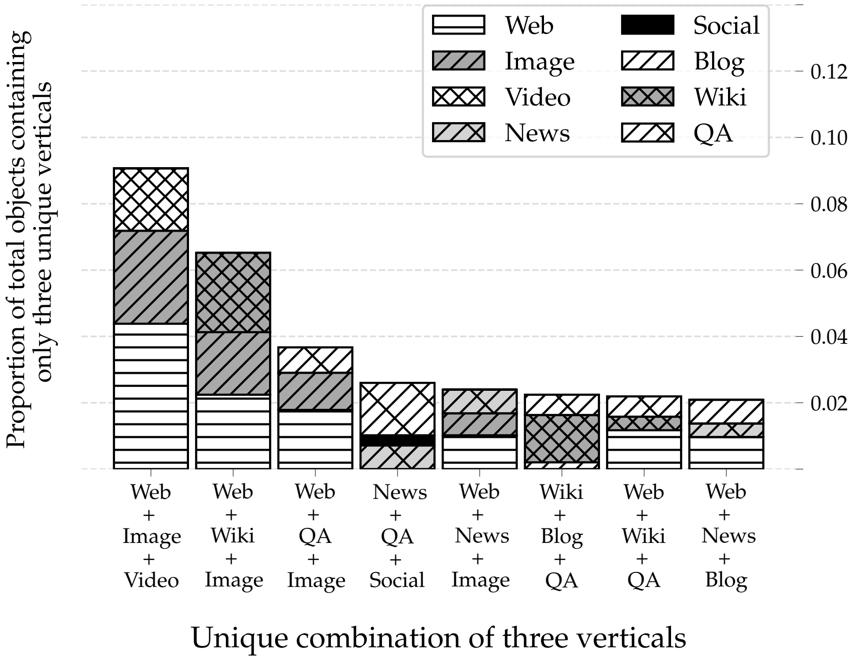


Figure 4.5: Frequency of vertical combinations in the subset of composite objects containing documents from exactly ***three*** distinct verticals. Each bar shows the frequency of vertical combinations, whereas the patches are proportional to individual verticals within the subgroup.

4.4.1.3 Document roles

Prior work on composite retrieval explored algorithmic formulations for the construction of composite objects that involved attaching complementary search results to a central item (Amer-Yahia et al., 2014). Inspired by this approach, we analyse different *roles* that documents have within user-generated composite objects. We distinguish between two separate roles:

- ***Pivot documents***: a document or set of documents that appear in multiple composite objects on the same subtopic of a given topic, where subtopic agreement is established as previously described in section 4.4.1.1.

- ***Ornament documents***: a document or set of documents which originate from different verticals than an object's pivot document set, and which are not explicitly assessed as irrelevant by the author of the object.

Given our definitions of document roles, it is clear that not all documents within composite objects (e.g., irrelevant documents) are assigned a role label. Although our definitions do not necessarily reflect all possible relationships between documents, we focus on those document roles that are, perhaps, more distinctive and of interest to our research goals.

To assess the effect of *pivot* documents on composite object structure, we used the same methodology as described in Section 4.4.1.1 to compute the semantic similarity of *pivot document titles* to composite object titles, assigned to objects by their authors. We also analysed pivot documents' explicit relevance assessments.

Composite object that were determined as being related, through the similarity of their titles, were used to identify and analyse pivot documents. In total, 47% of the objects we determined as being about the same subtopic contained at least one pivot document (i.e., they had at least one document in common). On average, the related objects contained one pivot document (mean $\mu = 1.27$, $\sigma = 0.56$), with the largest pivot document set containing 4 documents. Our results show that pivot document titles are significantly (determined using a one-tailed *t-test*: $t(1, 435) = 31.764$, $p < 0.01$) more similar to object titles, and are also significantly ($t(1, 435) = 70.831$, $p < 0.01$) assessed by users as being more relevant than other documents within composite objects.[3] Overall, roughly 21% of the composite objects constructed by our study participants are focused around at least one pivot document which is central to the composition process and determines the object title and its overall topical focus.

[3] In both tests, we use the *Welch–Satterthwaite* (Satterthwaite, 1946) method to approximate degrees of freedom, given unbalanced groups.

In addition, we analysed the vertical origin of pivot document sets. It is perhaps not surprising that the majority of pivot documents originated from *General Web* (61% of pivots) and *Wiki* (23% of pivots) verticals, considering their broader scope and perhaps higher semantic load than multimedia, *QA* or *Blog* documents.

To determine the *ornament* make-up of composite objects, we analysed similar objects, that contained at least one pivot document, and extracted documents that originated from other verticals than the pivots, and which were assessed by users as not completely irrelevant. Our intention was to determine which documents provide value through *"composition"* rather than explicit relevance, by complementing object content. Our results show that similar objects, which contain at least one pivot, have an average of 4 documents ($\mu = 3.60$, $\sigma = 4.26$) that match our ornament definition, originating from the *Image* (23%), *Video* (19%), *Wiki* (17%) and *QA* (16%) verticals.

	Pivot type	
	GW	***Wiki***
GW		24.6%
Image	23.5%	31.1%
Video	21.3%	18%
News	7.1%	1.6%
Social	1%	6.6%
Blog	9%	11.5%
QA	17.4%	4.9%
Wiki	19.7%	

Table 4.5: Vertical origin of ornament documents in composite objects containing *General Web* or *Wiki* pivot documents.

We investigated the relationship between pivots and ornaments by analysing the vertical distributions of ornaments in composite objects with different types of pivots. In particular, we focused on the two main pivot types (*GW* and *Wiki*) and analysed the types of ornaments associated with these types of pivots. Table 4.5 shows that pivot type affects ornament diversity to an extent. There is a weak trend that suggest a complementarity relationships between different types of pivots and ornaments. *Images* appear more frequently in composite objects centred around *Wiki* pivots, whereas *QA* documents complement *General Web* pivots more often. Table 4.6 also shows that, as object vertical diversity increases, ornament documents (e.g., *Video*) tend to be assessed as more relevant, whereas *GW* are assessed as being less relevant. This suggests that relevance becomes distributed across different verticals in composite objects that are more diverse.

	Number of verticals in object	
	Two verticals	***Three verticals***
GW	3.87	3.67
Image	3.21	3.35
Video	3.23	3.65
News	2.95	3.16
Social	2.67	2.20
Blog	3.59	3.40
QA	2.56	2.65
Wiki	3.55	3.58

Table 4.6: Mean document relevance, by vertical, in multi-vertical composite objects. Our data suggest that as object vertical diversity increases, ornament documents (e.g., *Video*) tend to be assessed as more relevant, whereas *GW* are assessed as being less relevant.

4.4.2 Composite Object Characteristics

In addition to examining object contents, our research aims include analysing user assessments of composite object characteristics (i.e., *Relevance, Diversity, Coherence* and *Freshness*). Our intention is to assess how users rate the objects they constructed with respect to these four criteria and determine a potential hierarchy of characteristics, as well as uncover correlations among these characteristics.

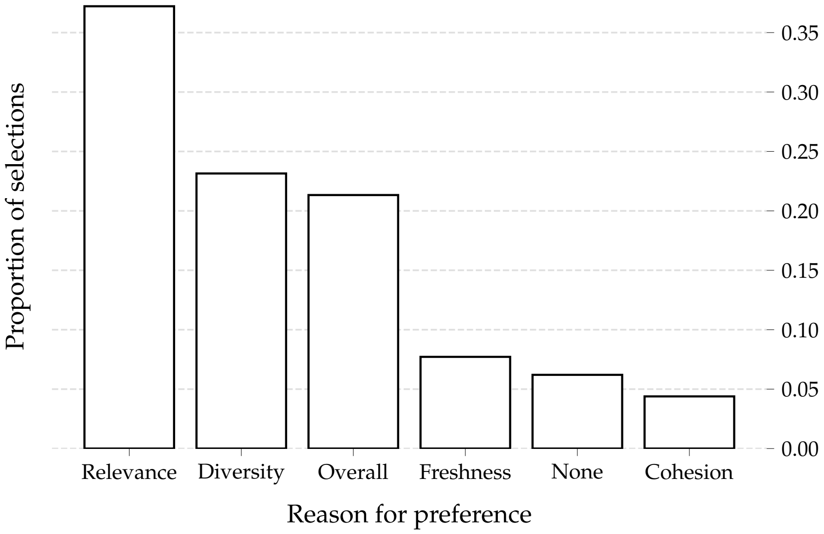


Figure 4.6: User indicated most influential criterion for object preference.

To this end, after constructing their objects, participants were required to make explicit preference judgements between all possible object pairs (i.e., all possible pairs generated from the objects they created) and motivate their preference by

	Pearson's R	
	All	***Chosen***
Relevance	0.332	0.496
Cohesion	0.228	0.432
Diversity	0.334	0.487
Freshness	0.208	0.213
Overall	0.453	0.454

Table 4.7: Correlation of criteria assessments with object pairwise preference

indicating one of the criteria above as influential on their choice (options *None* and *Overall* were also available). As shown in figure 4.6, based on pairwise assessments, *Relevance* (37%) and *Diversity* (23%) were most frequently indicated as the influential criteria for user preference. In addition, 21% of the participants indicated *Overall* (i.e., all of the criteria) as the motivation for their preference.

We also analysed the correlation between users' pairwise preference and user assigned object ratings — as mentioned in section 4.3.1, in addition to pairwise preference assessments, participants were required to explicitly rate the objects they created, on a five-point scale, with respect to the *Relevance*, *Coherence*, *Diversity*, *Freshness*, and *Overall* quality of the object. In particular, we examined whether pairwise preference of object A over object B correlates with higher criteria assessments for object A than for object B for ***All*** criteria – i.e., if object A is preferred over object B, does object A have higher ratings on ***All*** of the five criteria than object B. Additionally, we examined this correlation taking into account the criterion ***Chosen*** by users as most influential — i.e., if *Relevance* is indicated by the user as the influential criterion for preference of object A over object B, is the user assessed *Relevance* of A higher than that of B? Table 4.7 shows that there is modest correlation between preference and user assessment of object characteristics, even in cases where the specific characteristics are indicated as the reason behind object preference. In roughly 50% of the cases, even though users explicitly mention *Relevance* as the motive behind their preference, they prefer the object assessed as less relevant. Although part of this can be due to noise in our data, this highlights the difficulty of determining the characteristics of composite objects that are most important to users.

Finally, we investigated the correlation between different pairs of object characteristics, shown in table 4.8. It is worth noting that there is strong correlation between several object characteristics, the strongest correlation being that between *Relevance* and *Overall* ($Pearson's\ R = 0.630$). This suggests that object characteristics are difficult to assess independently and, combined with results

	Relevance	*Diversity*	*Coherence*	*Freshness*	*Overall*
Relevance	–	0.272	0.538	0.334	0.630
Diversity	0.272	–	0.144	0.485	0.478
Coherence	0.538	0.144	–	0.250	0.548
Freshness	0.334	0.485	0.250	–	0.537
Overall	0.630	0.478	0.548	0.537	–

Table 4.8: Linear correlation (*Pearson's R*) of explicit user ratings of composite object properties and overall quality.

mentioned above, collectively contribute to user preference. Even so, *Relevance*, *Coherence* and *Diversity* are correlated with both user preference and among themselves, which suggests their combined importance to users when assessing the quality of composite objects.

4.5 Discussion and Conclusions

In this study, we analysed the contents and characteristics of composite objects manually constructed by study participants, using pre-retrieved, heterogeneous Web documents. Our primary interest was to determine how composite objects are generated by users with regard to their topical focus and document composition, and how participants assess composite objects with respect to *relevance*, *coherence* and *vertical diversity*. Our results suggest the following trends:

***(RQ1)* Do users agree with each other with respect to the subtopics they form composite objects on?** With respect to our first research question, we found that there is agreement between users on the topical focus of composite objects, given a certain topic. This suggests that composition of search results can be focused on distinctive facets of given topics and search systems looking to integrate composite object within their result pages can explore the possibility of constructing objects around popular subtopics of a searcher issued query.

***(RQ2)* How do users aggregate information to construct composite objects? How vertically diverse are the composite objects generated by users?** With respect to our second research question, we observe there is a trend for composite objects to contain central documents, or *"pivots"*, that are more relevant and reflect the object's topical focus. These documents tend to originate from verticals with higher semantic load (such as *General Web* or *Wiki*). Furthermore, ornament documents, which tend to be less relevant than pivots and more vertically diverse, are also included within user-generated composite objects. In our case, *Image*, *Video* & *QA* verticals are popular origins of ornament documents. The above results suggest that one effective strategy for result composition is to first select a small subset of key pivot documents, and then explicitly attach other documents that complement the pivots, in order to enhance coverage, complementarity and vertical diversity of objects. This suggests that, even though relevance is crucial, less relevant documents are explicitly attached to composite objects by participants as they can provide value by complementing pivots and by providing diversity.

***(RQ3)* Which composite object characteristics are most important to users? What are the interactions between these characteristics?** With respect to our last research question, although our results do not establish a clear hierarchy of object characteristics, we make similar findings as prior work (Bailey et al., 2010*a*,*b*) and determine that *relevance*, *coherence* and *diversity* are important to participants, but are difficult to assess independently. Corroborated with the above-mentioned insights on vertical diversity, this implies that, although explicit relevance is crucial to users, composition of diverse results can generate additional value.

In terms of future work, many open questions remain. Our work so far has investigated user generation of composite objects, but has not explored the usefulness of these objects in a traditional search scenario. Further work is needed to understand the complex aspects related to the presentation and integration of composite objects within a traditional results page, as it is possible that presentation factors can influence both the perceived relevance and user interaction with this type of result aggregates. The following chapter of this thesis attempts to study the role of composite objects (i.e., entity cards) in a traditional search scenario in an effort to address some of these questions.

4.6 Chapter Summary

Retrieving results from heterogeneous sources and presenting them in a unified interface is a difficult problem. Aggregated search has become the most prevalent method for selecting and displaying results from different sources on a singular page, but is limited to merging blocks of homogeneous content within a heterogeneous ranking. While more complex result aggregates, containing and highlighting the connections between heterogeneous documents, can be constructed, it is unclear what content these aggregates should contain or, indeed, what types of results searchers typically expect to find within such aggregates. Even more, the importance of various properties of such aggregates — *relevance*, *coherence* and *diversity* — to users, and their interplay, is another understudied aspect of aggregating heterogeneous content within complex information objects to be integrated in a unified results page.

This chapter presents an exploratory user-study where 40 participants were asked to manually generate composite objects, using pre-retrieved heterogeneous results from 8 different verticals. The main goal of this study was to analyse the contents and characteristics of user-generated composite objects.

Our results show that ***(i)*** participants tend to agree on the topical focus of composite objects, ***(ii)*** composite objects tend to be structured around a central document or set of documents (i.e., *pivots*) and ***(iii)*** determining a hierarchy or weighting of object properties with respect to their importance to users is not obvious. This evidence supports our assertion that documents that play a central role within composite objects have a greater impact on object usefulness and that, although a clear hierarchy of object properties is not obvious, a wider range of object properties, in addition to relevance, is important to users.

The work discussed in this chapter represents a first study of users' perspectives on result composition for web search. Further work is required to understand the role of composite objects in a more traditional, *query-to-ranking* web search scenario, rather than the *composition* scenario we employed in our study. The following chapter of this thesis attempts this task by investigating the effect a type of composite object (i.e, entity cards) has on search behaviour and user perceived workload, in a traditional web search scenario.

Chapter 5

The Effect of Entity Cards on Search Behaviour and Workload

In addition to merging results of different types (e.g., images, videos or news items) into a ranked list, modern search engines have also started displaying *entity cards*[1] on the results page. Entity cards are intended to enhance search experience in several ways: ***(i)*** they help searchers navigate diversified results, ***(ii)*** provide a summary of relevant content directly on the results page and ***(iii)*** support exploratory search by highlighting entities associated with a given query.

In this chapter, we present a large-scale crowd-sourced user study, with more than 500 unique searchers, aimed at investigating the effect entity cards integrated into search result pages have on user behaviour and perceived workload. In our definition, entity cards are a type of composite object (i.e., entity cards are a type of composite object but not all composite objects are entity cards) and we study their influence on search behaviour, under the assumption that our findings regarding entity cards generalise to other potential types of composite objects integrated in a similar way on search engine result pages.

Our findings suggest that the presence of entity cards has an effect on both the way users interact with search results and their perceived task workload. Furthermore, by manipulating entity card properties (*content*, *coherence* and *vertical diversity*), we uncover different effects of card properties on measures of search

[1] Entity cards are also known in the research literature as *knowledge graph results*. We use the former term to refer to such objects on one hand because the term *knowledge graph* is the name of the knowledge base used by Google to assemble these objects, and therefore is more of a proprietary brand name rather than a generic term, and on the other hand, the term entity cards is useful in distinguishing between the actual search interface element and the underlying data representation model used to construct these objects (i.e., a *knowledge graph* or *knowledge base*) — whether a knowledge base is used to create such objects is not necessarily a primary aspect of our research.

behaviour and workload. Our study contributes an in-depth analysis of the effects entity cards have on user interaction with modern web search interfaces. Research presented in this chapter is based on previously published work available in Bota et al. (2016).

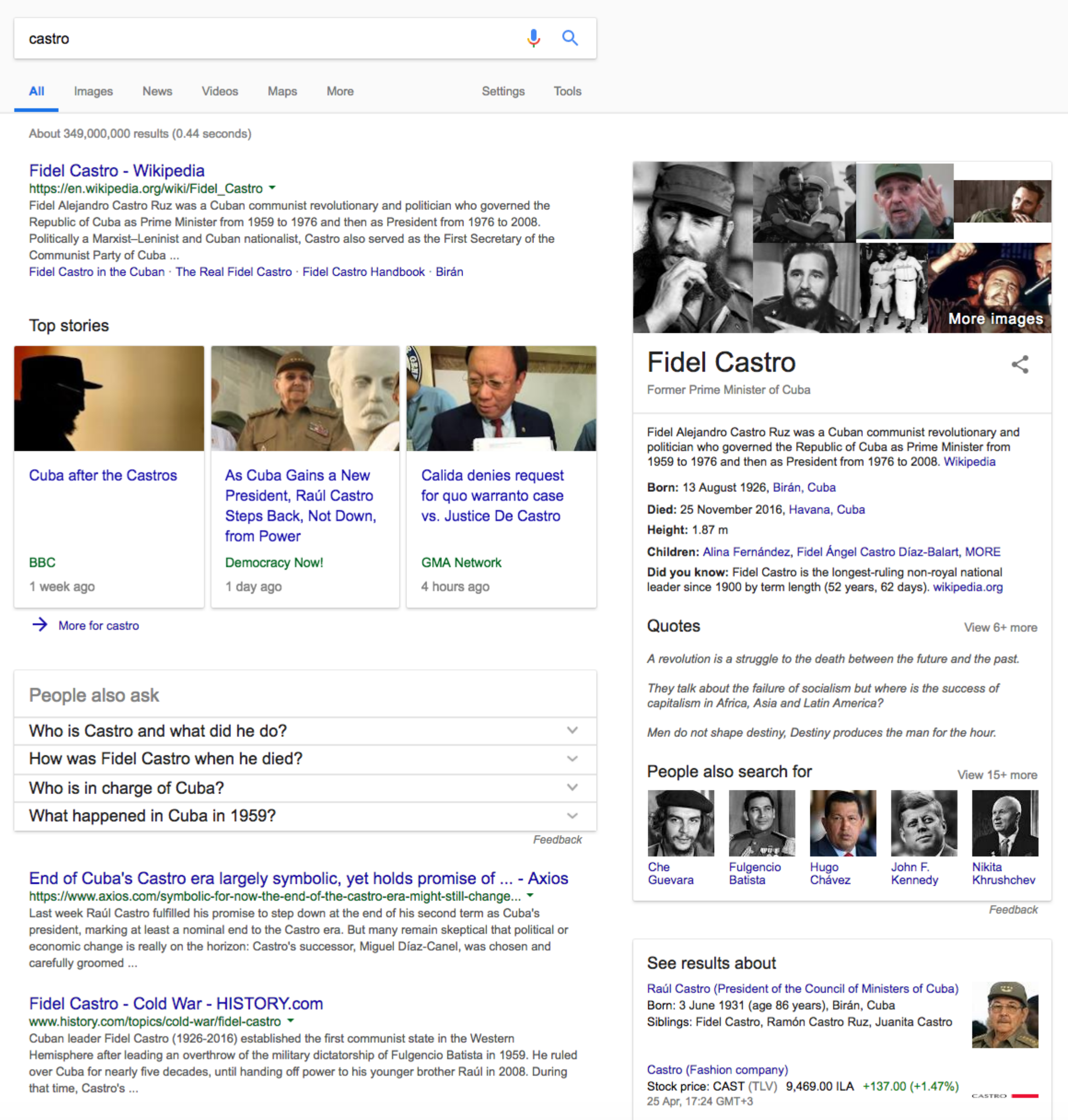
5.1 Introduction

Current search engines (e.g., Google or Bing) provide users with access to a wide range of specialised search services, in addition to web search. These specialised services, also known as *verticals*, allow users to direct their searches towards specific types of documents, such as *images*, *videos*, *news* and others. To help users explore relevant heterogeneous content, modern web search engines merge results from different verticals into a single results page (SERP) in a search paradigm that is known as *aggregated search* (Arguello, 2017).

Recently, besides aggregating results, modern search engines have started displaying complex information objects, or *entity cards*, on the results page. Similar to aggregated search, entity cards are intended to augment search results pages with diverse information, gathered from a variety of heterogeneous sources. Unlike aggregated search, they are assembled using different *semantic* retrieval techniques available to search engines[2] and are displayed as contextual elements (at the top-right of the SERP), rather than an in-line result or block of results, as is the case with aggregated search. Figure 5.1 shows the results page generated for the query *"castro"* by the Google search engine, which contains an entity card displayed in parallel to the list of web results.

Entity cards are intended to enhance search experience in several ways. Firstly, they help disambiguate underspecified information needs by highlighting different facets of a user's query. For instance, figure 5.1 shows the example of an entity card displayed for the query *"castro"* on Google's search interface. Although the central component of the card focuses on one particular facet of the query (in this case "Fidel Castro"), the *"See Results About"* component allows searchers to easily navigate search results about other facets of the query (e.g., "Castro" the fashion company). Secondly, entity cards summarise relevant content about a given topic by aggregating information from a variety of sources, such as images, Wikipedia or social media. Lastly, they support exploratory search by highlighting relationships between different entities associated with a given query.

[2] For example, Google uses its Knowledge Graph to create this type of information objects; similarly, the Satori Knowledge Base is used by Bing.



About 349,000,000 results (0.44 seconds)

Fidel Castro - Wikipedia
https://en.wikipedia.org/wiki/Fidel_Castro
Fidel Alejandro Castro Ruz was a Cuban communist revolutionary and politician who governed the Republic of Cuba as Prime Minister from 1959 to 1976 and then as President from 1976 to 2008. Politically a Marxist–Leninist and Cuban nationalist, Castro also served as the First Secretary of the Communist Party of Cuba ...
Fidel Castro in the Cuban · The Real Fidel Castro · Fidel Castro Handbook · Birán

Top stories

Cuba after the Castros
BBC
1 week ago

As Cuba Gains a New President, Raúl Castro Steps Back, Not Down, from Power
Democracy Now!
1 day ago

Calida denies request for quo warranto case vs. Justice De Castro
GMA Network
4 hours ago

→ More for castro

People also ask
- Who is Castro and what did he do?
- How was Fidel Castro when he died?
- Who is in charge of Cuba?
- What happened in Cuba in 1959?

End of Cuba's Castro era largely symbolic, yet holds promise of ... - Axios
https://www.axios.com/symbolic-for-now-the-end-of-the-castro-era-might-still-change...
Last week Raúl Castro fulfilled his promise to step down at the end of his second term as Cuba's president, marking at least a nominal end to the Castro era. But many remain skeptical that political or economic change is really on the horizon: Castro's successor, Miguel Díaz-Canel, was chosen and carefully groomed ...

Fidel Castro - Cold War - HISTORY.com
www.history.com/topics/cold-war/fidel-castro
Cuban leader Fidel Castro (1926-2016) established the first communist state in the Western Hemisphere after leading an overthrow of the military dictatorship of Fulgencio Batista in 1959. He ruled over Cuba for nearly five decades, until handing off power to his younger brother Raúl in 2008. During that time, Castro's ...

Fidel Castro
Former Prime Minister of Cuba

Fidel Alejandro Castro Ruz was a Cuban communist revolutionary and politician who governed the Republic of Cuba as Prime Minister from 1959 to 1976 and then as President from 1976 to 2008. Wikipedia

Born: 13 August 1926, Birán, Cuba

Died: 25 November 2016, Havana, Cuba

Height: 1.87 m

Children: Alina Fernández, Fidel Ángel Castro Díaz-Balart, MORE

Did you know: Fidel Castro is the longest-ruling non-royal national leader since 1900 by term length (52 years, 62 days). wikipedia.org

Quotes — View 6+ more

A revolution is a struggle to the death between the future and the past.

They talk about the failure of socialism but where is the success of capitalism in Africa, Asia and Latin America?

Men do not shape destiny, Destiny produces the man for the hour.

People also search for — View 15+ more

Che Guevara · Fulgencio Batista · Hugo Chávez · John F. Kennedy · Nikita Khrushchev

See results about

Raúl Castro (President of the Council of Ministers of Cuba)
Born: 3 June 1931 (age 86 years), Birán, Cuba
Siblings: Fidel Castro, Ramón Castro Ruz, Juanita Castro

Castro (Fashion company)
Stock price: CAST (TLV) 9,469.00 ILA +137.00 (+1.47%)
25 Apr, 17:24 GMT+3

Figure 5.1: Results page assembled by Google in response to the query *"castro"*. Entity card is displayed as a contextual element of the page, to the right of the web results ranking, offering an easily accessible summary of relevant and diverse content directly on the results page. Captured in April 2018.

Several factors suggest that studying the effect of entity cards on search behaviour is an important practical problem. Firstly, ambiguous queries are frequent. Sanderson and Croft (2012) examined a commercial search engine query log and found that 16% of all head queries issued by searchers are ambiguous. Entity cards are increasingly being used by modern search engines to assist users in disambiguating their information need. Given the frequent ambiguous queries modern search engines receive and because entity cards are becoming a tool frequently used to help users navigate diversified search results, understanding the impact of entity cards on search behaviour is crucial to understanding modern web search. Secondly, users' perception of entity cards can potentially influence their interaction with the search system as a whole. In fields other than information retrieval, research has shown that people associate attributes of a contextual stimulus to an object being judged. This effect has been observed in people's judgements on the quality of products (Morales and Fitzsimons, 2007) or businesses (Meyers-Levy and Sternthal, 1993), and is known as the *assimilation effect*. In information retrieval, this effect has been studied in connection to aggregated search, where the vertical search result blocks merged into SERPs were treated as contextual stimuli and web results as the objects being judged (Arguello and Capra, 2014, 2016; Liu et al., 2015). Given the contextual placement of entity cards on the SERP, we intend to determine whether entity cards trigger an assimilation effect and influence searchers' perception of web results, as reflected in their search interactions, something that has not been studied in prior work. Therefore, we focus on studying the effect entity cards have on user search behaviour and perceived workload by conducting a large-scale, crowd-sourced user study. We aim to answer the following research questions:

(RQ1) How does card presence and content influence users' search behaviour and perceived workload?

(RQ2) Do knowledge card properties (*coherence* and *vertical diversity*) moderate the effect cards have on workload and search behaviour?

The contributions of our work are twofold: ***(i)*** we examine the effect of ECs on user search interactions by analysing various different search behaviour signals, and perceived task workload, in a study with more than 500 unique searchers; ***(ii)*** we conduct a detailed analysis on the influence of entity card properties (*coherence* and *vertical diversity*) on search behaviour.

5.2 Prior Work

Understanding user search behaviour is a key component of modelling and evaluating search engine performance. Numerous recent studies have investigated different aspects of user search behaviour in information seeking tasks (Arguello and Capra, 2016; Diaz et al., 2013; Lagun et al., 2014; Navalpakkam et al., 2013). Extending these, our study brings together two different lines of research on user behaviour in information seeking: ***(i)*** the effect of *aggregated search* on user behaviour, with focus on aggregated search *coherence*; and ***(ii)*** searcher behaviour on *non-linear results pages* in the context of heterogeneous information access.

Aggregated search. In the context of aggregating heterogeneous information on the results page, user behaviour has been shown to differ significantly compared to the more traditional ten blue links environment. We discuss aggregated search results in detail in section 3.2.3. Overall, the presence of heterogeneous content displayed on the results page (i.e., through aggregated search) has been shown to influence user behaviour in various ways. Unlike aggregated search, entity cards merge results from different sources into a single contextual block that is usually shown at the top right of the results page, in parallel to the list of organic web results. There is limited understanding of search behaviour in this context and we review previous studies on non-linear results pages later in this section.

Our approach is similar to prior work (Arguello et al., 2012) in which the effects of task complexity and vertical display are investigated. In their study, Arguello et al. (2012) use task duration, number of queries issued, number of clicks on web and vertical results, and user explicit preferences to analyse the various effects of task complexity and vertical display on search behaviour. Similarly, our work studies the effect of entity cards on user search behaviour, but also looks at different entity card coherence and diversity manipulation in a novel search scenario: SERPs displaying entity cards shown as contextual, non-linear elements. We also report users' subjective assessments on different dimensions of task workload, operationalised through the application of the NASA TLX questionnaire (Hart and Staveland, 1988). Workload measurements using a similar technique have been used widely to understand search tasks in interactive information retrieval (Brennan et al., 2014).

Non-linear results pages. Web search interfaces are becoming increasingly complex, on both desktop and mobile, displaying heterogeneous results, entity cards, query suggestions and rich format ads in non-linear layouts. Understanding user behaviour on these novel interfaces is becoming essential for modelling and evaluating search engine performance. We discuss prior research on entity cards and non-linear result page elements in more detail in sections 3.2.4 and 3.2.5.

In addition to previous work, our study goes into more detail with regard to card content and structure, given that we manipulate not only card relevance, but also card coherence and vertical diversity. Even more, our study examines additional search behaviour signals, such as mouse hovers, scroll depth or query reformulations.

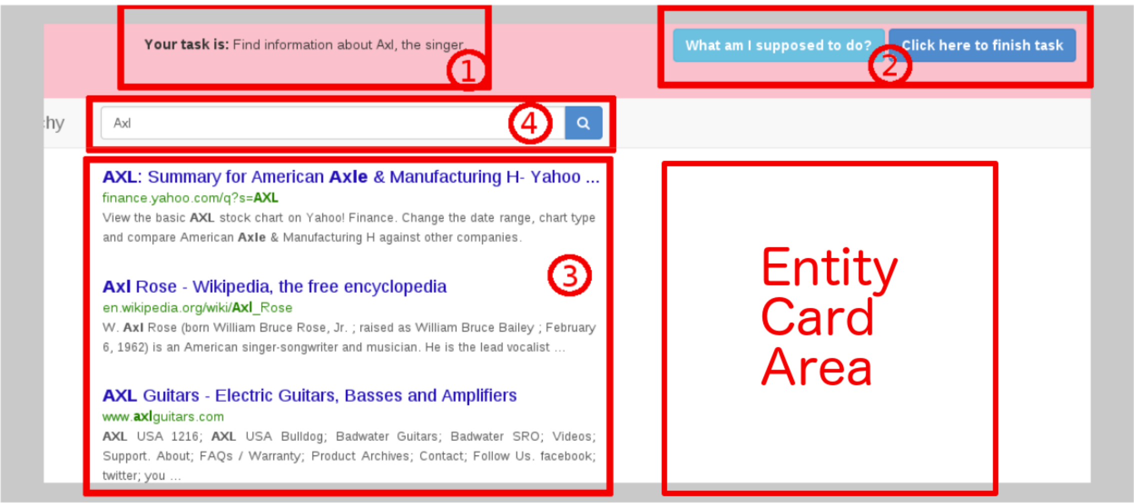
5.3 Experimental Methodology

We conducted a crowd-sourced experiment to investigate the effects of entity cards on search behaviour. In the following sections, we provide an overview of the experiment (section 5.3.1), a discussion on the experimental variables, including search tasks (section 5.3.2), and give details on the crowd-sourcing methodology we employed (section 5.3.4).

5.3.1 Overview

A practical scenario that relates to our study is the following: a user issues an ambiguous query to a search engine, either because: ***(i)*** their information need is well defined (e.g., "Find info about Fidel Castro, the Cuban president"), but their query is underspecified (e.g., "castro"); or because ***(ii)*** their information need is ambiguous (e.g., "What does Castro mean?") and their query reflects this ambiguity (e.g., "castro"). In response to the ambiguous query, the search interface displays an entity card that focuses on a particular entity (e.g., "Fidel Castro"). In this context, the user must decide to interact with the entity card or web results, or to reformulate the query. Our research questions focus on whether cards, and their properties (*content* or *topical focus*, *coherence* and *vertical diversity*), influence users' behaviour and workload during their search tasks.

Our study participants were given access to a live web search engine and asked to find results that were relevant to a search task defined by us. Our primary goal was to study user behaviour when the SERP displays an entity card. We followed a similar experimental protocol as previous work on aggregated search coherence (Arguello and Capra, 2014). In addition, we also manipulated card coherence and card diversity, as described in section 5.3.2. After being given general instructions for their tasks, participants were redirected to our search engine, where they were shown an *initial SERP*, which contained a list

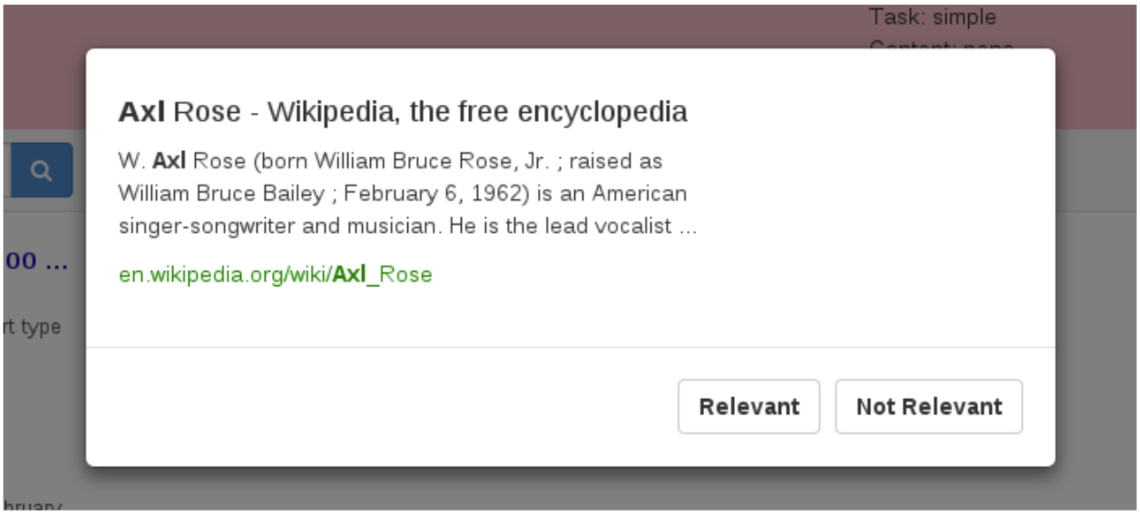

Your task is: Find information about Axl, the singer
What am I supposed to do?
Click here to finish task
Axl
1
2
3
4
Entity Card Area

AXL: Summary for American **Axle** & Manufacturing H- Yahoo ...
finance.yahoo.com/q?s=**AXL**
View the basic **AXL** stock chart on Yahoo! Finance. Change the date range, chart type and compare American **Axle** & Manufacturing H against other companies.

Axl Rose - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Axl**_Rose
W. **Axl** Rose (born William Bruce Rose, Jr. ; raised as William Bruce Bailey ; February 6, 1962) is an American singer-songwriter and musician. He is the lead vocalist ...

AXL Guitars - Electric Guitars, Basses and Amplifiers
www.**axl**guitars.com
AXL USA 1216; **AXL** USA Bulldog; Badwater Guitars; Badwater SRO; Videos; Support. About; FAQs / Warranty; Product Archives; Contact; Follow Us. facebook; twitter; you ...

(a) Search interface used in our study – highlighted elements: ***(1)*** Task description; ***(2)*** Task control buttons, allowing participants to access additional information about their task, by clicking the *"What am I supposed to do?"* button, or end their task by clicking the *"Click here to finish task"* button; ***(3)*** Search results ranking pre-populated with results related to the participant's assigned query, on the *initial SERP* where all experimental manipulation occurred; ***(4)*** Search bar pre-populated with participant's assigned query, on the *initial SERP* where all experimental manipulation occurred.



Axl Rose - Wikipedia, the free encyclopedia

W. **Axl** Rose (born William Bruce Rose, Jr. ; raised as William Bruce Bailey ; February 6, 1962) is an American singer-songwriter and musician. He is the lead vocalist ...

en.wikipedia.org/wiki/**Axl**_Rose

(b) In our study interface, clicking on a web result displayed a dialogue which contained some information about the result (i.e., title, URL and snippet) as well as two buttons for labelling the result as either *"Relevant"* or *"Not Relevant"*. Participants were given access only to web results, and not to actual web documents.

Figure 5.2

of results and an entity card. The initial SERP was the only part of the experiment in which participants interacted with the entity card, and it was where all experimental manipulation took place. In addition, the initial SERP displayed a standard querying interface, a topic accompanied by a detailed description and an *initial query* for which the results on the page had been retrieved. On the initial SERP, we displayed the top-50 results returned by the Bing web Search API, without pagination support or vertical results (i.e., aggregated search blocks). We decided to show additional results on the page, and not just the traditional ten blue links, because we wanted to assess the effect of entity cards on the effort searchers are willing to expend during their tasks, as reflected by scroll depth. Figure 5.2a displays the search interface used in our experiments.

Participants were asked to *"find information"* about their assigned topic, by marking web results they considered informative as relevant. Clicking on a web result displayed a dialogue which contained some information about the result — title, URL and highlighted snippet — and two buttons for labelling the result as either *"Relevant"* or *"Not Relevant"*, as shown in Figure 5.2b. Clicking on different blocks of the entity card displayed a similar dialogue, containing a highlighted version of the card block (but participants were not asked to inspect the entity card in any way).

Participants were told to search freely, by issuing their own queries or by inspecting the results on the initial SERP. A button in the top-right corner of the search interface allowed participants to end their task when they had "found enough information" about their given topic. Clicking this button displayed a post-task questionnaire, where they were asked to fill in their subjective workload assessments. The task ended after participants submitted their responses.

Before discussing experimental manipulations, it is worth reviewing card structure and components. The cards that we used in our experiment are made up of four major blocks: *Images*, *Wiki*, *Related Entities* and *See Results About*. Each individual block contains several items from its respective vertical. The *Wiki* block displays, in addition to a brief summary about a given entity, a list of short facts specific to that particular entity – for example, the entity card for "Fidel Castro" displays a list of facts about his *date of birth*, *height*, *siblings*, and others. In our experiment, all manipulations of the *Wiki* block refer to manipulations on this list of facts. Figure 5.3 shows a structure diagram for a typical entity card used in our experiment. More details on card content and structure are provided in the following sections.

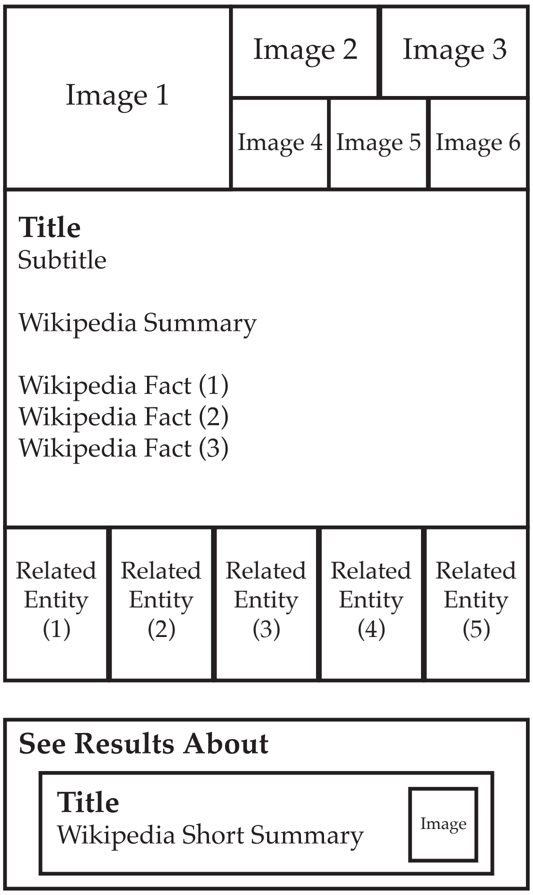


Figure 5.3: Structure of entity cards used in our experiment. There is some variation in the way different cards are presented on modern web search interfaces, depending on the sources of information that are relevant to the searcher's query. Some cards contain maps or social media results. For our experiment, we decided to investigate only cards that follow the structure outlined here, and, where necessary, replaced maps with images, as well as removed social media content.

5.3.2 Input Variables

In this section, we describe our experimental manipulations. The variables we control relate to card *content* — relevance or topical focus — and card *properties and structure*. The user interaction outcome measures that we employed as a proxy for search behaviour and our approach to assessing workload are described in this section as well.

5.3.2.1 Tasks

For our study, we chose to select 40 different ambiguous search topics and attach task descriptions to each individual topic. All topics used in our study are listed in Table 5.1. The topics used in our experiment were selected by following a procedure similar to Sanderson and Croft (2012), and employed in several studies on aggregated search (Arguello and Capra, 2012, 2014, 2016; Arguello et al., 2013; Capra et al., 2013). Firstly, we selected a set of ambiguous entities by identifying all Wikipedia disambiguation pages. On Wikipedia, a disambiguation page is a page that displays links to specialised articles related to different meanings of an ambiguous entity, and serves as a navigational hub[3]. A total of 162, 987 Wikipedia disambiguation pages were identified. Secondly, we selected only the ambiguous entities that appeared in the AOL query-log: 36, 910 ambiguous entities (roughly 22% of the initial set) had an exact match in the AOL query-log. We then sorted the set based on entity popularity[4] and selected the top 10% most popular ambiguous entities on Wikipedia that appear in the AOL query-log as our selection pool. Even though entity cards are displayed for non-ambiguous queries as well, we chose to focus on ambiguous queries because this allows us to realistically explore the effects of both *on-topic* and *off-topic* entity cards on search interactions.

Because we wanted to select topics that potentially trigger the presence of entity cards, we issued each entity in our set of 3, 691 entities to a popular commercial search engine and, using a scraper, downloaded all the corresponding SERPs. Using regular expressions, we identified a total of 2, 485 entities (roughly 67% of our filtered set) that triggered entity cards on the results page. The high percentage of ambiguous queries that trigger cards further supports our assertion that entity cards are increasingly being used by commercial search engines to enhance user experience and that studying their effect on search behaviour is an important practical problem.

Finally, we randomly selected 40 entities that had at least two different senses

[3] For example http://en.wikipedia.org/wiki/Castro.

[4] The number of page visits for its top-3 most visited disambiguated pages, from https://dumps.wikimedia.org.

Query terms	Task description	On topic entity	Off topic entity
CK	Find information about CK, the fashion company.	Calvin Klein (Fashion company)	Louis CK (Comedian)
Voodoo	Find information about Voodoo, the Caribbean r...	Haitian Vodou (Religion)	Voodoo (Song)
American Tragedy	Find information about American Tragedy, the n...	An American Tragedy (Novel by Theodore Dreiser)	American Tragedy (Studio album by Hollywood Un...
Andy	Find information about Andy, the tennis player.	Andy Murray (Tennis player)	Andy Warhol (Visual Artist)
Ararat	Find information about Ararat, the mountain in...	Mount Ararat (Peak in Turkey)	Ararat (2002 film)
Armstrong	Find information about Armstrong, the manufact...	Armstrong World Industries (Manufacturing comp...	Lance Armstrong (Professional Road Racing Cycl...
Avenger	Find information about Avengers, the action mo...	The Avengers (Fantasy / Action Film)	Dodge Avenger (Mid-size sedan)
Axl	Find information about Axl, the singer.	Axl Rose (Singer-songwriter)	AXL Guitars (Company)
Bergman	Find information about Bergman, the director.	Ingmar Bergman (Director)	Ingrid Bergman (Film actress)
Black Moon	Find information about Black Moon, the movie.	Black Moon (1975 film)	Black Moon (Musical Group)
Bloody Mary	Find information about Bloody Mary, the cocktail.	Bloody Mary (Cocktail)	Mary I of England (Queen of England)
Blue Hills	Find information about Blue Hills, the ski area.	Blue Hills Ski Area (Canton, MA)	Blue Hills Regional Technical School (High sch...
Brotherhood	Find information about Brotherhood, the Americ...	Brotherhood (American television series)	The Brotherhood of War (2004 film)
Carter	Find information about Carter, the apparel com...	Carter's, Inc. (Apparel company)	Jimmy Carter (39th U.S. President)
Castro	Find information about Castro, the former pres...	Fidel Castro (Former President of Cuba)	The Castro (Neighbourhood in San Francisco, Cal...
Challenger	Find information about Challenger, the sports ...	2015 Dodge Challenger (Sports car)	Space Shuttle Challenger (Spacecraft)
City of Angels	Find information about City of Angels, the 199...	City of Angels (1998 film)	City of Angels (Musical by Larry Gelbart)
City of God	Find information about City of God, the 2002 f...	City of God (2002 film)	City of God (Book by Augustine of Hippo)
Clue	Find information about Clue, the 1985 film.	Clue (1985 film)	Clue (Video game)
Colt	Find information about colt, the animal.	Colt (Animal)	Colt's Manufacturing Company (Corporation)
Congo	Find information about the Republic of Congo, ...	Congo (Country in Africa)	Congo (1995 film)
Cyrus	Find information about Cyrus, the king.	Cyrus the Great (King)	Miley Cyrus (Singer-songwriter)
Dark Angel	Find information about Dark Angel, the America...	Dark Angel (American television series)	Dark Angel (Band)
Dead man	Find information about Dead Man, the 1995 film.	Dead Man (1995 film)	Deadman (Fictional Character)
Dead or Alive	Find information about Dead or Alive, the vide...	Dead or Alive (Video game series)	Dead or Alive (Band)
Dido	Find information about Dido, the singer.	Dido (Singer-songwriter)	Dido (Queen of Carthage)
Doom	Find information about Doom, the video game.	Doom (Video game)	Doom (Band)
Fame	Find information about Fame, the 1980 film.	Fame (1980 film)	Fame (Duo)
Fargo	Find information about Fargo, the 1996 film.	Fargo (1996 film)	Fargo (City in North Dakota)
Gallagher	Find information about Gallagher, the comedian.	Gallagher (Comedian)	Brendan Gallagher (Ice hockey player)
Gazza	Find information about Gazza, the soccer player.	Paul Gascoigne (Soccer player)	Gazza (Musical Artist)
Gettysburg	Find information about Gettysburg, the 1993 film.	Gettysburg (1993 film)	Gettysburg (Town in Pennsylvania)
Gojira	Find information about Gojira, the 1954 film.	Gojira (1954 film)	Gojira (Band)
Homefront	Find information about Homefront, the 2013 film.	Homefront (2013 film)	Homefront (Video game)
Jaya	Find information about Jaya, the singer.	Jaya (Singer)	Jaya Bhaduri Bachchan (Indian Politician)
JJ	Find information about JJ, the director.	J.J. Abrams (Director)	JJ (Swedish band)
JK	Find information about JK, the novelist.	J. K. Rowling (Novelist)	J.K. (Singer)
Irving	Find information about Irving, the basketball ...	Kyrie Irving (Basketball player)	Washington Irving (Author)
Locke	Find information about Locke, the 2013 film.	Locke (2013 film)	John Locke (Philosopher)
King Stephen	Find information about King Stephen, the king ...	Stephen, King of England (King)	Stephen King (Author)

Table 5.1: Queries and task descriptions used in our study. *On* and *Off* senses, as reflected in our manipulations of entity cards, also shown in the two right most columns.

determined by the presence of a *See results about* block on the entity card, which allows searchers to navigate to a different sense of a given query. For each search task, we manually assembled entity cards, containing similar information[5], displayed in the same style, as the cards shown on the commercial SERPs we scraped, whereas the *off topic* cards contain information about the related topic, shown in the *See results about* of the scraped card. All card content, including images, was cached on our servers.

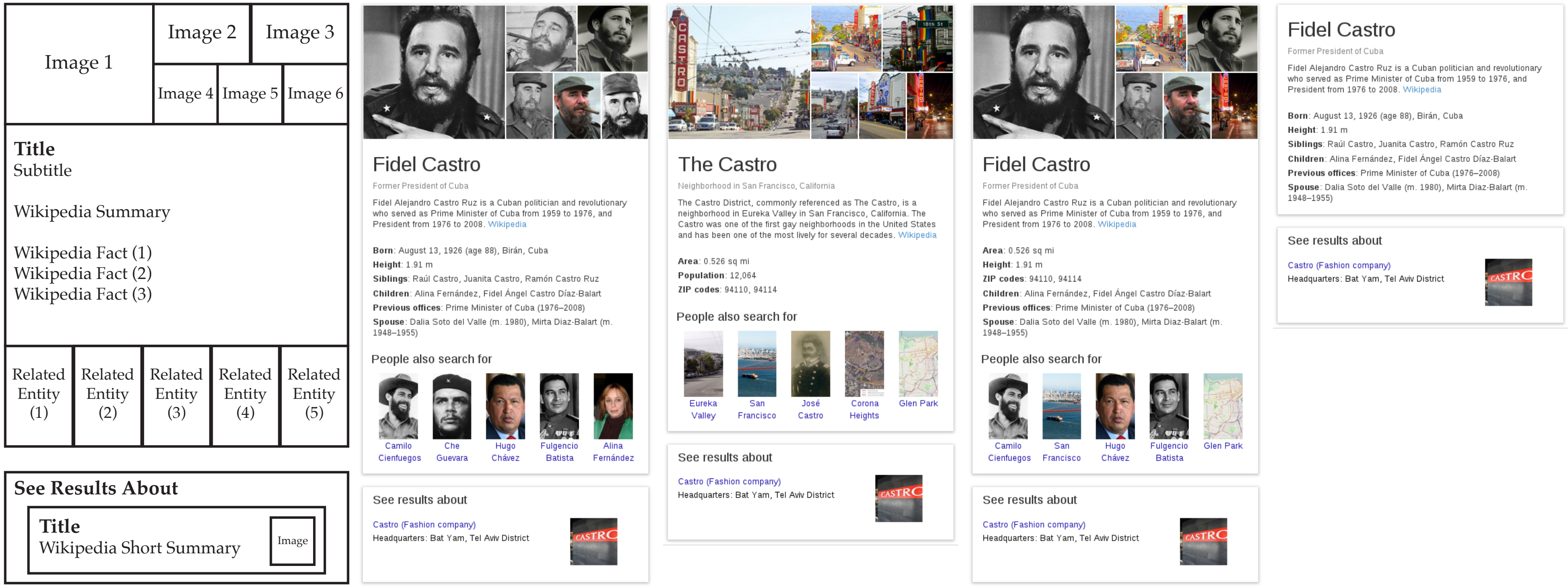
5.3.2.2 Cards and Card Properties

In this section, we describe our experimental manipulations of cards and card properties. We explored three different dimensions of card manipulation: ***(i)*** card content — whether the card is *on-topic* or *off-topic* with regard to the user's assigned search topic; ***(ii)*** card coherence — whether card blocks are coherent and all focus on the same topic of a user's assigned topic; ***(iii)*** card diversity — whether cards contain visually salient blocks of elements, such as *Images*.

The ***card content*** variable manipulated information displayed on the entity card and the card's presence on the SERP. For our tasks, entity cards were either ***(i)*** on-topic, displaying information about the user's assigned search topic, ***(ii)*** off-topic, displaying information about a different facet of the user's assigned topic, or ***(iii)*** completely absent from the SERP. Figure 5.4 shows examples of *on* and *off-topic* entity cards for the query "castro". Our decision to manipulate card content is primarily motivated by the frequency with which entity cards are being displayed for ambiguous user queries, as previously discussed. Inferring intent from ambiguous queries has been widely studied in recent work (Ashkan et al., 2009; Brenes et al., 2009), and remains one of the important problems of modern web search. Given that query disambiguation is problematic, and that entity cards are becoming widely used as disambiguation helpers, understanding user interaction with *on* and *off-topic* cards is an important aspect of modelling and understanding search behaviour.

In addition to card content, we also manipulated ***card coherence*** to study the assimilation effects of unreliable cards on user behaviour and workload. In aggregated search, if web results and vertical results on a SERP focus on different senses of a given query, then the overall results can be described as having *low* coherence (Arguello and Capra, 2012). Unlike aggregated search, our manipulation of card coherence refers to the internal components of the card, and not the card's

[5] There is some variation in the way different cards are presented. Some cards contain results from the maps or social media verticals. For our experiment, we decided to investigate only cards that follow the structure outlined in figure 5.4, and, where necessary, replaced maps with images, as well as removed social media content.



Fidel Castro

Former President of Cuba

Fidel Alejandro Castro Ruz is a Cuban politician and revolutionary who served as Prime Minister of Cuba from 1959 to 1976, and President from 1976 to 2008. Wikipedia

Born: August 13, 1926 (age 88), Birán, Cuba
Height: 1.91 m
Siblings: Raúl Castro, Juanita Castro, Ramón Castro Ruz
Children: Alina Fernández, Fidel Ángel Castro Díaz-Balart
Previous offices: Prime Minister of Cuba (1976–2008)
Spouse: Dalia Soto del Valle (m. 1980), Mirta Diaz-Balart (m. 1948–1955)

People also search for

See results about

Castro (Fashion company)
Headquarters: Bat Yam, Tel Aviv District

The Castro

Neighborhood in San Francisco, California

The Castro District, commonly referenced as The Castro, is a neighborhood in Eureka Valley in San Francisco, California. The Castro was one of the first gay neighborhoods in the United States and has been one of the most lively for several decades. Wikipedia

Area: 0.526 sq mi
Population: 12,064
ZIP codes: 94110, 94114

People also search for

See results about

Castro (Fashion company)
Headquarters: Bat Yam, Tel Aviv District

Fidel Castro

Former President of Cuba

Fidel Alejandro Castro Ruz is a Cuban politician and revolutionary who served as Prime Minister of Cuba from 1959 to 1976, and President from 1976 to 2008. Wikipedia

Area: 0.526 sq mi
Height: 1.91 m
ZIP codes: 94110, 94114
Children: Alina Fernández, Fidel Ángel Castro Díaz-Balart
Previous offices: Prime Minister of Cuba (1976–2008)
Spouse: Dalia Soto del Valle (m. 1980), Mirta Diaz-Balart (m. 1948–1955)

People also search for

See results about

Castro (Fashion company)
Headquarters: Bat Yam, Tel Aviv District

Fidel Castro

Former President of Cuba

Fidel Alejandro Castro Ruz is a Cuban politician and revolutionary who served as Prime Minister of Cuba from 1959 to 1976, and President from 1976 to 2008. Wikipedia

Born: August 13, 1926 (age 88), Birán, Cuba
Height: 1.91 m
Siblings: Raúl Castro, Juanita Castro, Ramón Castro Ruz
Children: Alina Fernández, Fidel Ángel Castro Díaz-Balart
Previous offices: Prime Minister of Cuba (1976–2008)
Spouse: Dalia Soto del Valle (m. 1980), Mirta Diaz-Balart (m. 1948–1955)

See results about

Castro (Fashion company)
Headquarters: Bat Yam, Tel Aviv District

(a) Card structure (b) On-topic card (c) Off-topic card (d) Non-coherent card (e) Non-diverse card

Figure 5.4: General entity card structure and different experimental manipulations of entity cards used in our study for the query *"Castro"*: ***(a)*** General card structure; ***(b)*** On-topic card, focusing on the primary topic of the query (in this case, *"Fidel Castro, Former President of Cuba"*); ***(c)*** Off-topic card, focusing on the secondary topic of the query (in this case, *"Castro, Neighbourhood in San Francisco"*); (d) Non-coherent card, focusing on the primary topic of the query, but displaying images, Wikipedia facts and related entities on the secondary meaning of the query; ***(e)*** Non-diverse card, focusing on the primary topic of the query, without displaying images or related entities.

relation to web results. In our work, a card is *non-coherent* when it contains both *on-topic* and *off-topic* content within the same card. This exploration of card coherence is partially motivated by previous work on aggregated search (Arguello and Capra, 2012), and also by our empirical observations that cards scraped from commercial search engines (as described in the previous section) can contain non-coherent elements. To create *non-coherent* cards, we replaced two individual elements within each card block – *Images*, *Wiki* and *Related entities* – on the on-topic card, with corresponding elements from the off-topic card. Figure 5.4d shows an example of a *non-coherent* card for the search topic *"Castro"*: each individual block within the card contains both *on-topic* and *off-topic* elements (e.g., in Figure 5.4d, images two and six are replaced with their off-topic counterparts, displaying views of *Castro, the neighbourhood in San Francisco* rather than *Castro, the former President of Cuba*). In our investigation of card coherence, we only manipulated the top three blocks of each card, without modifying the *See Results About* block. Coherent cards are *on-topic* cards with unmodified content – all elements within their blocks focus on the same aspect of a user's assigned search topic.

In addition to card coherence, we also manipulated card ***vertical diversity*** to determine the effect of cards' content diversity and visual saliency on user behaviour. In our study, *diverse* cards displayed all the blocks, whereas *non-diverse* cards displayed the *Wiki* block, but not the *Images* or *Related Entities* blocks, as shown in figure 5.4e. Note that our definition of card *vertical diversity* implies that non-diverse cards contain less information than diverse cards. Both types of cards displayed the *See Results About* block.

5.3.3 Output Variables

In this section, we describe the interaction signals we used to investigate user search behaviour and the way we operationalised workload assessments.

Search behaviour was investigated only on the initial SERP, given that our main goal was to investigate the effect of entity cards on user interaction with web results and entity cards were only shown on the initial SERP. The experimental output measures we used can be partitioned into related groups of variables:

- ***Card interactions***: relating to searchers' engagement with entity cards, as reflected in the number of mouse clicks on the card (`num_card_click`) or in the number of mouse hovers over card elements (`num_card_hovers`), where a hover is determined by the presence of a mouse cursor above the card for more than 2 seconds. Note that in our instructions, we did not ask participants to interact with the card in any way, only to mark informative web results as relevant.

- ***Web interactions***: relating to searchers' engagement with the overall search page, as reflected in the number of mouse clicks on web results (`num_web_clicks`), the number of mouse hovers over web results (`num_web_hovers`), where a hover is determined by the presence of a mouse cursor above a web result surrogate (i.e., title or snippet) for more than 2 seconds, the number of results marked as relevant (`num_rel_web_docs`), the number of reformulated queries (`num_reform_queries`), scroll depth in the results list (`web_scroll_depth`). In addition, we also measured the time spent interacting with the initial SERP (`time_init_serp`), which is where all experimental manipulations of entity cards occurred, as well as the duration of time taken before issuing a first click on the results list (`time_to_web_click`). All time measures are reported in seconds.

These experimental outcome measures extend prior work on aggregated search that considered only clicks and bookmarks (Arguello and Capra, 2012; Arguello et al., 2013), and indicate different levels of engagement with search systems.

Workload was operationalised by applying the *raw NASA Task Load Index (TLX)* (Hart and Staveland, 1988) questionnaire. The TLX is a tool used to assess perceived workload, and measures various types of demands imposed on participants during their task, as well as self-assessed effort, frustration and performance. It is one of the most widely used workload measurement scales (Megaw, 2005) and has been applied in various studies related to information seeking (Brennan et al., 2014; Speier and Morris, 2003; Stasi et al., 2011). Our study participants completed the workload questionnaire at the end of each task, recording self-assessed perceived workload in various dimensions (`mental`, `physical`, `temporal`, `performance`, `effort` and `frustration`) by answering 6 questions on a scale from 1 to 20 (low to high). Table 5.2 lists workload questions, as shown to participants in our study. The overall workload was computed as the mean of individual responses.

5.3.4 Study Details

Our crowd-sourcing study was run as an external task, using both Amazon's Mechanical Turk (MTurk) and CrowdFlower (CF) platforms. We decided to publish the study on both platforms because we wanted a diverse pool of workers attempting our tasks. On both platforms, we employed similar quality control mechanisms, as outlined below.

Workload dimension	*Workload question*
Mental	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal	How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Table 5.2: Questions associated with individual dimensions of perceived task workload, as shown to study participants at the end of their task.

Card content	*Card coherence*	*Card diversity*	*HITs analysed*
Absent	–	–	284/400
Off-topic	Coherent	Diverse	269/400
On-topic	Coherent	Non-diverse	279/400
On-topic	Coherent	Diverse	289/400
On-topic	Non-coherent	Diverse	267/400

Table 5.3: All experimental conditions explored in our study, and associated number of HITs analysed, after removing submissions from workers flagged as unreliable, out of a total of 400 HITs collected in each condition.

Each Human Intelligence Task (HIT) corresponded to a single search task. Because we wanted to increase the number of individual participants attempting our tasks, workers were allowed to complete no more than 5 of our HITs. The study design was not fully crossed, because we wanted to investigate the effect of card coherence only on cards that can potentially improve search experience (i.e. *on topic* cards). Similarly, we only studied the effect of card diversity on coherent, on topic cards. Table 5.3 shows a list of all experimental conditions and variable manipulations in our study. In total, we had 200 experimental conditions: 5 different conditions for card presence and properties × 40 search topics. For each experimental condition, we collected 10 redundant data points, for a total of 2000 HITs. Each HIT was priced at $0.15 USD. Our tasks were completely *external* to the crowd-sourcing platforms they were deployed on, meaning that only worker recruitment and compensation was managed by the platforms. Our system assigned workers to experimental conditions, logged user interaction, and dynamically managed quality control. This allowed us to capture all the user interaction measures described previously. Additionally, each worker was assigned a different search topic for each task attempted. In total, we report on data collected from 510 unique workers.

Quality control is one of the major components of running crowd-sourcing tasks. We approached quality control from several angles. Firstly, on both platforms we allowed only workers from English speaking countries[6] to attempt our tasks. Secondly, we only allowed workers with acceptance rate above 85% on AMT, or Level 2 on CF, to complete our task. Thirdly, because entity cards occupy a large area on the SERP, we needed to ensure that searchers can at least view them. Therefore, we disabled access to our HITs from mobile devices, or from browser windows with display width lower than 800 pixels. Finally, because our tasks were external to the crowd-sourcing platforms, we were able to dynamically manage quality control on two dimensions. Firstly, we inserted two results retrieved for a different query on each individual SERP, at random positions between ranks 3 and 10. Workers marking any of these results as relevant were labelled as unreliable and were not allowed to attempt any other HITs. Secondly, the workload questionnaire contained an additional question, ranked low to high, asking workers "How much attention are you paying right now?". For each task, the question was displayed at a random position in the list of questions. All workers with answers below 50% for this question were not allowed to attempt other HITs. All HITs attempted by unreliable workers (as determined by the methods outlined above) were removed from the final analysis – this pro-

[6] USA, UK, IRL, AU, NZ, CAN

cess did not affect the balance of our data: after removing unreliable HITs, each condition had approximately the same number of associated HITs. In total we analysed 1661 HITs, approximately 70% of the total data we collected.

5.3.5 Statistical Methodology

In an experimental setting, a statistical model is required to assess the differences between two groups (e.g., whether one group is, on average, larger than the other) because measurements are typically accompanied by error and noise that obstruct drawing conclusions from the observed data directly. Different methods of comparing groups of experimental data exist, but the *de facto* approach to statistical comparison of two or more groups is the statistical test (e.g., *Student's T-test*). Typically, this requires defining a clear *null hypothesis* and the use of a statistic computed from the data (e.g., *t statistic*) to determine whether differences between groups are higher than some arbitrary (pre-specified) threshold, in which case the differences are deemed significant. However, this type of approach to assessing differences in data is difficult to conduct correctly, as it relies on a number of subjective choices (e.g., test to use, null hypothesis, significance level) which are typically justified based on customary techniques (i.e., what everyone else uses) that are *"entirely arbitrary"* (Johnson, 1999) and that overstate the evidence against the null hypothesis (Goodman, 1999). For example, comparing many different groups and variables, as in our experiment, can lead to overestimated confidence (or underestimated confidence, in the case of corrected comparisons (Narum, 2006)) in differences between groups of data. Indeed, completely abandoning this approach to statistical testing is being discussed widely (Goșa et al., 2018; McShane et al., 2017).

A more effective approach has been proposed in the form of *estimating the difference* between groups, rather than testing this difference (Kruschke, 2013; Straw et al., 2015), driven by Bayesian probability theory. This approach aims to estimate how different groups of data are, which is more informative than simply stating that they are different (significantly or not), and also addresses the problem of multiple comparisons implicitly (Gelman and Tuerlinckx, 2000). We use this approach in our analysis, following the work conducted in Kruschke (2013). Similar approaches have been used in information retrieval studies previously (Carterette, 2015). As an example, in the remainder of this section we attempt to work through estimating differences in the number of mouse hovers on web results (`num_web_hovers`) when the card is in one of the following experimental conditions: *absent*, *on topic* or *off topic*.

The first step in our approach is to specify a full probability model for the variable under analysis. In this example, we are interested in modelling a count variable (i.e., the number of times a study participant hovered over web results) and therefore use a Poisson random variable and its associated probability distribution (Poisson and Schnuse, 1841) as the assumed underlying model of our data. The Poisson distribution is often used to model the number of events occurring in a fixed period of time, when the times at which events occur are independent. In our case, the fixed period of time is the experiment duration, and we assume hovers over web results to be independent.

A Poisson random variable has one parameter λ which corresponds to the average number of events per given time interval. Thus, we specify the likelihood function of our models as:

$$\begin{aligned} y^{(\text{absent})} &\sim Poisson(\lambda_{\text{absent}}) \\ y^{(\text{on topic})} &\sim Poisson(\lambda_{\text{on topic}}) \\ y^{(\text{off topic})} &\sim Poisson(\lambda_{\text{off topic}}) \end{aligned}$$

where $y^{(k)}$ is the number of web hovers, and $k \in \{\text{absent}, \text{on topic}, \text{off topic}\}$. In other words, we assume that the number of web hovers can be approximated by a Poisson distribution. The parameter λ is real-valued, therefore we apply a normal prior, setting the hyper-parameters to the pooled mean (μ) and three times the pooled standard deviation (σ) of the empirical data, such that:

$$\lambda_k \sim Normal(\mu, 3 \cdot \sigma)$$

which applies very diffuse information to the number of hovers occurring per unit of time, and does not favour a particular value a priori (Straw et al., 2015). Having specified our model, we can then estimate the posterior distribution of our parameters using a sampling approach[7] and the data we observed. In this example, λ_k is the expected number of web hovers under different experimental conditions. Comparing the posterior distributions of λ_{absent} and $\lambda_{\text{on topic}}$, for instance, allows us to estimate differences in the expected number of web hovers between the two conditions. In addition, this allows us to explicitly quantify uncertainty due to our lack of knowledge of model parameters, and uncertainty due to the inherent randomness of experimental settings.

[7] We use PyMC3 with the No-U-Turn Sampler to execute our tests, for each test generating 5500 samples, burning the initial 500 samples. We note that all our parameter estimates, in all tests, converge using this sampling approach.

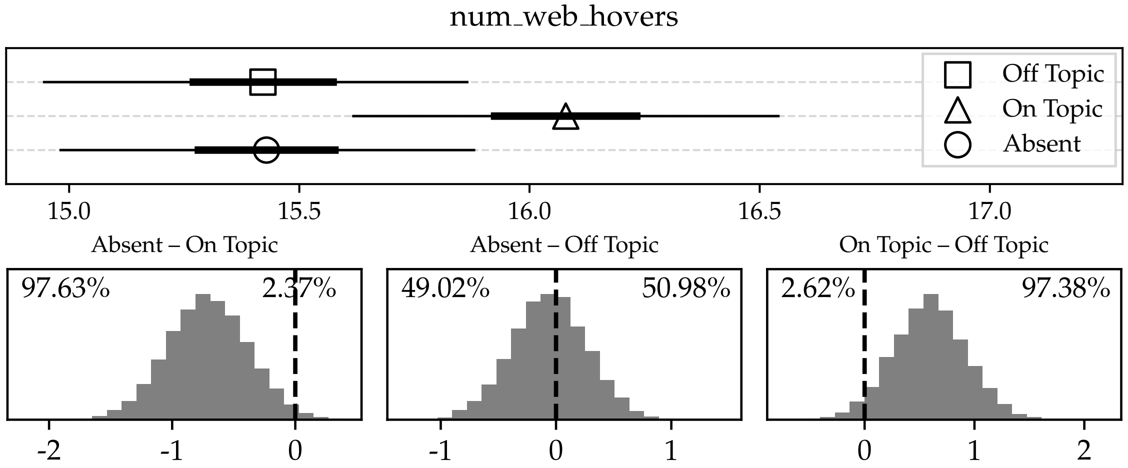


Figure 5.5: Example showing the posterior distribution of the estimated expected number of web result hovers in different experimental conditions.

The ***top panel*** of the figure shows the posterior distribution of the parameters λ_k — in this case, the expected number of web result hovers in different experimental conditions — with the interquartile range of the parameters' distributions shown as a dark bar, and the 95% Bayesian credible interval of the parameters' distributions shown as a light bar.

The ***lower panels*** show the posterior distribution of parameter differences across condition pairs. Specifically, the ***left-most lower panel*** shows the distribution of parameter differences (i.e., difference in expected number of web hovers) between the *Absent* and the *On Topic* conditions. Using 0 as a reference value (because if no differences between parameters across the two conditions are estimated, the distribution of differences should be centred on 0) we can observe that a majority of the difference distribution lies below 0, indicating that the differences we observe between these two conditions is meaningful. The panel title indicates the directionality of the result: in this example, it is the number of hovers in the *Absent* condition minus the number of hovers in the *On Topic* condition (rather than its converse). The panel labels (in this case, 97.63% and 2.37%) indicate what proportion of the difference distribution lies below or above zero, respectively (i.e., the cumulative probability for the posterior distribution on either side of the reference value). All lower panels present the same type of information, but with respect to different experimental condition pairs. Similarly, each panel title indicates the directionality of differences between conditions, and panel labels indicate what proportion of the difference distribution is below or above zero. For instance, the ***centre lower panel*** indicates no differences between *Absent* and *Off Topic* conditions, whereas the ***right-most lower panel*** indicates a positive difference between the number of web hovers in the *On Topic* and *Off Topic* conditions, which lies almost entirely above the reference value of 0, indicating that, on average, the *On Topic* condition is very likely to lead to more web hovers compared to the *Off Topic* condition.

Figure 5.5 shows the format in which we present our results in the following sections. The top panel shows the posterior distribution of the expected rate of web result hovers, across conditions, whereas the lower panels show the posterior distribution of differences between condition pairs. For example, the left panel shows the posterior difference in the expected number of web hovers between the *absent* and *on-topic* card conditions: 97.63% of the distribution is below zero, which means that, on average, it is very probable that displaying an on-topic entity card increases the number of web result hovers, as compared to the absent condition. Indeed, Gelman and Tuerlinckx (2000) suggest that claims based on 95% posterior intervals can be made *"with confidence"*, and we follow their recommendation (i.e., when more than 95% of the posterior difference of a given outcome measure between two experimental conditions is above or below 0, we conclude that experimental manipulations had an effect on that particular outcome measure that is unlikely due to chance). We also report *Cohen's d* (Cohen, 1977, 1992) associated with each difference of outcomes, as a measure of effect size. *Cohen's d* is a standardised measure of the difference between two groups, and is computed by:

$$d = \frac{\mu_i - \mu_j}{\sqrt{\frac{1}{2}(\sigma_i^2 + \sigma_j^2)}}$$

where μ and σ are the mean and standard deviations of the two groups. Given that, in our case, this statistic is computed using distributions of means and distributions of standard deviations, it also has a posterior distribution, but in our results, we report the mean of this distribution as the mean effect size.

The Poisson distribution can be used to model the number of events occurring in a fixed period of time, however our experimental outcome variables are not all counts of events. For example, the duration of time spent on the initial SERP (`time_on_init_serp`) can take any positive real value, whereas answers to workload question can only take one of twenty different values (i.e., participants are asked to answer on a scale from 1 to 20). As such, depending on the type of outcome measure, we use different distributions to model differences between experimental conditions: for count measures (i.e. outcome variables that start with `num_`), we use the Poisson distribution, as discussed above; for time measures (i.e., outcome variables that start with `time_`), we use the exponential distribution, which is commonly used in modelling the time between events; for the workload answers, we use the beta-binomial distribution which is commonly used to model variables that occur within a finite range of non-negative integers;

and for the overall workload, we use the normal distribution, which is commonly used to model variables whose distributions are not known. In all cases, we use similar methodology as described here to assess the difference of expected means between groups. Our distribution choices are subjective and informed by observations of our empirical data. We make use of this statistical methodology in the following section, where we present the results of our experiment.

5.4 Experimental Results

Our goal was to investigate the effect entity cards have on searchers' interactions with the results page. We intended to answer the following questions:

(RQ1) How does card presence and content influence users' search behaviour and perceived workload?

(RQ2) Do card properties, such as card *coherence* and *vertical diversity*, moderate the effect cards have on search behaviour and workload?

For our analysis, user search behaviour is defined as the interaction with different components of the *initial SERP*, as quantified by the measures outlined in section 5.3.3. Similarly, perceived workload is operationalised through post-task questionnaire responses, as described in the same section.

5.4.1 Preliminary Results

Table 5.5 displays summary statistics for all outcome measures across experimental conditions. We point out that participants in our study tend to not interact with entity cards and tend to not reformulate queries (i.e., the median values for `num_card_clicks`, `num_card_hovers` or `num_reform_queries` are all 0). Participants interacted with the card in less than 9% of HITs, and they issued another query to the system in less than 11% of HITs. This is not unexpected as, on one hand, participants were not asked to interact with the card, but with the web results list, and on the other hand, even though retrieved using an ambiguous query and therefore topically diverse, web results were at least partially related to the primary meaning of their assigned topic, offering enough information to complete the task. Given the few card and re-querying interactions we observed, we do not analyse card or query interaction measures further.

		Absent	*On topic*	*Off topic*	*Diverse*	*Non-diverse*	*Coherent*	*Non-coherent*
`num_card_clicks`	μ	0.000	0.239	0.126	0.239	0.190	0.239	0.195
	$\tilde{x}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	σ	0.000	0.636	0.579	0.636	0.483	0.636	0.527
`num_card_hovers`	μ	0.000	0.578	0.420	0.578	0.609	0.578	0.667
	$\tilde{x}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	σ	0.000	1.032	0.941	1.032	0.934	1.032	1.129
`num_web_clicks`	μ	13.806	13.176	12.691	13.176	12.681	13.176	12.652
	$\tilde{x}$	7.000	7.000	7.000	7.000	6.000	7.000	6.000
	σ	16.673	15.845	15.874	15.845	15.207	15.845	16.071
`num_web_hovers`	μ	15.433	16.076	15.416	16.076	15.398	16.076	15.554
	$\tilde{x}$	10.000	11.000	10.000	11.000	9.000	11.000	9.000
	σ	15.598	15.741	16.108	15.741	16.075	15.741	15.495
`num_rel_web_docs`	μ	6.165	5.820	5.539	5.820	6.251	5.820	5.517
	$\tilde{x}$	3.000	4.000	4.000	4.000	4.000	4.000	4.000
	σ	6.745	6.174	6.071	6.174	6.959	6.174	6.137
`num_reform_queries`	μ	0.099	0.100	0.160	0.100	0.151	0.100	0.075
	$\tilde{x}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	σ	0.343	0.364	0.481	0.364	0.414	0.364	0.291
`web_scroll_depth`	μ	24.979	23.585	23.532	23.585	23.613	23.585	23.708
	$\tilde{x}$	19.000	15.000	16.000	15.000	18.000	15.000	18.000
	σ	20.307	20.004	19.541	20.004	19.459	20.004	19.735
`time_to_web_click`	μ	22.625	24.894	28.438	24.894	23.656	24.894	24.511
	$\tilde{x}$	11.644	10.672	14.385	10.672	12.985	10.672	12.893
	σ	36.058	45.593	54.980	45.593	35.267	45.593	35.835
`time_on_initial_serp`	μ	178.121	162.374	170.334	162.374	160.919	162.374	180.094
	$\tilde{x}$	141.272	125.709	133.146	125.709	125.414	125.709	146.873
	σ	138.011	125.826	135.779	125.826	125.370	125.826	154.030

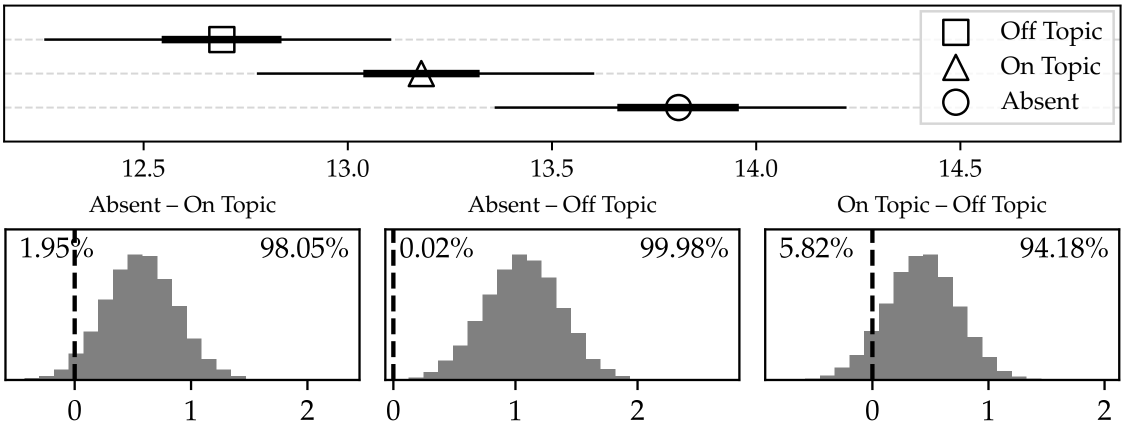
workload	μ	6.816	6.836	7.506	6.836	8.059	6.836	7.411
	$\tilde{x}$	6.667	6.333	7.333	6.333	8.000	6.333	7.500
	σ	3.915	3.834	3.801	3.834	3.559	3.834	3.915
workload_mental	μ	9.000	9.242	10.015	9.242	10.785	9.242	9.865
	$\tilde{x}$	9.000	9.000	11.000	9.000	11.000	9.000	11.000
	σ	5.985	5.719	5.841	5.719	5.077	5.719	5.687
workload_physical	μ	4.694	4.931	5.227	4.931	6.373	4.931	5.697
	$\tilde{x}$	3.000	3.000	3.000	3.000	4.000	3.000	3.000
	σ	4.907	4.995	5.310	4.995	5.657	4.995	5.437
workload_temporal	μ	7.560	7.381	8.606	7.381	9.036	7.381	8.584
	$\tilde{x}$	6.000	7.000	9.000	7.000	9.000	7.000	9.000
	σ	5.563	5.347	5.620	5.347	5.219	5.347	5.613
workload_performance	μ	2.937	3.163	3.487	3.163	3.265	3.163	3.004
	$\tilde{x}$	2.000	2.000	2.000	2.000	2.000	2.000	2.000
	σ	3.470	3.840	3.911	3.840	3.228	3.840	3.164
workload_effort	μ	10.979	10.820	12.063	10.820	12.556	10.820	11.360
	$\tilde{x}$	12.000	12.000	13.000	12.000	13.000	12.000	12.000
	σ	6.385	5.964	6.056	5.964	5.220	5.964	5.956
workload_frustration	μ	5.729	5.478	5.636	5.478	6.341	5.478	5.955
	$\tilde{x}$	4.000	3.000	3.000	3.000	5.000	3.000	5.000
	σ	5.088	5.299	5.087	5.299	5.328	5.299	5.085

Table 5.5: Sample summary statistics with respect to experimental output variables we measured (μ is the mean, $\tilde{x}$ is the median, and σ is the sample standard deviation). Note that the ***On topic***, ***Diverse*** and ***Coherent*** conditions are exactly the same, as defined in our experimental design (i.e., the ***On topic*** cards are both *diverse* and *coherent*) but are duplicated here for ease of comparison.

5.4.2 Card Content

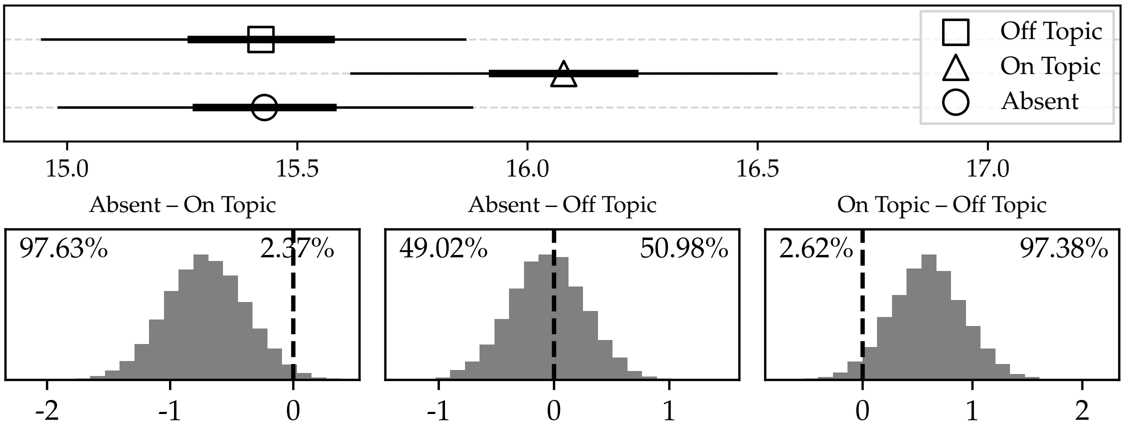
In this section, we present our analysis with respect to manipulations of card content. For each experimental measure, we briefly comment on our notable findings below. Overall, we find that card content has an effect on users' search interactions and a much stronger effect on perceived task workload.

num_web_clicks

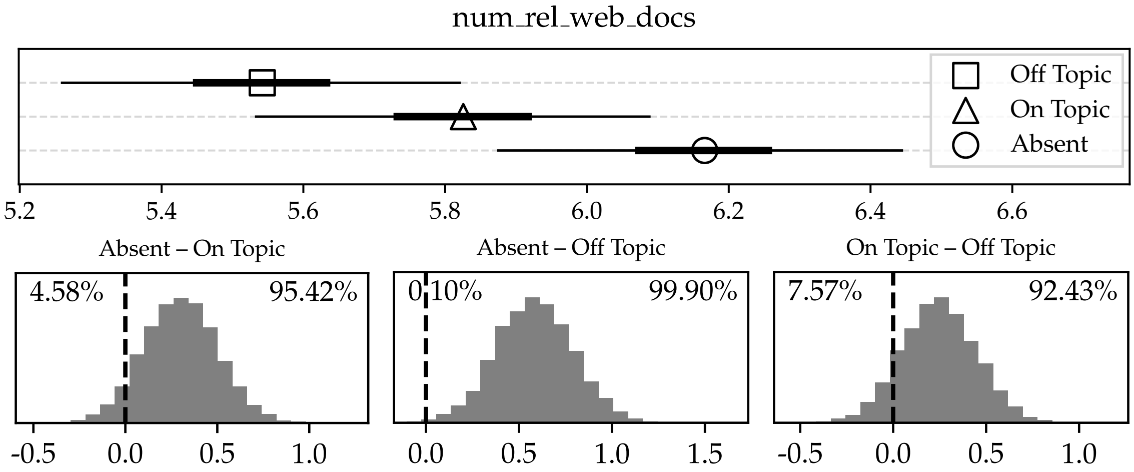


Our analysis suggests that searchers typically click on more results when the card is absent from the page, as compared to when either an on topic or off topic card is shown – based on our analysis, on average, searchers inspect one less document when the card is off topic compared to when the card is absent.

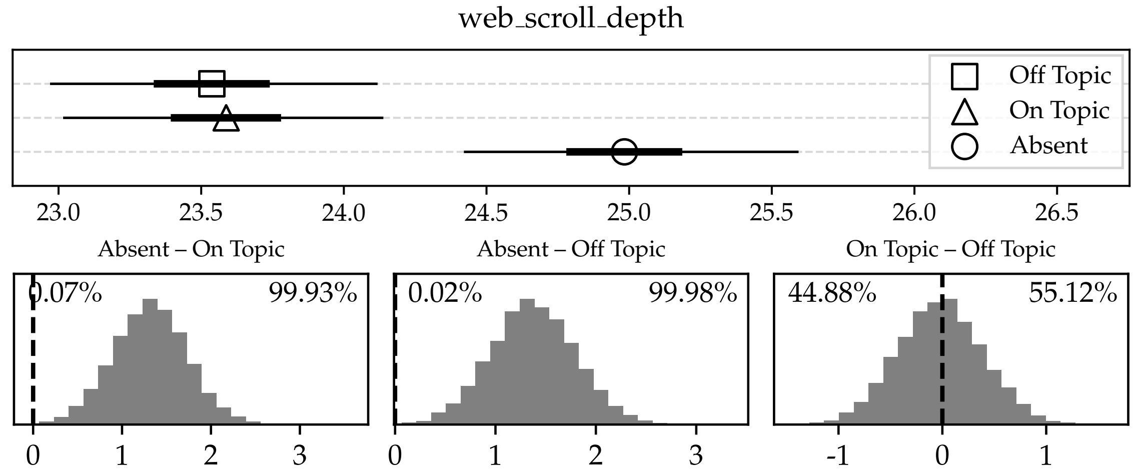
num_web_hovers



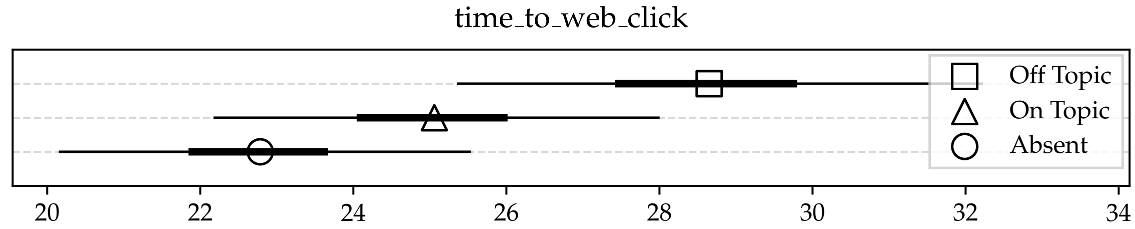
The presence of an on topic card increases the number of mouse hovers over web results, compared to both the absent and off topic conditions. This suggests that searchers, even though click on fewer results, inspect the results in more detail when an on topic card is displayed on the page. This finding supports prior work in aggregated search (Arguello and Capra, 2016), where on topic aggregated search blocks lead to more engagement with the results page in general.

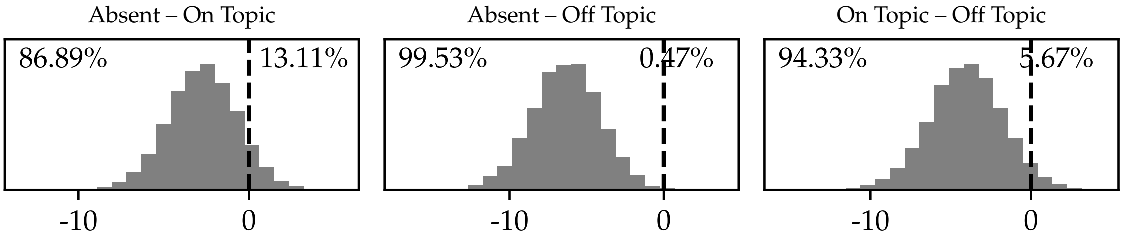


The number of results marked as relevant is, on average, lower when entity cards, either on or off topic, are shown on the results page. An explanation for this might be that for off topic cards, the entire page is perceived as lower quality and therefore abandoned sooner, whereas for on topic cards, searchers' information need is partially addressed through the card, thus, fewer documents are inspected.

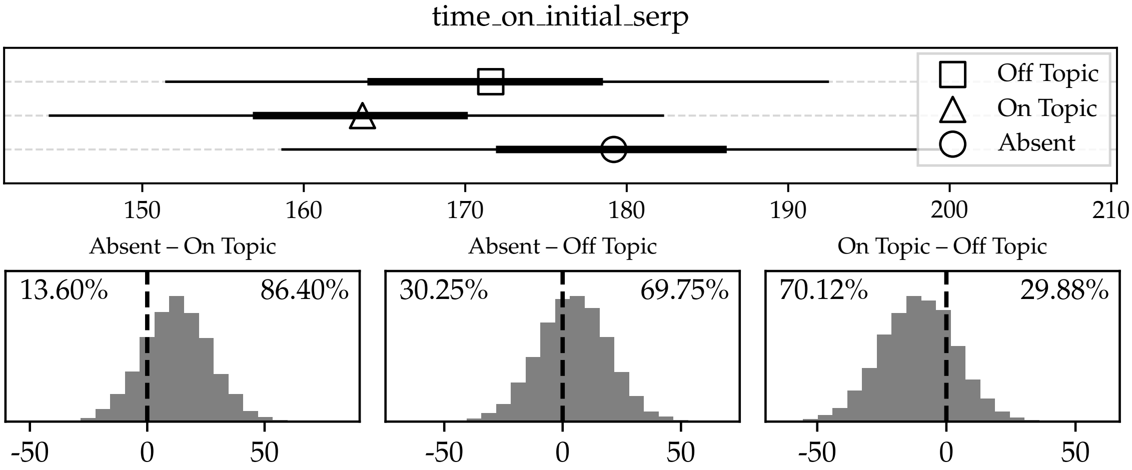


Scroll depth is higher when cards (either on or off topic) are absent from the results page. As with the number of relevant documents, an explanation for this might be that the on topic card satisfies the information need, whereas the off topic card leads to a faster assessment of page quality, both of which, in turn, lead to a decrease in the number of documents explored.

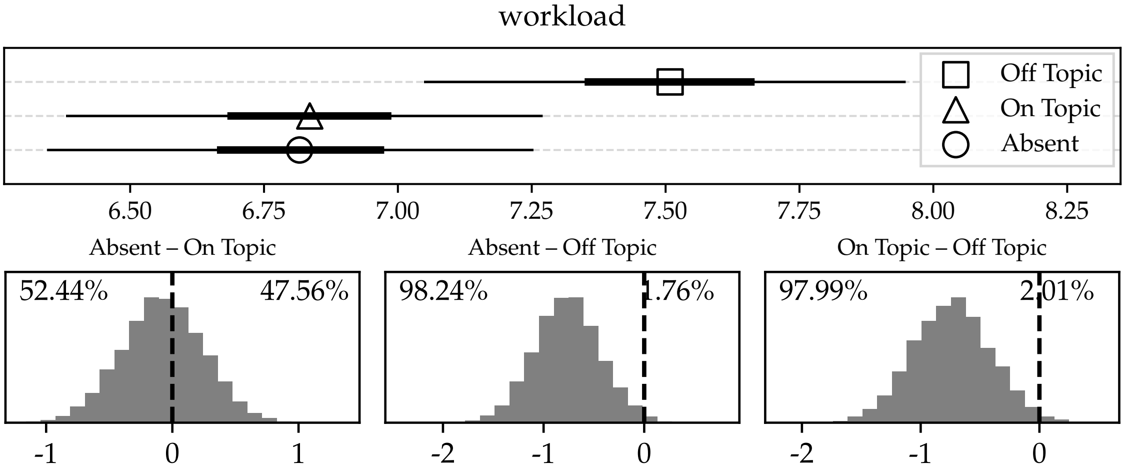




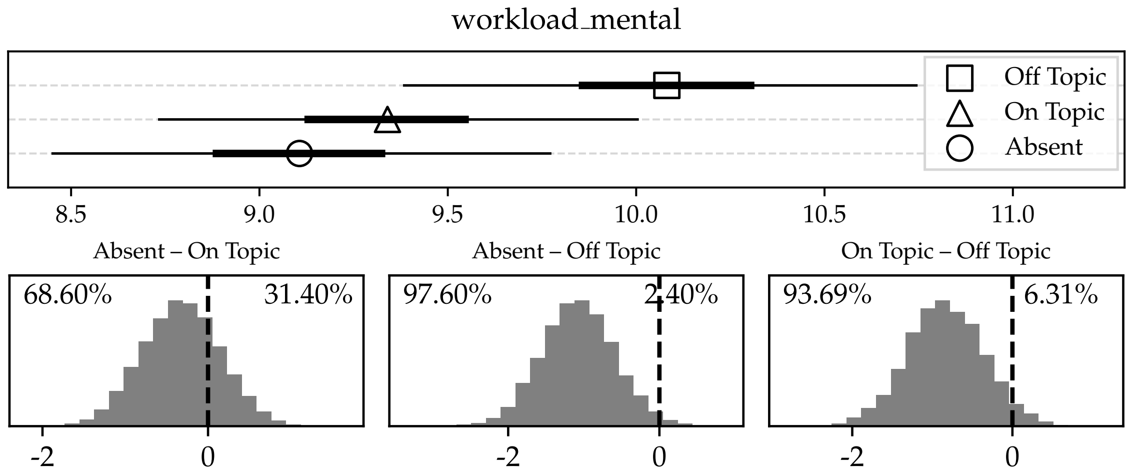
Searchers typically take more time to issue a first click on web results when an off topic card is present on the results page. Similar effects are observed between on topic and off topic cards. Based on our estimation, the off topic card increases time to first click, on average, compared to the absent condition, by approximately six seconds.



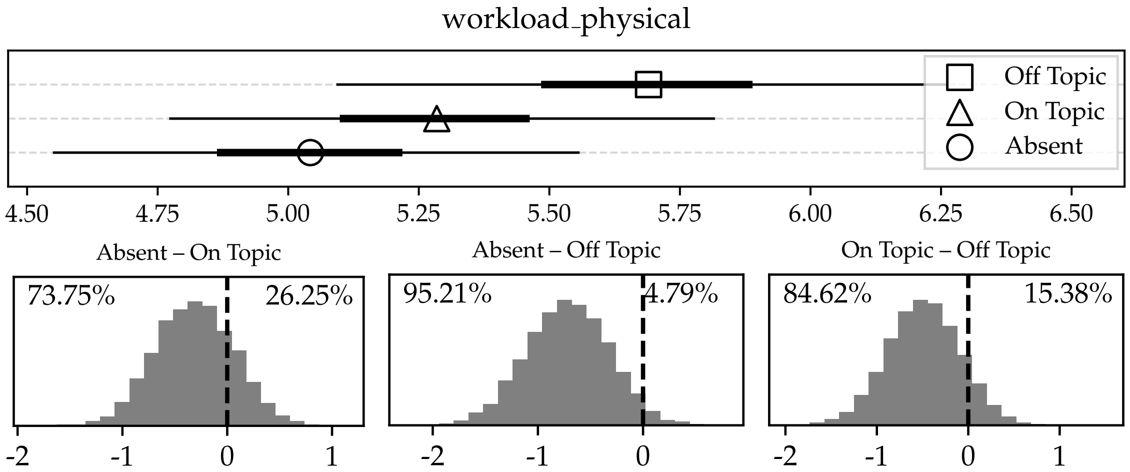
Less noticeable effects can be observed on time spent exploring the initial SERP (or total task duration, when no additional queries were issued by the searcher) across experimental conditions.



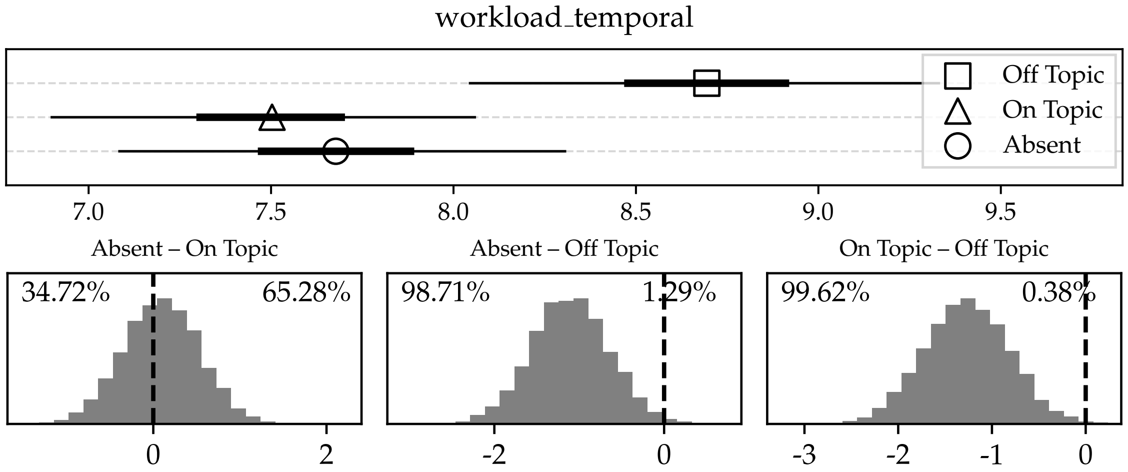
With respect to overall workload, we find that the off topic card increases perceived workload over both on topic and absent conditions. We also observe that the on topic card does not increase workload compared to the absent condition, even though the card requires searchers to process additional information. We observe similar findings with respect to several individual components of perceived workload (mental, temporal, performance and effort) as shown next.



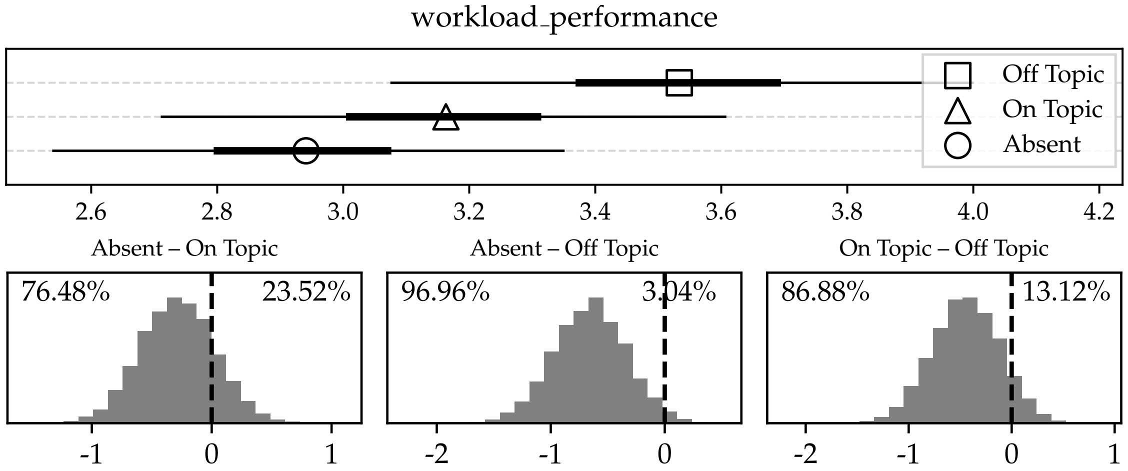
With respect to mental and perceptual activity, the off topic card increases perceived workload over both on topic and absent conditions, on average. This is not surprising, given the dissonance between card information and assigned topic.



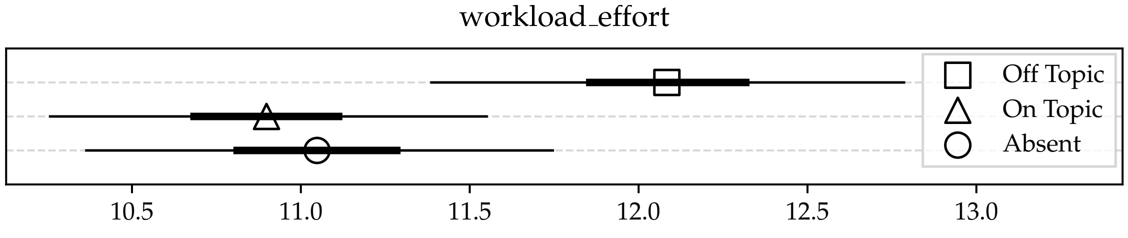
As with mental workload, participants perceive their tasks as being more laborious when the entity card shown on the page is off topic, compared to the absent condition.

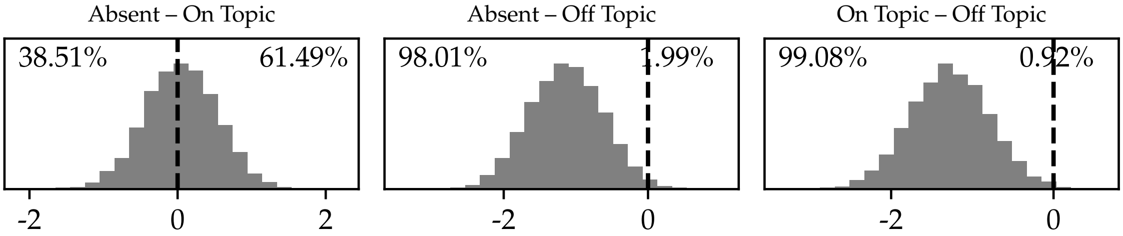


With respect to the time pressure experienced by participants, the off topic card, on average, increases perceived workload relative to both on topic and absent experimental conditions. An explanation for this might be that participants spend more time processing the information in the off topic card (as suggested above by a longer period of time taken to issue a first click on web results when the card is off topic) and therefore perceive to have less time for their main assigned task, which is selecting documents.

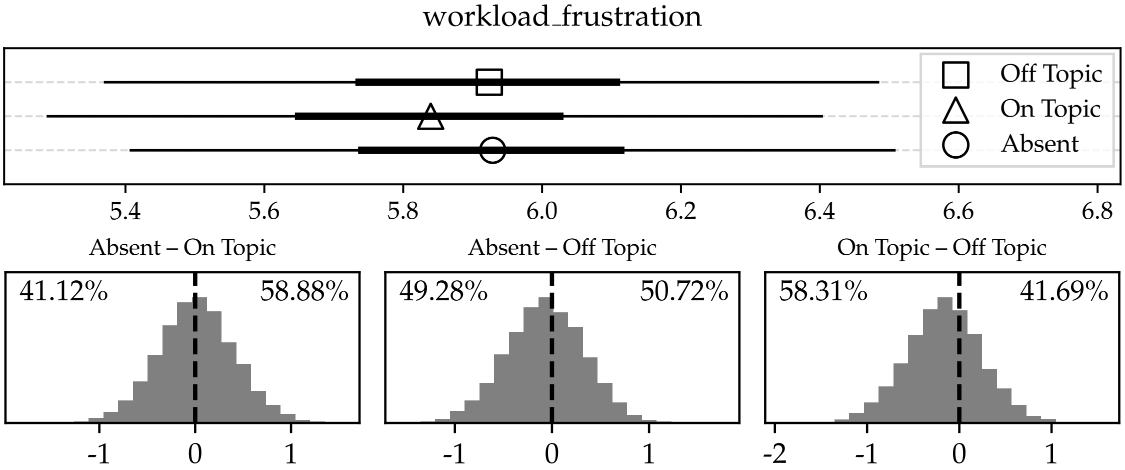


The presence of an off topic card on the results page, on average, leads to participants perceiving their task completion as less successful compared to when the results page that does not display a card at all.



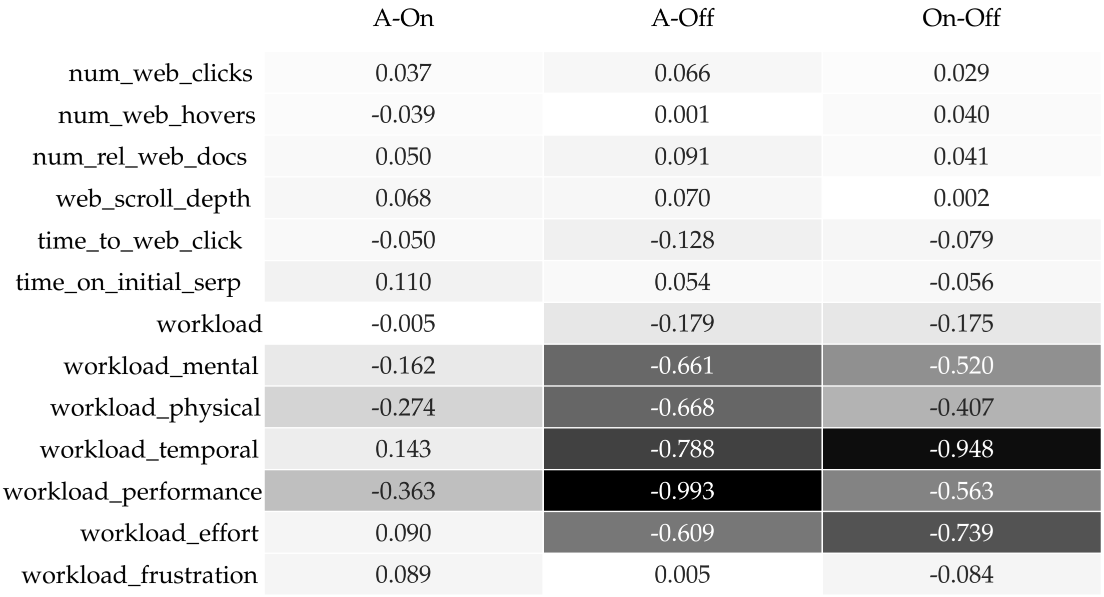


The off topic card also leads to participants perceiving more effort necessary in accomplishing their task, overall, when compared to both the absent and the on topic experimental conditions. Together with the temporal and performance aspects of workload, this suggests that when the card is off topic, task effort is not only increased, but also partitioned between processing the information in the card and completing the task (i.e., finding relevant documents), instead of focused on the main goal of the task.



Our analysis reveals no differences, on average, with respect to how insecure, discouraged or irritated participants assessed themselves to be in their task, across experimental conditions.

To summarise, with respect to search interactions, we find that in the absent condition, more clicks are issued on web results, more documents are marked as relevant, and searchers scroll to a lower depth on the page than when an off topic entity card is displayed. An explanation for this behaviour can be that when an off topic card is displayed, users perceive the entire results page as less reliable and decide to terminate their task with fewer results examined. It is interesting to note that when an on topic entity card is shown, on average, participants hover over web results more, yet click on fewer results and mark fewer documents as relevant than in the absent condition. This might suggest that searchers inspect results more carefully when an on topic entity card is present. In addition, we also find that the display of an off topic entity card increases the time taken to



	A-On	A-Off	On-Off
num_web_clicks	0.037	0.066	0.029
num_web_hovers	-0.039	0.001	0.040
num_rel_web_docs	0.050	0.091	0.041
web_scroll_depth	0.068	0.070	0.002
time_to_web_click	-0.050	-0.128	-0.079
time_on_initial_serp	0.110	0.054	-0.056
workload	-0.005	-0.179	-0.175
workload_mental	-0.162	-0.661	-0.520
workload_physical	-0.274	-0.668	-0.407
workload_temporal	0.143	-0.788	-0.948
workload_performance	-0.363	-0.993	-0.563
workload_effort	0.090	-0.609	-0.739
workload_frustration	0.089	0.005	-0.084

Figure 5.6: Mean effect size (*Cohen's d*) for card content experimental manipulations (*A* – absent, *On* – on topic, *Off* – off topic). Sign shows the direction of the effect. Cell colour is proportional to the absolute effect size. Cohen (1992) and Sawilowsky (2009) suggest that absolute effect sizes larger than 0.8 can be interpreted as large effects. Manipulating card content has relatively weak effects on measures of search behaviour, however, stronger effects can be observed on different dimensions of perceived workload.

issue a first click on web results, on average, by approximately 6 seconds, which suggests that searchers are taking additional time to process the card that is dissonant with their assigned search topic, as compared to the absent condition.

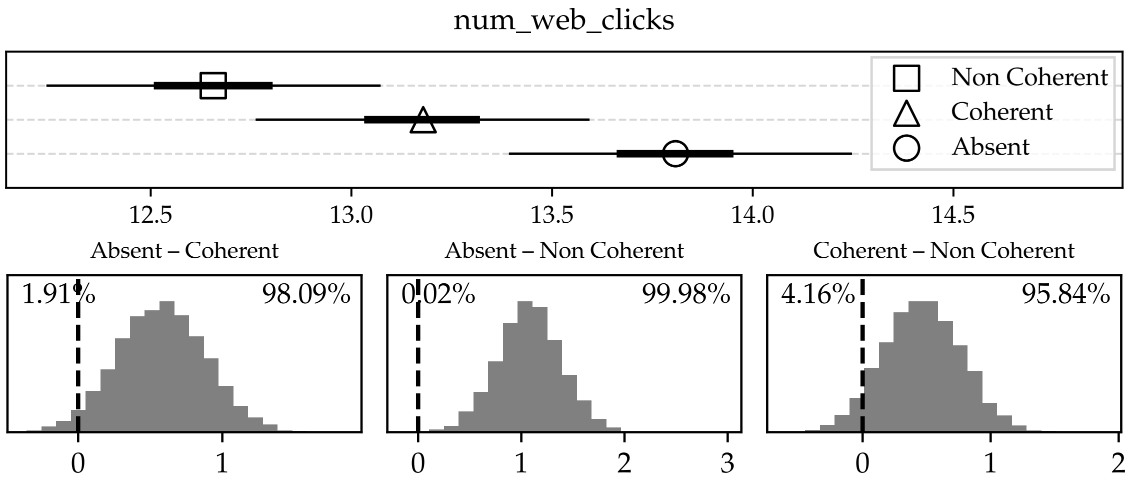
With respect to task effort, we find that off topic cards increase almost all dimensions of perceived workload. In the *temporal* and *effort* dimensions, off topic cards increase workload compared to their on topic counterparts also. This is interesting because searchers appear less engaged in their task (i.e., fewer clicks, lower scroll depth), on average, when the card is off topic, but overall perceive higher levels of workload across dimensions. It is also interesting to note that on topic cards do not seem to increase workload, even though they require searchers to process additional information, which is not directly related to successful task completion (i.e., selecting results).

Finally, the effects we discuss above vary in size, and as such, in figure 5.6 we report *Cohen's d* as a standardised measure of (mean) effect size in order to assess relative effect sizes across outcome measures. We note that entity card content effects on measures of search interaction are relatively small. However, effects on different dimensions of perceived workload are much stronger – Cohen

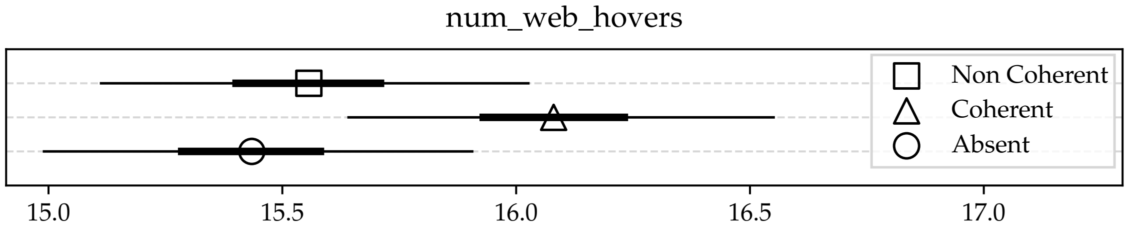
(1992) and Sawilowsky (2009) suggest that absolute effect sizes larger than 0.8 can be interpreted as large effects. Overall, we conclude that the effect entity card content has on both search interactions and perceived workload is not due to chance, but that this effect is much stronger on perceived workload than on measures of search behaviour.

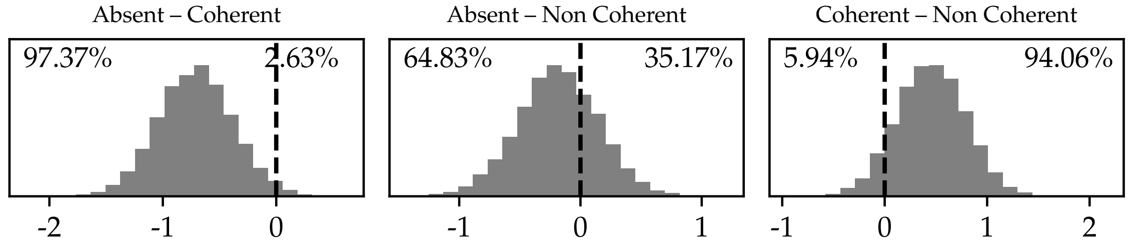
5.4.3 Card Coherence

In this section, we present our analysis with respect to manipulations of card coherence. Similar to the previous section, we compare differences between the *absent*, *coherent* and *non-coherent* experimental conditions. Note that the *coherent* condition is identical to the *on topic* condition in the previous section, yet the results presented here (i.e., when comparing *absent* to *coherent*) vary marginally from the previous section due to sampling error in our modelling approach. We focus primarily on differences between *coherent* and *non-coherent* cards, as estimating these differences is our main objective.

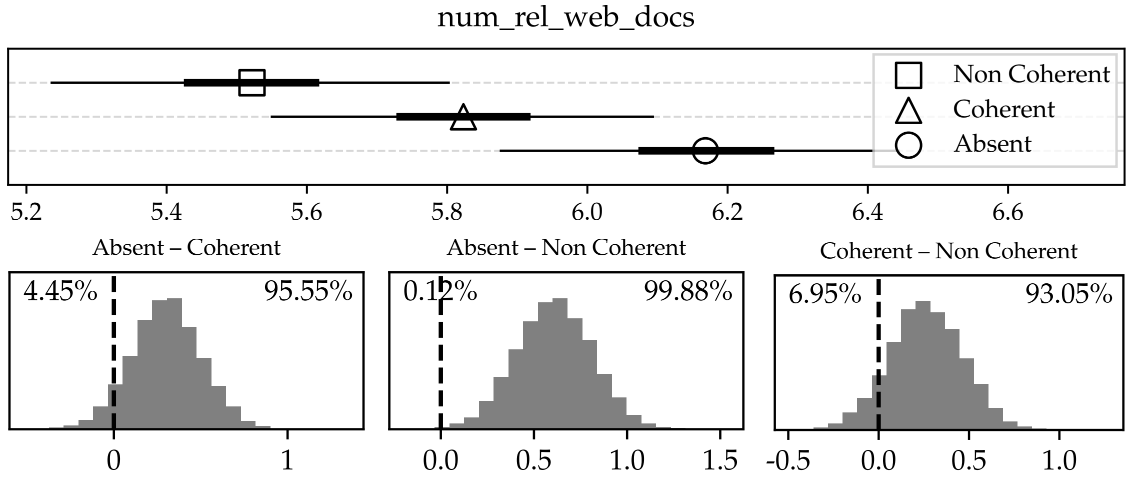


Our analysis suggests that non-coherent entity cards lead to fewer examinations of web results, on average, as reflected in the number of clicks, when compared to coherent or absent cards. As with off topic cards, searchers might assess the quality of the page as being unreliable, based on the quality of the card, and therefore explore the results page less when a non-coherent card is displayed.

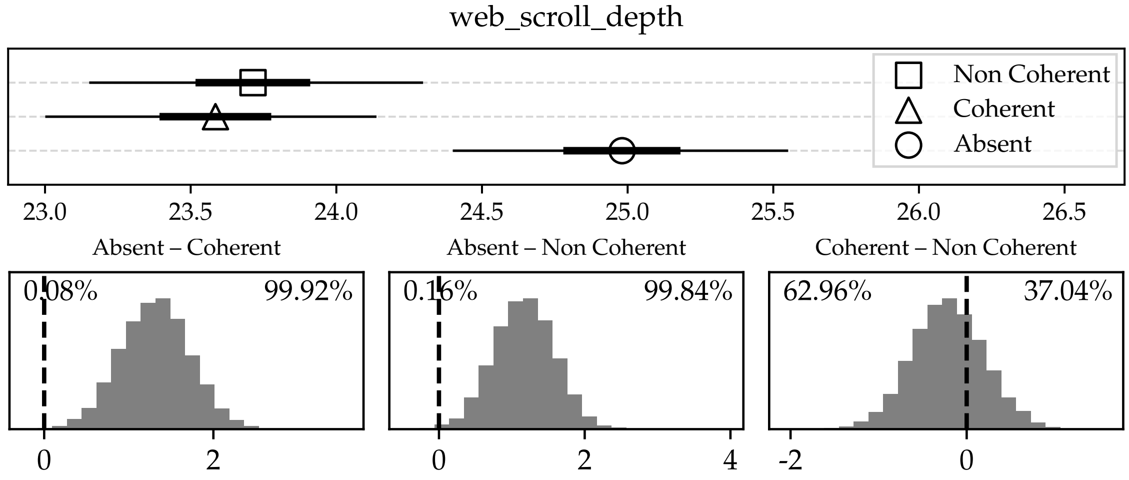




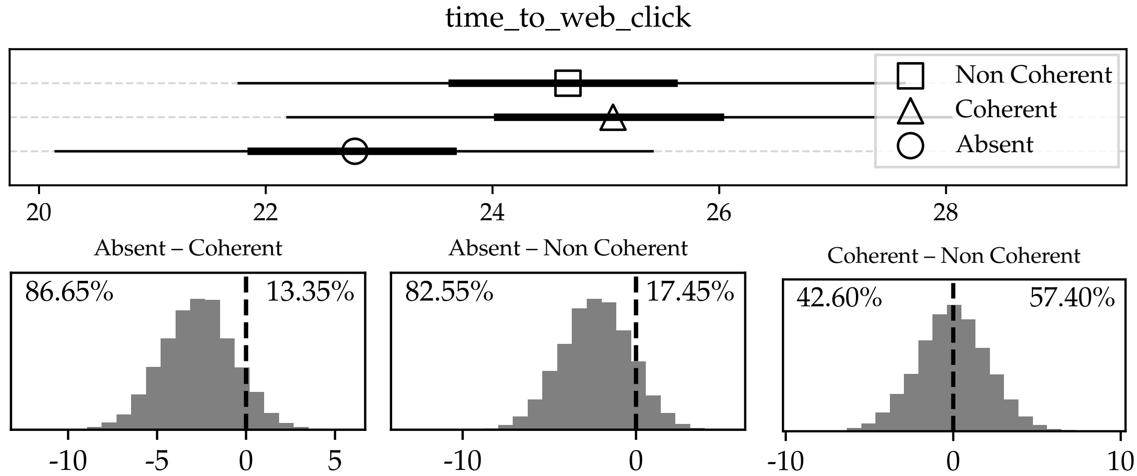
Similarly, reflected in the number of mouse hovers on web results, there is an indication that non-coherent cards lead to fewer inspections of web results, on average, compared to coherent cards.



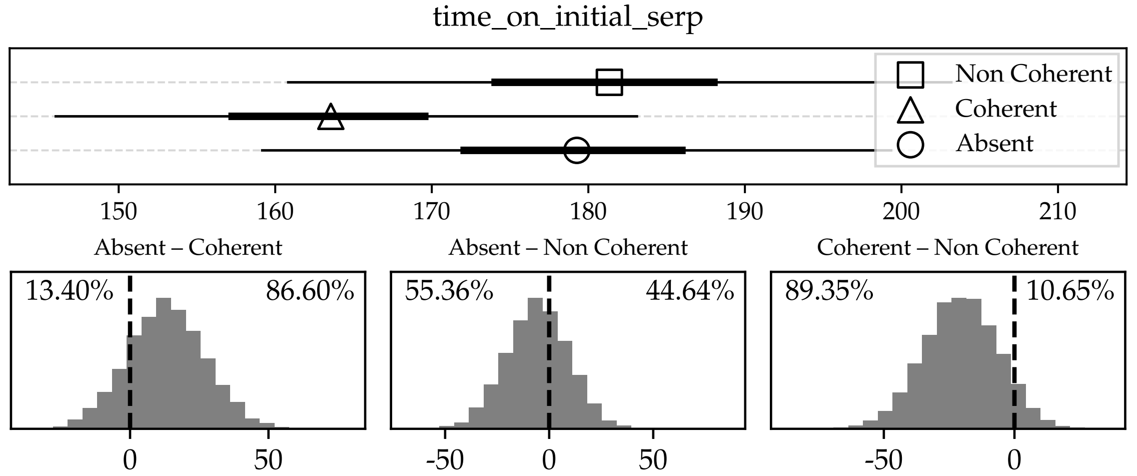
A similar indication is reflected in the number of web results marked as relevant, with non-coherent cards potentially leading to fewer documents inspected and marked as relevant, on average.



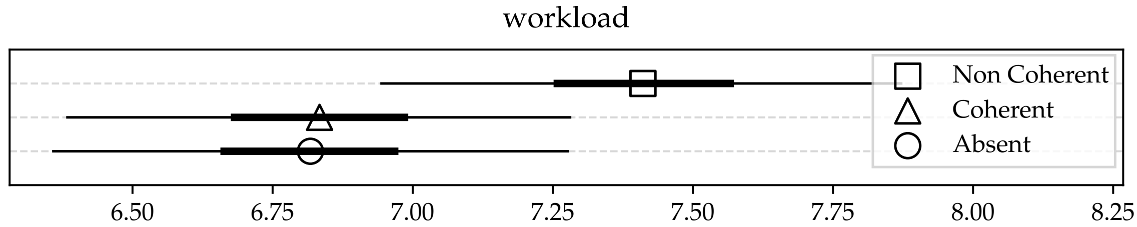
With respect to scroll depth, no clear differences between coherent and non-coherent cards, on average, are revealed by our analysis.

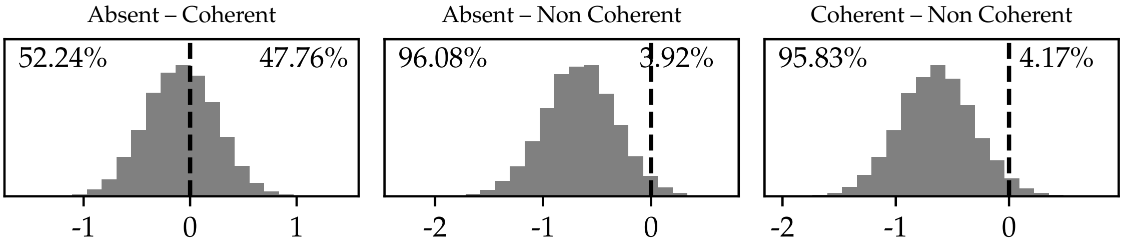


As with scroll depth, no clear differences in time taken before engaging with web results, on average, are revealed by our analysis.

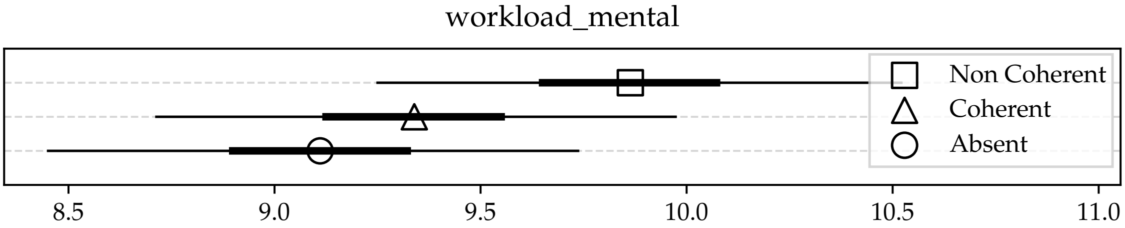


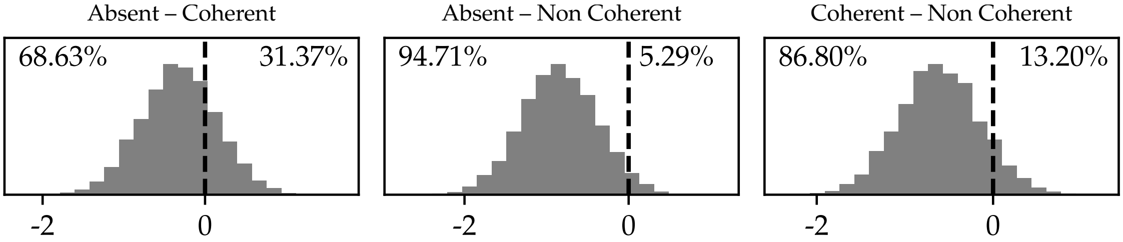
Similarly, total time spent on the initial SERP is not revealed to be different across conditions, on average, given our data and modelling approach. Overall, with respect to measures of search behaviour, non-coherent cards appear to generate similar effects as off topic cards, with the exception of task duration and time taken before a first click is issued on the results. This suggests that searchers notice the dissonant elements of the card, which in turn affects their assessment of page quality, but given that most of the content within the card is informative and on topic, less time is spent processing the entity card overall.



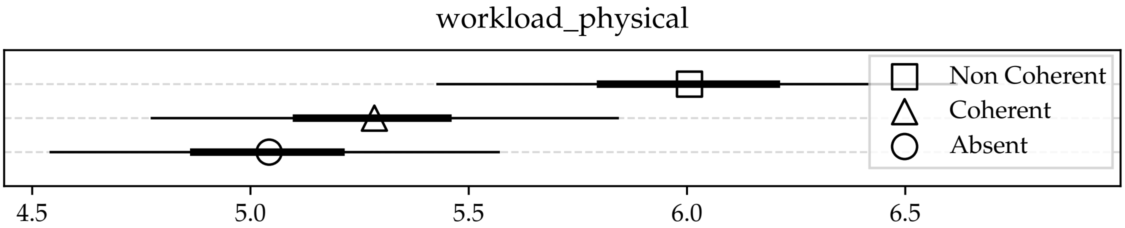


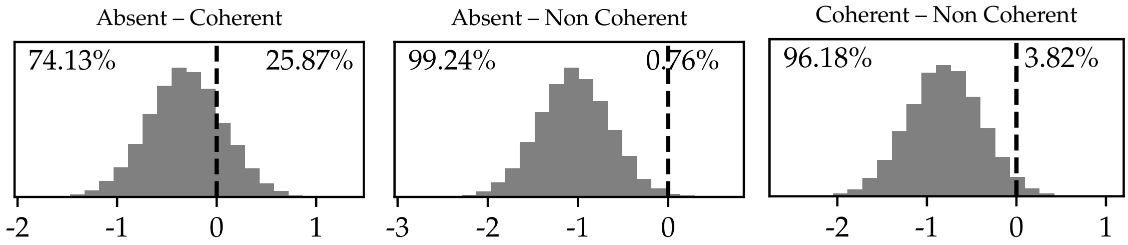
With respect to overall workload, we find that non-coherent cards, on average, increase searchers' perceived workload, compared to coherent cards. This can be explained by the dissonant nature of information contained within the card, which requires searchers to actively determine whether the card is useful to their assigned topic or not. We also find that workload is increased, on average, compared to the absent condition as well, which suggests that placing a non-coherent card on the results page can be detrimental to user experience, on average. However, few dimensions of workload are affected by card coherence (physical and temporal) as shown next.



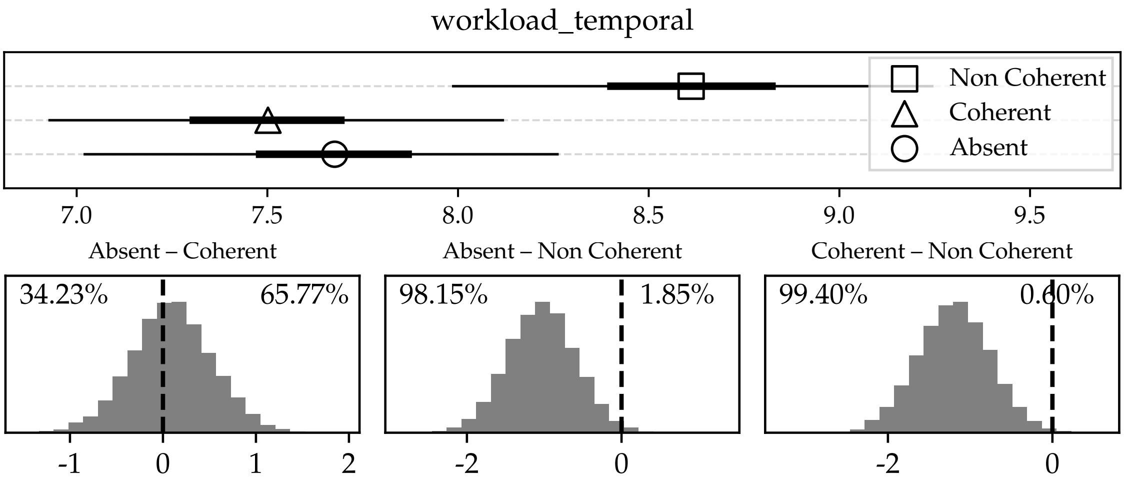


Our analysis does not reveal differences regarding perceived mental effort between the coherent and non-coherent experimental conditions, on average.

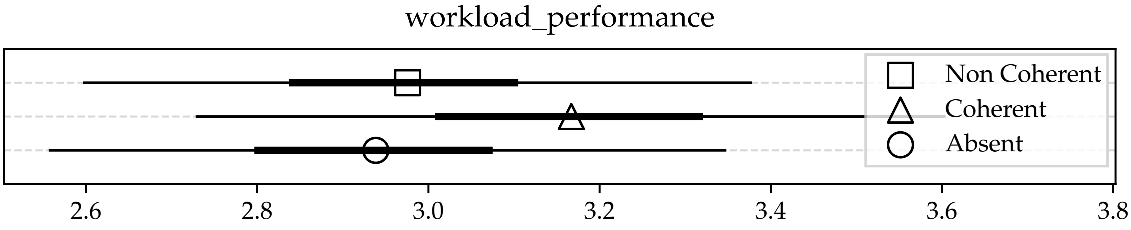


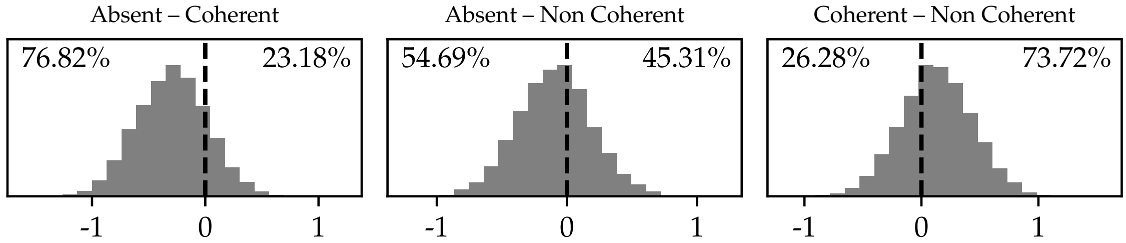


Interestingly, the physical dimension of workload is, on average, perceived as higher when non-coherent cards are displayed on the results page, compared to coherent cards. This suggests that searchers consider their tasks as more laborious when the entity card displayed on the results page is not coherent, perhaps because they need to actively engage in determining whether the card is useful to their assigned search topic or not.

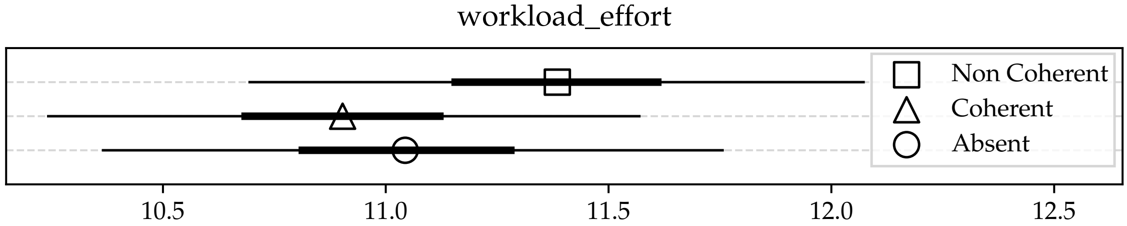


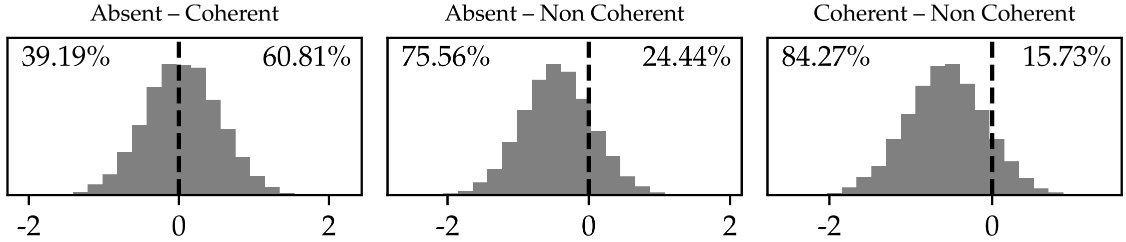
Similarly, more time pressure is perceived (i.e. searchers perceive the task as being time constrained), on average, when a non coherent card is shown on the results page, compared to a coherent card. As with off topic cards, an explanation for this might be that searchers spend more time processing the information within the non-coherent card, in an effort to determine whether the card is useful, and spend less time on their main task of finding relevant web results.



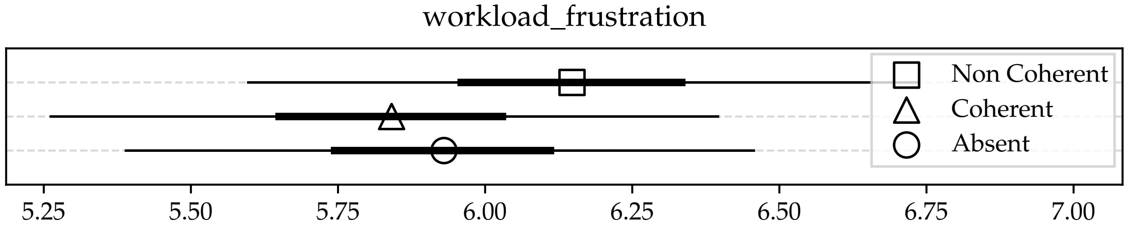


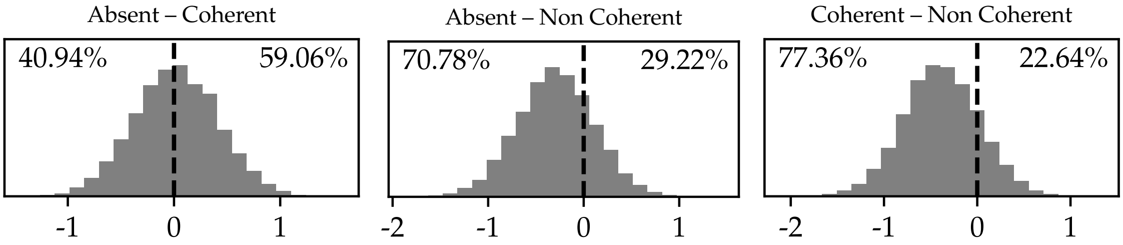
Our analysis does not reveal differences with respect to perceived performance, on average, between coherent and non-coherent conditions. Compared to off topic cards, it seems that, on average, participants perceive non-coherent cards as less of an impediment in their attempt to successfully complete the assigned task.





Our analysis does not reveal differences with respect to overall perceived effort levels, on average, between coherent and non-coherent cards, which is surprising given that with respect to perceived physical workload, non-coherent cards appear to increase demand. This is perhaps an indication of how difficult assessing measures of workload independently is in this context. In addition, compared to off topic cards, which increase perceived levels of effort on average (i.e., effort dimension of workload), non-coherent cards appear to have a weaker effect, which again suggests that searchers are able to process the results page with less effort when non-coherent entity cards are present, as compared to off topic cards.



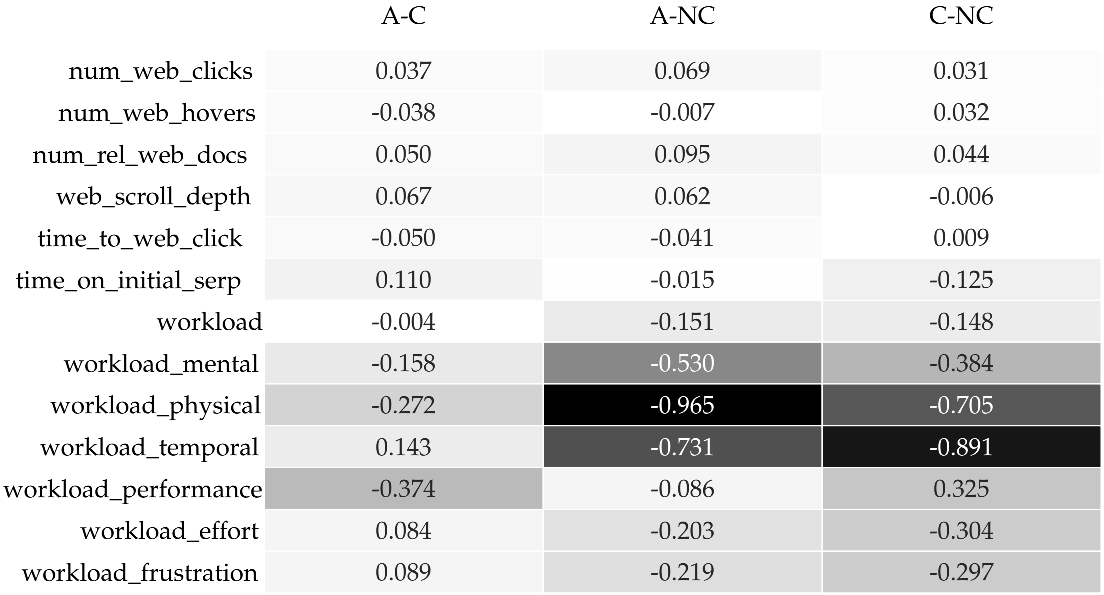


Our analysis does not reveal differences, on average, between coherent and non-coherent entity cards with respect to perceived frustration.

In summary, with respect to search interactions, we find that non-coherent cards tend, on average, to lead to fewer clicks and hovers on web results, and a lower number of documents marked as relevant, compared to coherent cards. Although it is unlikely that these differences occur due to chance, they are, however, small and difficult to interpret practically: for example, on average, the number of web clicks on pages displaying coherent cards is approximately 5.8, whereas for non-coherent cards it is approximately 5.5. These findings suggest that non-coherent cards generate effects similar to those generated by off topic cards, with respect to search behaviour. Unlike off topic cards, non-coherent cards do not appear to increase total time spent on the initial SERP, on average, or the duration of time before a click is issued on the results page. Together, these findings suggest that non-coherent influence searchers' perception of the results page as being unreliable to a lesser extent than off topic cards.

With respect to overall perceived workload, our analysis reveals that non-coherent cards tend to lead, on average, to higher self-assessed estimates of task effort compared to coherent or absent cards. However, compared to card content manipulations, and in particular off topic cards, we find that fewer dimensions of workload are affected by the coherence of entity cards.

Figure 5.7 reports estimates of standardised effect sizes for differences in the outcome measures we explored. Our manipulations of card coherence have strong effects on the physical and temporal dimensions of perceived workload, but have weaker effects on workload overall, and very weak effects on measures of search interaction. Like in the case of card content, we conclude that the effects of entity card coherence are not due to chance, but that these effects are limited to fewer dimensions of search behaviour and workload than in the case of card content manipulations, and that the strongest effects of card coherence are on the perceived physical and temporal aspects of workload.

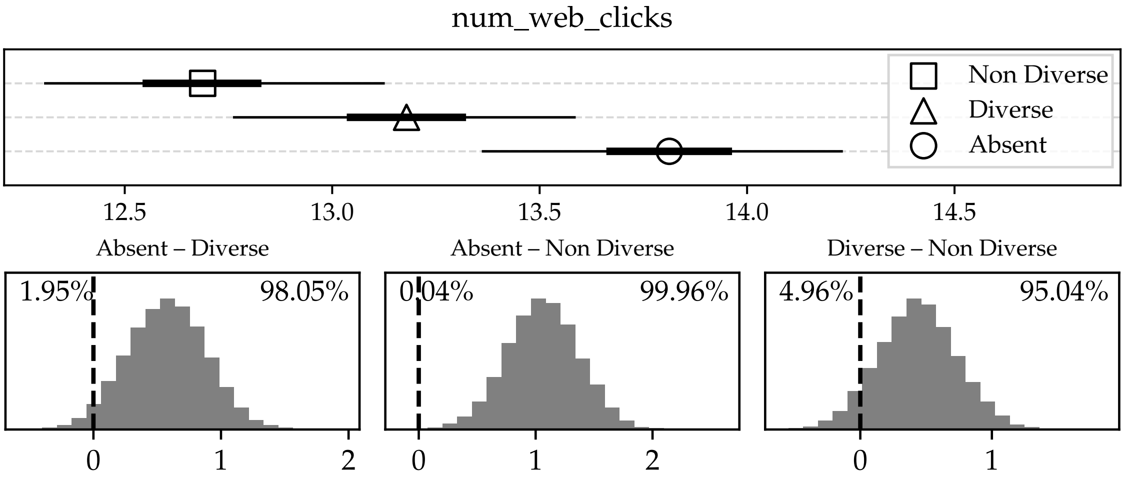


	A-C	A-NC	C-NC
num_web_clicks	0.037	0.069	0.031
num_web_hovers	-0.038	-0.007	0.032
num_rel_web_docs	0.050	0.095	0.044
web_scroll_depth	0.067	0.062	-0.006
time_to_web_click	-0.050	-0.041	0.009
time_on_initial_serp	0.110	-0.015	-0.125
workload	-0.004	-0.151	-0.148
workload_mental	-0.158	-0.530	-0.384
workload_physical	-0.272	-0.965	-0.705
workload_temporal	0.143	-0.731	-0.891
workload_performance	-0.374	-0.086	0.325
workload_effort	0.084	-0.203	-0.304
workload_frustration	0.089	-0.219	-0.297

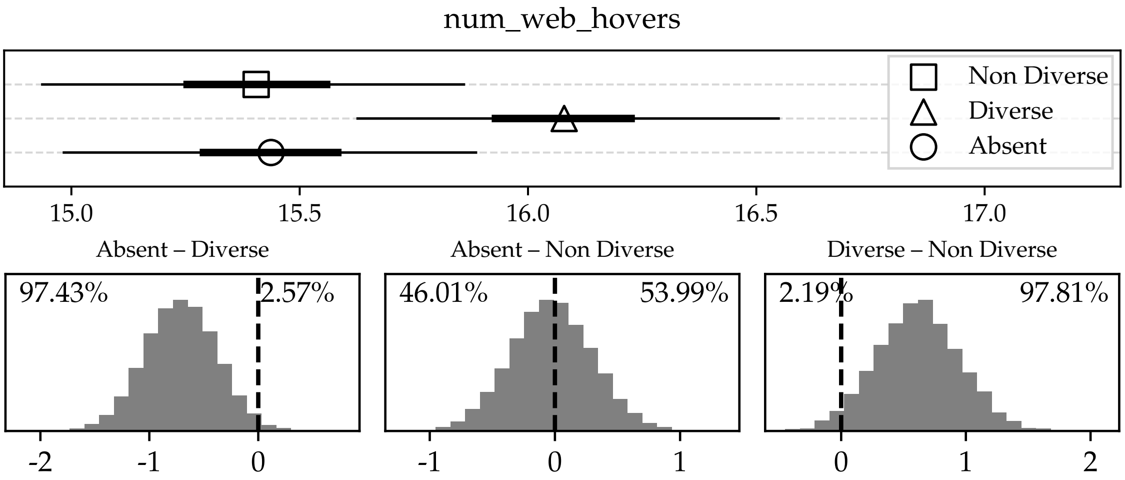
Figure 5.7: Mean effect size (*Cohen's d*) for card coherence experimental manipulations (*A* – absent, *C* – coherent, *NC* – non-coherent). Sign shows the direction of the effect. Cell colour is proportional to the absolute effect size. Cohen (1992) and Sawilowsky (2009) suggest that absolute effect sizes larger than 0.8 can be interpreted as large effects. The effects of entity card coherence manipulations are limited to fewer dimensions of search behaviour and workload than in the case of card content manipulations, with the strongest effects of card coherence being on the perceived physical and temporal aspects of workload.

5.4.4 Card Diversity

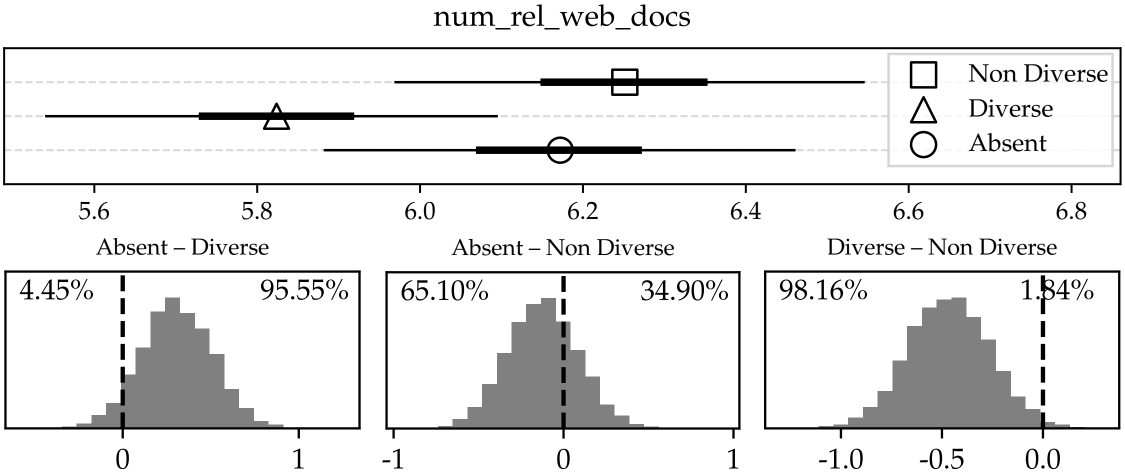
In addition to card coherence, we examined the effect card *vertical diversity* has on search interaction and workload. Similar to the previous sections, we make comparisons between the *absent*, *diverse* and *non-diverse* conditions. Note that the *diverse* condition is identical to the *on topic* or the *coherent* conditions in the previous sections (e.g., when comparing *absent* to *on topic*), yet the results presented here (i.e., when comparing *absent* to *diverse*) vary marginally due to sampling error in our modelling approach. We focus primarily on differences between *diverse* and *non-diverse* cards, as estimating these differences is our main objective.



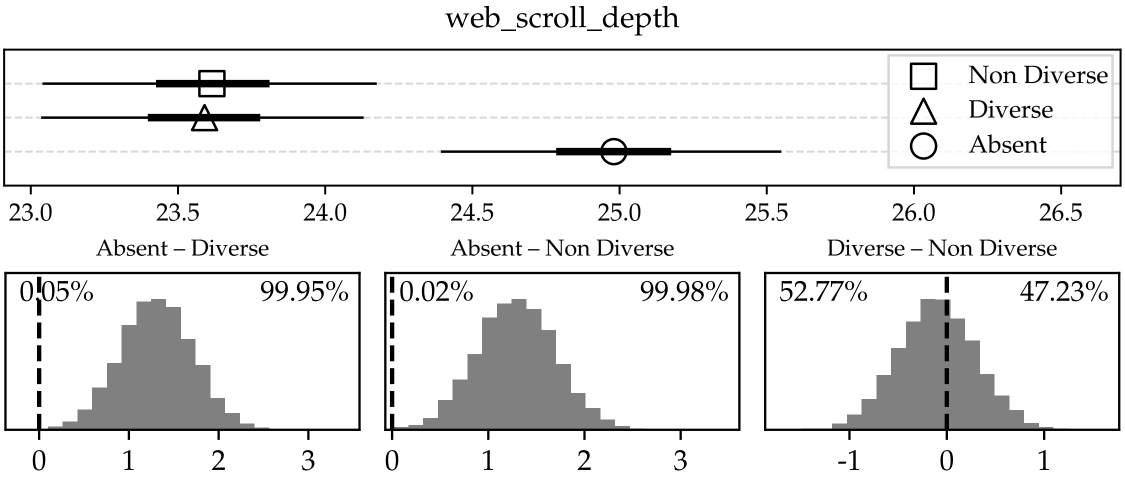
The number of clicks issued on web results is lower, on average, when entity cards shown on the results page are non-diverse, as compared to when the card is diverse. It is interesting to note that this effect is similar to the effect generated by off topic or non-coherent cards, as shown previously, even though the content within non-diverse cards is informative and related to searchers' assigned topic.



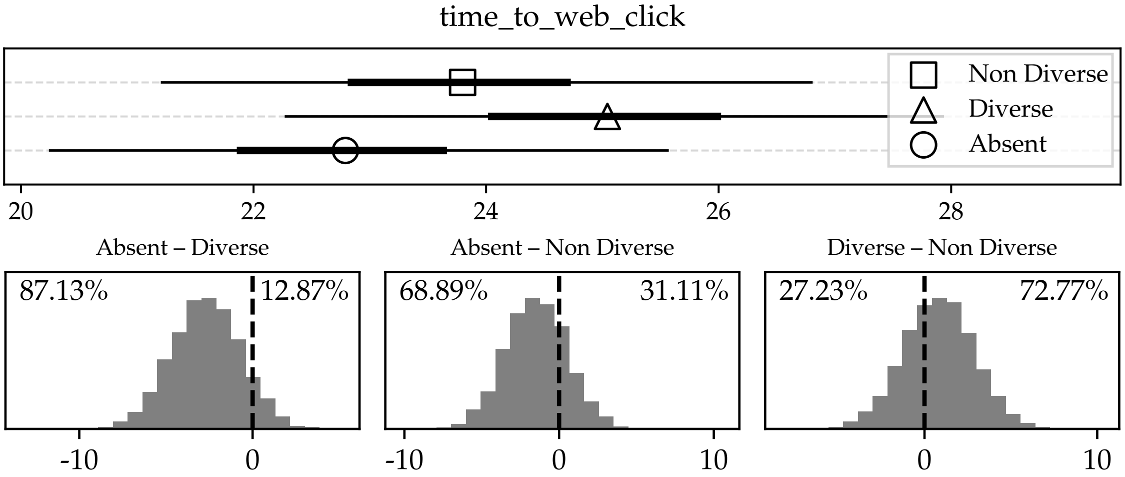
The number of mouse hovers over web results is lower when the card is diverse, on average, as compared to both the non-diverse and the absent conditions. Again, this is a similar effect to that observed for both off topic and non-coherent entity cards.



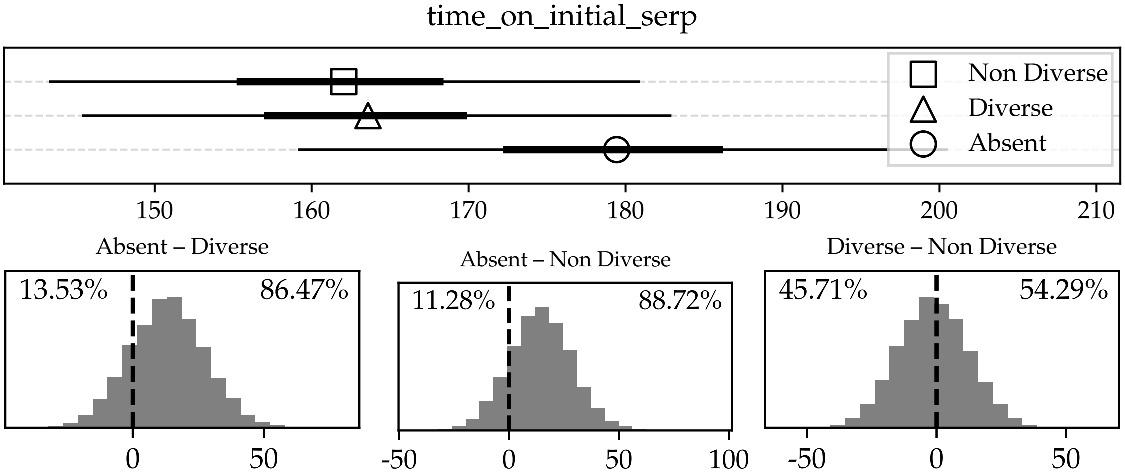
The number of documents marked as relevant is, on average, higher in the non-diverse condition, compared to the diverse condition, which suggests that the non-diverse entity card is perceived as less informative than its diverse counterpart. However, unlike the previous measures of search behaviour, this effect is dissimilar to that generated by off topic or non-coherent cards, where the number of results marked as relevant is higher in the on topic or coherent conditions. This suggests, that unlike off topic or non-coherent entity cards, non-diverse cards do no lead to overall assessments of poor page quality (in which case the page is abandoned with fewer inspected results), but rather do not satisfy searchers' information needs to the same extent as diverse cards, which leads to more results being marked as relevant.



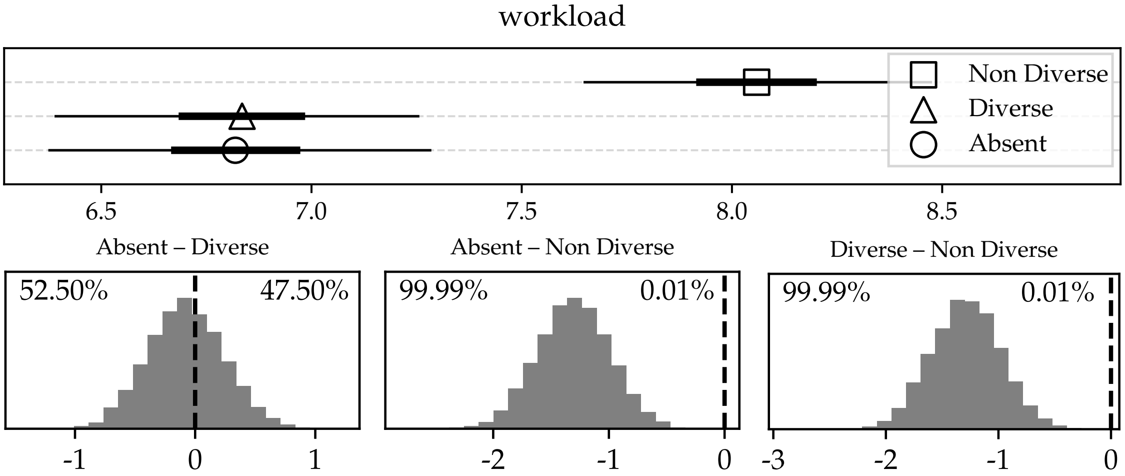
Our analysis does not reveal differences between diverse and non-diverse entity cards with respect to scroll depth, on average. Again, this is similar to previous findings regarding manipulations of card content or card coherence. Overall, our results suggest that the presence of an entity card on the results page reduces scroll depth, on average, by approximately 1 position, irrespective of card content, coherence or diversity.



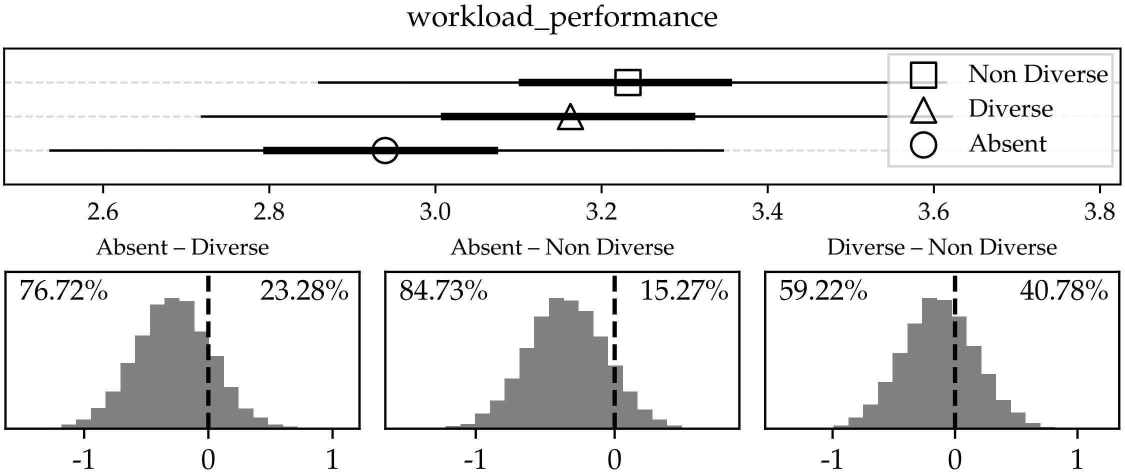
Our analysis does not reveal differences between diverse and non-diverse entity cards, on average, with respect to the duration of time before a first click is issued on web results. This is similar to the effect (or absence of effect) generated by non-coherent cards, but is dissimilar to the effect of off topic cards, which, on average, lead to higher durations before a first click.



Our analysis does not reveal differences, on average, between diverse and non-diverse entity cards with respect to the duration of time spent on the initial SERP (similar to the case of off topic or non coherent entity cards).



Our results suggest that non-diverse cards increase overall perceived workload, on average, by approximately 1.5 points (on a 20 point scale), compared to both diverse and absent conditions. This indicates that removing vertically diverse results from entity cards can be detrimental to search user experience, as these results help searchers process entity card information with less perceived effort. This is again similar with the cases of off topic or non-coherent entity cards, which both increase perceived workload compared to their on topic or coherent counterparts, respectively. We observe a very similar effect of card diversity on *all* dimensions of perceived workload, with the exception of performance, and as such report only perceived performance workload next.



Similar to both off topic or non-coherent cards, and their on topic or coherent counterparts, card vertical diversity manipulations do not lead to lower assessed task performance levels, on average, given our analysis and collected data.

In summary, with respect to measures of search behaviour, our analysis suggests that non-diverse cards are similar to both off topic and non-coherent cards, leading to fewer clicks or mouse hovers on web results, on average, than diverse

	A-D	A-ND	D-ND
num_web_clicks	0.037	0.066	0.029
num_web_hovers	-0.039	0.002	0.040
num_rel_web_docs	0.051	-0.012	-0.062
web_scroll_depth	0.067	0.066	-0.001
time_to_web_click	-0.049	-0.022	0.027
time_on_initial_serp	0.112	0.122	0.010
workload	-0.004	-0.330	-0.330
workload_mental	-0.158	-1.253	-1.128
workload_physical	-0.267	-1.556	-1.313
workload_temporal	0.142	-1.163	-1.347
workload_performance	-0.366	-0.628	-0.144
workload_effort	0.094	-0.993	-1.178
workload_frustration	0.090	-0.620	-0.676

Figure 5.8: Mean effect size (*Cohen's d*) for card diversity experimental manipulations (*A* – absent, *D* – diverse, *ND* – non-diverse). Sign shows the direction of the effect. Cell colour is proportional to the absolute effect size. Cohen (1992) and Sawilowsky (2009) suggest that absolute effect sizes larger than 0.8 can be interpreted as large effects. Similar to both card content and card coherence manipulations, our analysis reveals weak effects of card diversity manipulation on measures of search behaviour, but much larger effects on measures of perceived workload. Overall, our findings suggest that the display of non-diverse entity cards can lead to suboptimal user experience.

cards. However, with respect to the number of web results marked as relevant, diverse cards generate an effect that is dissimilar to that generated by off topic or non-coherent cards, in that on result pages with non-diverse cards, the number of results marked as relevant is higher than in the case of pages displaying diverse cards, on average. This suggests that, unlike off topic or non-coherent entity cards, non-diverse cards do no lead to overall assessments of poor system quality (in which case the search page is abandoned with fewer results marked as relevant or inspected in general), but rather do not satisfy searchers' information needs to the same extent as diverse cards, which in turn leads to more results being marked as relevant.

With respect to task effort, our analysis suggests that card diversity manipulations have an effect on perceived workload that is not due to chance. Our results indicate that non-diverse cards increase overall workload, on average, by approximately 1.5 points (on a 20 point scale), compared to both diverse and absent conditions. In addition, non-diverse cards appear to have an effect on *all* dimensions of task workload, with the exception of perceived performance levels. This finding supports the claim that removing vertically diverse results from entity cards can be detrimental to search user experience, as these results can help searchers process entity card information with less perceived effort. This is again similar with the cases of off topic or non-coherent entity cards, which both increase overall perceived workload compared to their on topic or coherent counterparts.

As in the previous sections, we report the mean standardised effect size of card diversity manipulations on experimental outcome measures in Figure 5.8. Similar to both card content and card coherence manipulations, our analysis reveals weak effects of card diversity manipulation on measures of search behaviour, but much larger effects on measures of perceived workload. Sawilowsky (2009) suggests that effect sizes above 1.2 can be considered very large and our analysis reveals multiple such effects. In particular, our findings suggest that card diversity has a very large effect on the physical and temporal dimensions of perceived workload, increasing self-assessed measures by 1.5 points, on average, compared to diverse entity cards displayed on the results page.

5.5 Discussion and Conclusions

In this section we discuss the implications of our findings. We follow a similar structure as in the previous section, discussing our results regarding card content followed by the implications of our findings related to card properties.

Card content. In the previous sections, we reported on the effects of card content on search interactions. With respect to ***RQ1***, our results suggest that the presence of entity cards on the SERP affects searchers' behaviour. In particular, we find that when an entity card is displayed on the results page, fewer clicks are issued on web results, fewer documents are marked as relevant, and searchers scroll to a shallower depth on the result page, irrespective of card content, than when the card is entirely absent. In the case of off topic entity cards, this can be attributed to the *assimilation effect* discussed in section 5.1: searchers perceive the entire page as less reliable and overall engage with results less. Our results also show that, when an off topic card is displayed, searchers typically take longer to issue their first click on the results page, but overall take just as long as when an on topic card is displayed (or the card is absent) to complete their task. This might suggest that they take longer to process the off topic card, which is dissonant to their assigned search topic, and therefore use less time to explore the results ranking.

When the card is on topic, our results reveal that, on average, participants hover over web results more, yet click on fewer results and mark fewer documents as relevant than in the absent condition. This might suggest that participants inspect web results more attentively when an on topic entity card is displayed, a finding that supports previous work on aggregated search (Arguello and Capra, 2016), where the presence of an on topic aggregated search block lead to higher engagement, overall, with the results page.

Our results also indicate that off topic cards tend to increase almost all dimensions of perceived task workload, as assessed by searchers directly. In the *temporal* and *effort* dimensions, off topic cards increase workload compared to their on topic counterparts also, which is not surprising given that searchers take longer to issue a first click on web results when the card is off topic, which in turns leads to less perceived time available for completing their primary task, which is finding and marking relevant web documents (even though our tasks were not time limited). This finding is also interesting because searchers appear less engaged in their task (i.e., fewer clicks, lower scroll depth), on average, when the card is off topic, but due to increased effort in processing dissonant elements of the results page, they perceive higher levels of workload across dimensions. Our results also suggest that on topic entity cards do not seem to increase workload, on average, compared to the absence of cards, even though they require searchers to process additional information on the results page which is not directly related to successful task completion (i.e., selecting relevant results).

Our estimates of effect sizes suggest that card content manipulations have a weak effect on search behaviour, but a much stronger effect on perceived workload. This is due to higher variance in searcher behaviour measures, possibly caused by larger differences between participants in how they approached our tasks, than in self-assessed workload measurements (as indicated by standard deviation values in table 5.5), which were limited by design to a 20 point scale. Indeed, in approximately 10% of our tasks, searchers inspected all 50 web results shown on the page, which is perhaps a reason for the higher variance in user search behaviour we observed. Even so, our analysis suggests that differences in measures of search behaviour across card content manipulations are not due to chance, even though effect sizes are relatively small. Effects are, however, much stronger regarding perceived demand, with off topic cards increasing several dimensions of workload, over on topic or absent cards.

Finally, with respect to ***RQ1***, we conclude that card content has an effect on both search behaviour and perceived workload that is unlikely due to chance. This effect appears to be stronger in the case of off topic cards and stronger with respect to measures of workload, rather than search interactions. Our findings overall suggest that placing off topic entity cards on results pages can be detrimental to user search experience, on average. In addition, our results have broader implications for modelling and evaluating modern web search systems.

Card coherence. The results of our experimental manipulations of card coherence suggest that non-coherent cards generate effects similar to those generated by off topic cards, with respect to both search behaviour and workload. Unlike off topic cards, non-coherent cards do not appear to increase the duration of time before a click is issued on the results page, on average, suggesting that they are easier to process than completely off topic cards. Also unlike off topic cards, non-coherent cards have overall weaker effects on measures of workload, and, overall, tend to increase searchers' perceived task effort in fewer dimensions of workload.

We attribute the weaker effect of card coherence on search behaviour and workload to the more subtle manipulation of card display features (i.e., card coherence is, perhaps, not immediately obvious to searchers) and also to the fact that users were assigned clearly formulated information seeking tasks, and therefore were able to extract useful information from entity cards even when they displayed non-coherent information.

Entity card coherence can be considered a spectrum, with on topic and off topic cards being opposite ends of this spectrum, and non-coherent cards somewhere in-between. This view of card coherence is supported by our results, which show similar effects for non-coherent and off topic cards, but overall weaker effects (and in fewer dimensions) in the case of non-coherent cards. We conclude that card coherence, as put into practice in our study, has an effect on both search behaviour and workload that is unlikely due to chance, however this effect is smaller than in the case of card content or card diversity.

Card diversity. Our analysis suggests that non-diverse cards are similar to both off topic and non-coherent cards. In particular, on topic but non-diverse cards lead to fewer clicks or mouse hovers on web results, on average, than diverse cards. Unlike off topic or non-coherent cards, non-diverse cards tend to lead, on average, to a higher number of results marked as relevant than their diverse counterparts, even though fewer results seem to be inspected by searchers. This might suggest that non-diverse cards provide assistance to searchers in identifying relevant content on the results page, perhaps by making their assigned search topic more salient on the page (i.e., an interface element reminding them what their assigned task topic is).

This is supported by our findings regarding perceived workload. Our results suggest that non-diverse cards increase overall workload, on average, by approximately 1.5 points (on a 20 point scale), compared to both diverse and absent conditions. This might be due to our experimental setting, with non-diverse cards being perceived as another element of the experimental page that instructs searchers what information to look for, rather than an informative element in itself. Even so, our findings supports the claim that removing vertically diverse results from entity cards can be detrimental to user experience, as these results potentially help searchers process entity card information with less perceived effort. Overall, our findings suggests that heterogeneous content displayed within entity cards is both informational and diverting from task demand.

It is also interesting to note that entity cards lead to shallower scroll depth, on average, irrespective of card content, coherence or diversity. This might suggests that entity cards act as a visual anchor at the top of the page, independent of card information or visual saliency.

Our findings have implications for the designer of web search interfaces, and suggest that displaying more diverse content within entity cards, when query intent prediction is accurate, is to be preferred. In broader terms, our findings have implications for modelling and evaluating modern web search systems.

5.6 Chapter Summary

Modern search engines have started integrating entity cards on the results page in response to users' queries. Entity cards enhance search experience in several ways, by helping users navigate a diverse information space and by providing a summary of relevant and diverse content directly on the results page.

In this chapter, we report on a large-scale crowd-sourced user study, with more than 500 unique searchers, investigating the effect entity cards integrated into result pages have on searcher behaviour and perceived workload. Our results show that the presence of an entity card on the results page influences searcher behaviour with respect to both engagement with web results, as well as perceived task demand. Our analysis of card properties suggests that card *diversity* is more influential on search behaviour and workload than card *coherence*. Our findings have important practical implications for modern web search, in particular with respect to user modelling and the evaluation of non-linear result pages.

Within our *Composite Web Search* framework, entity cards represent instances of composite objects and therefore, our study of entity card influence on search behaviour is run under the assumption that our findings generalise to other potential types of composite objects that are integrated in the results page in a similar style (i.e., as contextual elements, clearly delimited from the general web results ranking through display properties, and not a simple ranking of items). Our assertions that composite object usefulness is constrained by ***(a)*** the relevance of a document (or group of documents) that play a central role within the object and ***(b)*** a complex interplay of object properties are supported by the experimental evidence presented in this chapter. With respect to ***(a)***, our results show that the effect of manipulating composite object *coherence*, by explicitly deteriorating the quality of results contained within the object, on search behaviour and perceived workload, is more limited, as long as the central component of the object (in the case of entity cards, the entity title and summary) is on-topic (i.e., relevant to the user's query). With respect to ***(b)***, our results show that composite object properties — relevance (or content) and diversity — influence user behaviour and perceived workload in subtle but distinct ways.

The following chapters of this thesis focus on algorithmic aspects of result composition. How to represent heterogeneous documents within a unified features space in order to assess their similarity, how to operationalise and assess object properties and how to construct composite objects are the topics we address next. We begin by exploring the application of a general *composite retrieval* (Amer-Yahia et al., 2014) framework in a heterogeneous web search environment with the goal of constructing *relevant*, *coherent* and *diverse* composite objects in the following chapter.

Chapter 6

Composite Retrieval of Heterogeneous Web Search

Current web search systems generally present a ranked list of documents in response to user queries. In aggregated search systems, results from different and increasingly diverse sources of information (e.g., *image*, *video* or *news* verticals) are returned to users — for instance, many current web search engines return to users both images and web documents in response to the query *"cat"*, merged within a unified ranking of results.

In this chapter, we present an experiment inspired by previous work on *composite retrieval* (Amer-Yahia et al., 2013, 2014) which investigates the algorithmic assembly of complex information objects — what we define as *composite objects* — containing results from multiple heterogeneous sources. In our experiments, rather than merging blocks of results from different verticals within a single ranking, as is the case with aggregated search, we propose to return to users a set of *composite objects*, where each object contains results from several verticals, assembled around a common topic. For example, for the query *"London Olympics"*, one composite object per sport could be returned, each containing results extracted from the *news*, *videos*, *images*, or *Wikipedia* verticals.

In our experiment, we propose and evaluate a variety of approaches to constructing *relevant*, *coherent* and *diverse* composite objects. How these properties of composite objects are operationalised is a core aspect of the work presented here. Compared with three baselines (general web only ranking, federated search ranking and aggregated search), our evaluation demonstrates performance improvement for a highly heterogeneous web collection. Research presented in this chapter is based on previously published work available in Bota et al. (2014).

6.1 Introduction

Consider the following user information need *"finding all information to plan a trip to Korea"*. Answering this information need typically involves submitting several queries to gather information about airports and visa policies, to read online reviews about hotels, and to check the geographic proximity of places to visit. Current search engines aggregate results from multiple verticals. However, the presentation of search results is limited to blocks where each block contains homogeneous information of one type.

As the web has made available a large variety of diverse search engines — or *verticals* — it is becoming important to return to users organised answers, made of information extracted from heterogeneous data sources. Doing so will not only support users in complex search tasks, but also allow them to understand the diversity of the information space and access relevant and diverse information in a structured way, directly on the results page.

For example, users typing the query *"olympics"* during the London 2012 Olympic Games may have been interested in different on-going game results with detailed statistics, video summaries and players' post-match commentary quotes. Returning such results per sport could be helpful to users, allowing them to explore all the information regarding each individual sport or focus on specific sports, events or athletes. We propose to return to users a results page — what we refer to as a *composite page* — where results from diverse verticals are assembled into *composite objects*. Thus a composite page is a set of objects, where an object is composed of *"coherent"* information extracted from various sources (e.g., videos, statistics and quotes). What *coherent* means and how it is operationalised is addressed in the following sections of this chapter.

Previous work by Amer-Yahia et al. (2014) defined the task of searching for complementary items in a collection as *composite retrieval*, and proposed organising results into *bundles*, in an effort to provide an improved exploratory experience over the typical ranked list of results. In their work, composite retrieval is explored in the context of a homogeneous information space. However, modern web search involves searching and merging results across verticals of heterogeneous rather than homogeneous items.

To construct a high-quality composite page, several criteria should be satisfied: ***composite object relevance***, where the items within objects should be topically relevant to the query; ***composite object coherence***, where the items within the object should be similar with each other and therefore coherent to the topic; ***composite object diversity***, where the items within the object should cover a diverse set of results from various verticals; and ***page diversity***, where the objects within the page should cover various aspects/topics of the query. As in traditional search, *relevance* is a priority as this is the key factor to a satisfactory user experience; promoting cohesion and diversity is important but only when the results are relevant.

In this experiment, we study *composite retrieval* in a *heterogeneous* web search environment. We propose several approaches to constructing composite objects. Two challenges arise due to the heterogeneous nature of the data. First, relevance score distributions are not comparable across verticals, making relevance estimation of composite objects more complex. Second, different factors can sometimes be contradictory with each other (relevance vs. diversity), thus determining an appropriate trade-off is required.

We build on an existing composite retrieval framework by Amer-Yahia et al. (2014), adapting it to a heterogeneous web search context. In addition, we put forward a new approach for result composition, which we refer to as the *central-satellite* result composition paradigm, where a retrieved object is defined as a central package (e.g., a set of general web documents) and a set of satellite packages consisting of items retrieved from verticals that are coherent to the central package. The study presented here addresses the following research questions:

(RQ1) Can we construct composite pages that are more relevant than existing solutions, such as "general web search only" ranking, aggregated search and federated search ranking?

(RQ2) When constructing composite objects, can we utilise query-related entities as an anchor to bridge the semantic gap between items retrieved from heterogeneous sources?

(RQ3) As composite objects are created by selecting items from traditionally ranked lists of documents, how robust are different result composition approaches to the quality of the initial ranking of documents?

The contributions of our study are the following: we propose a novel approach to result composition and demonstrate its effectiveness. We conduct extensive experiments comparing our proposal with a number of baselines and approaches. To our knowledge, this is the first endeavour to use entities to bridge the semantic gap between items retrieved from heterogeneous web resources. Moreover, we investigate and provided insights on the usefulness of different sources of evidence for result composition.

The rest of this chapter is organised as follows. Section 6.2 discusses previous work regarding composite retrieval. We describe our result composition framework in section 6.3 and propose a variety of approaches to forming composite objects in section 6.4. Details of the test collections and experimental setup are provided in section 6.5. Section 6.6 reports our experimental results, and we discuss the implications of our results and conclude in section 6.7.

6.2 Prior Work

Web search has become central to technology-mediated information interaction. Because of its central role in accessing and creating information, many aspects of web search have been studies extensively over the past several decades, from retrieval algorithms to user modelling and search user interfaces. The work presented in this chapter brings together two broad areas of study related to web search: ***(i)*** cluster-based retrieval and ***(ii)*** heterogeneous information access on the web.

Cluster-based retrieval algorithms. The cluster hypothesis — closely associated documents tend to be relevant to the same requests (van Rijsbergen, 1979) — gave rise to a large body of work (Kurland and Domshlak, 2008; Tombros et al., 2002) on using query-specific document clusters for improving retrieval effectiveness. Various approaches to clustering results have been explored in prior work (e.g., (Kurland and Domshlak, 2008; Tombros et al., 2002)), and discussed in more detail in sections 2.2.4 and 3.2.2.

Our work is similar to cluster-based retrieval as we form composite objects based on a cluster-inspired optimisation approach (selecting items that are similar to each other to form a composite object). Similar to cluster-based retrieval, we rank the verticals (i.e., each vertical can be considered a cluster of documents) based on their estimated relevance and ultimately select the top ranked verticals to choose items from. The heterogeneous nature of the data and our approach to constructing links between heterogeneous documents are what differentiate our work from traditional cluster-based retrieval.

Composite retrieval. Composite retrieval has been studied in recent years, however, the applications so far are in structured or semi-structured scenarios — such as recommending products or finding a restaurant. Amer-Yahia et al. (2014) studied the complexity of the problem of composing *"bundles"* — what we define as *composite objects* — with constraints, such as budget or item compatibility. Furthermore, they formally prove that the task of composite retrieval is $NP - hard$, and propose a greedy approximation approach to address the complexity of the problem by first producing and then selecting an approximately optimal set of composite objects (i.e., *"bundles"*). We discuss previous approaches to composite retrieval in more detail in section 2.2.4.

In this chapter, we study composite retrieval in the context of heterogeneous web search and provide solutions to tackle the challenges arising from the heterogeneous nature of the data. In addition, as relevance is highly correlated to a satisfactory user experience in search (Sanderson et al., 2010), different from other works, we treat this as our main criteria of optimisation, whereas criteria such as cohesion and diversity are considered secondary. We explore the framework developed in Amer-Yahia et al. (2014) and adapt it to suit the needs of heterogeneous scenarios. Specifically, we enhance the definition of composite objects by incorporating the concept of relevance explicitly. Secondly, we incorporate diversity and relevance at the time of ranking (choosing) composite objects, thus leading to effective solutions. Thirdly, we exploit the use of entities in linking relevant items across verticals, and we also incorporate query intent into the formation of composite objects. Extensive experiments on the TREC federated web track data set demonstrate the effectiveness of heterogeneous composite retrieval on the web.

Aggregated search. Aggregated search is the task of retrieving results from a variety of resources (i.e., verticals) and merging results within a unified page — in sections 2.2.3.2 and 3.2.3 we review prior work on aggregated search in more detail. In aggregated search, the most common presentation strategy is to group results into a ranked list of so-called blocks where each block contains homogeneous information of one type (i.e., retrieved from one relevant vertical). Similar to aggregated search, selecting and organising results from heterogeneous sources is the main focus of result composition. However, rather than presenting the results of each selected vertical as a homogeneous block, result composition aims to return results into coherent objects, where each objects contains heterogeneous items, retrieved from different verticals.

Diversity in information retrieval. Information retrieval research has investigated *"diversity-based"* or *"subtopic"* retrieval approaches for modelling user search intents during search tasks for ambiguous or multi-faceted information needs (Clarke et al., 2008; Santos, 2012). An intent-diversified result ranking can be created by interleaving documents sampled from possible search intents (subtopics), with the importance of each intent indicated by several features such as prior search intent click-through rate or original document relevance. Our result composition approach also takes *"subtopic"* diversity into account when forming result pages containing a diverse set of composite objects. However, rather than forming a homogeneous ranking list covering various subtopics, we construct a page consisting of composite objects where each object corresponds to a *"subtopic"* of the user's query.

6.3 Result Composition Framework

We formally define the result composition problem below, followed by a review of the associated challenges.

6.3.1 Problem Formulation

We propose a framework for heterogeneous web result composition which is similar to previous work by Amer-Yahia et al. (2014) on *composite retrieval*. The latter has been mostly studied in structured or semi-structured environments, and assumes item relevance as a given property of items. Hence, the application of their framework to heterogeneous web search is not straightforward. In addition, we consider *relevance* to be a crucial component of user experience in web search and therefore incorporate relevance estimates into the optimisation of our objective functions. The heterogeneous nature of the multi-vertical environment also requires novel ways to estimate the various components of the framework.

The goal of result composition in a heterogeneous environment is to assemble a set of composite objects $P = \{S_1, ..., S_k\}$ to form a composite page P, where an object $S_i \in 2^I$ is a set of items that originate from a subset of verticals $V = \{V_1, ..., V_n\}$. The objective of the optimisation is to find a set of composite objects to form a page $P = \{S_1, ..., S_k\}$ that maximises the utility $util(P)$. We assume that the utility of the page $util(P)$ solely depends on the following four criteria:

- **Relevance**: the expected probability of items in composite objects to be relevant to a searcher's query:

$$rel(P) = \frac{\sum_{1 \leqslant i \leqslant k} \sum_{u \in S_i} r(u|q)}{\sum_{1 \leqslant i \leqslant k} \sum_{u \in S_i}}$$

  where $r(u|q)$ is the probability that a user finds an item u relevant as a function of the editorial grade g_u of that item u. $r(u|q)$ can be chosen in different ways. Similar to the common gain function for DCG, we define it as:

$$r(u|q) = \frac{2^{g_u} - 1}{2^{g_{max}}}, \;\; g \in \{0, ...g_{max}\}$$

when the item is non-relevant ($g = 0$), the probability that a user finds it relevant is 0, whereas when the item is highly relevant ($g = 4$, if a 5 point scale is used), the probability of relevance is near 1. When a binary relevance grade is used, $rel(P)$ corresponds to precision-at-a-cut-off metric $P@n$ of the page P.

- **Topical Coherence**: the expected accumulated similarity of the items within the composite object:

  $$tcoh(P) = \frac{\sum_{1 \leqslant i \leqslant k} \sum_{u,v \in S_i} s(u,v)}{\sum_{1 \leqslant i \leqslant k} \sum_{u,v \in S_i}}$$

  The similarity $s(u,v)$ between an item pair (u,v) can be computed implicitly by a given representation of the items. This $tcoh(P)$ corresponds to the normalised expected coherence of composite objects.

- **Topical Diversity**: the expected inter-object separation for composite object pairs:

  $$tdiv(P) = \frac{\sum_{1 \leqslant i < j \leqslant k} (1 - \max_{u \in S_i, v \in S_j} s(u,v))}{\sum_{1 \leqslant i < j \leqslant k}}$$

  where the inter-object separation is defined as the minimum distance between two items from separate composite objects.

- **Vertical Diversity**: the expected number of verticals the relevant items belong to in the composite object:

  $$vdiv(P) = I\text{-}rec(P)$$

  where I-$rec(P)$ is the intent (vertical) recall metric (Zhai et al., 2015) for page P. Basically, it calculates the recall of those verticals that return relevant items on the composite page.

A page P with high $util(P)$ should consist of a set of *topically diverse* composite objects that each contain *relevant* items originating from a set of *diverse verticals* and are *coherent* to one aspect of the query. Since this is a novel search task, the importance of each factor to user experience has not been well understood. Indeed, based on our study of users' perspective on result composition, discussed in chapter 4, determining a hierarchy or a weighting of object properties based on their importance to users is not obvious. Therefore, we evaluate the performance of each factor separately. However, as relevance is key to the user experience in search (Sanderson et al., 2010) we use it as our main criteria of interest for the purpose of evaluation.

If query subtopics and relevance assessments corresponding to individual subtopics are available, as in collections used to evaluate diversity based information retrieval (Clarke et al., 2008), *topical relevance* and *topical coherence* can be evaluated using existing diversity metrics (e.g., intent-aware metrics (Agrawal et al., 2009)). However, we rely on a collection which has multiple verticals (i.e., a federated search collection) for which subtopic relevance assessments are not available. Considering the high cost involved in collecting subtopic relevance assessments, we evaluate those two criteria (*topical relevance* and *topical coherence*) using cluster quality evaluation metrics (i.e., cluster coherence and separation) as described above.

6.3.2 Challenges

Result composition is particularly challenging in the context of a heterogeneous web environment for several reasons: ***(i)*** computational complexity, ***(ii)*** term mismatch with heterogeneous information (i.e., cross-vertical vocabulary gap) and ***(iii)*** the appropriate estimation and optimisation of the multiple criteria used in our optimisation function (i.e., coherence, diversity and relevance).

For effective user experience, we aim to create optimal composite objects, ones that meet all our criteria. However, this in itself is an *NP-hard* optimisation problem. More precisely, it has been proven that optimising for *coherence* and *diversity* in the objective function of the composite retrieval approach proposed by Amer-Yahia et al. (2014) can be reduced from the well known $NP - hard$ problem of *Maximum Edge Subgraph*. Therefore, as our approach is parallel to that of composite retrieval, we require efficient greedy algorithms to optimise the utility of a composite results page.

Term distributions in different verticals vary widely (Santos et al., 2011) and are therefore not comparable. This makes it difficult to calculate the similarity $s(u, v)$ between two items u and v that originated from two different verticals. Therefore, it is important to bridge the gap between term distributions across diverse verticals in a way that enables item comparison.

Different sources of evidence (e.g., the query-item similarity, the vertical where the item originated from, etc.) have to be considered when estimating the relevance of an item and also when deciding whether to include it in a composite object, and in which object, or not. A comprehensive study is required to understand the optimal way to estimate the relevance of an item in heterogeneous result composition. In addition, how to appropriately combine these sources of evidence to account for coherence and diversity is not obvious.

Our contributions lie in addressing these three challenges: ***(i)*** we propose a new approach for composite object formation and experiment with a number of variations of the produce and choose approach from Amer-Yahia et al. (2014); ***(ii)*** we investigate entity based document representation as a solution for bridging the cross-vertical vocabulary gap; and ***(iii)*** we propose different approaches to incorporating relevance into the objective function, and analyse the usefulness of various features in estimating relevance.

6.4 Composite Object Selection & Ranking

We introduce our adapted greedy approach for optimisation and describe the methodology used for estimating different sources of evidence in this section.

As discussed above, the problem of result composition while optimising for the various criteria of interest (i.e., relevance, coherence and diversity) is NP-hard. Previous work (Amer-Yahia et al., 2013, 2014) showed that *Maximum Edge Subgraph* and *composite retrieval* are two counterparts. If we generate candidate composite objects and we consider each candidate object as a node of an object-graph, where inter-object distances are the edge weights, then the result composition problem can be reduced from the *Maximum Edge Subgraph* problem. This suggests that result composition can be approximated by generating a set of candidate objects and then selecting the best possible subset. Amer-Yahia et al. (2013, 2014) call this approach Produce-and-Choose (PAC) and we choose this paradigm as the basis for our investigation of result composition.

The PAC approach discussed in Amer-Yahia et al. (2013) consists of two stages: ***(i)*** produce composite objects and ***(ii)*** choose composite objects. In their work, they explore different ways of generating composite objects: *Bundles One-by-One (BOBO)* and *Constrained Clustering*. We employ just one of these approaches, BOBO (as discussed further in section 6.4.1.1), as it is representative of a wide class of clustering algorithms. However, the application of this framework in a heterogeneous environment is not trivial. We propose to address this problem by using an entity based representation for heterogeneous documents. This approach also allows us to diversify the results based on captured query intents that are represented by entities (discussed in section 6.4.2.1). In summary, unlike Amer-Yahia et al. (2014), at the production stage we compute coherence and vertical diversity for producing good composite objects and at the choose stage (ranking), we integrate topical diversity and relevance. In addition, based on the unique characteristics of the multi-vertical environment, we proposed a novel approach (Central-Plus-Satellite, CPS) that better suits a heterogeneous web environment, such as the one we are exploring in this study. For choosing composite object (discussed in section 6.4.2), previous research (Amer-Yahia et al., 2013, 2014) showed that the known results for *Maximum Edge Subgraph* can be exploited to preserve the approximation guarantees. We employ their approach and incorporate relevance into it.

In summary, unlike Amer-Yahia et al. (2014), we apply the composite retrieval framework in a heterogeneous environment, and tackle the problems that arise in multi-vertical environment by proposing a novel, entity based approach to assessing result similarity and relevance.

6.4.1 Produce Composite Objects

We introduce two approaches for producing composite objects: BOBO (adapted from previous approach, as discussed above) and CPS (newly proposed).

6.4.1.1 Objects One-by-One (BOBO)

This method of producing a set of candidate objects is inspired by *k-nn* clustering: a pivot is chosen at each step and a valid object is built around that pivot, by selecting its nearest neighbours in a unified feature space. If the object's internal coherence score is above a certain threshold μ, it is kept, otherwise it is discarded. The pseudo-code for this algorithm is shown in Algorithm 1.

The BOBO approach starts with an empty set of candidate composite objects, and considers each element in the item set as a possible pivot. The item set originates from the initial federated search rankings that merge results from all verticals. At each iteration an item is picked from the set of *Pivots*, and in our case, we choose the item (i.e., result) in *Pivots* with the highest relevance estimation. Once a pivot is selected, we build a composite object S around it. This is done by the routine *pick_object* described in Algorithm 2. The routine greedily picks the closest element to the pivot ω, as long as the composite object s does not exceed the pre-defined maximum number of items for an object. The function f in Algorithm 2 also ensures vertical diversity, by enforcing the constraints that the composite object s is required to contain items from at least two different verticals.

Algorithm 1: ***Produce Composite Objects (BOBO)***

```
Input: set of items $I$, a cost function $f$ that checks vertical diversity and
composite object size constraints, a threshold $\beta$ on the number of items in
a composite object, minimum composite object score $\mu$, number of
composite objects $c$
Output: a set of $c$ valid composite objects
$Cand \leftarrow \varnothing$
$Pivots \leftarrow I$
while $Pivots \neq \varnothing$ and $|Cand| < c$ do
    $\omega \leftarrow Pivots[0]$
    $I \leftarrow I \setminus \omega$
    $S \leftarrow pick\_object(\omega, I, f, \beta)$
    if $score(S) \geqslant \mu$ then
        $I \leftarrow I \setminus S$
        $Pivots \leftarrow Pivots \setminus S$
        $Cand \leftarrow Cand \cup S$
    else
        $Pivots \leftarrow Pivots \setminus \omega$
return $Cand$
```

Once a candidate object is created (by *pick_object*), the algorithm checks whether its internal coherence (*score* function in Algorithm 1) is larger than a pre-defined threshold μ. More precisely, to reflect coherence, the *score* function is defined as the expected similarity of item pairs $score(S) = \sum_{u,v \in S} s(u,v) / \sum_{u,v \in S}$. If this check has passed, then the object enters the candidate set *Cand* and the items within this object are removed from I and *Pivots* so that they can no longer be selected again for other candidate objects. Otherwise if the object S has a score lower than μ then it is discarded. In both cases the pivot ω is removed from *Pivots* so that it is no longer considered.

6.4.1.2 Central-Plus-Satellite (CPS)

We introduce a different method, suggested by the observation that established vertical selection methods are prone to noise interference, and inspired by the approach to composite retrieval described in Basu Roy et al. (2010). The basic idea is that we first produce composite objects in a *central vertical* using BOBO and then attach items from other *satellite verticals* onto the produced objects. Our approach combines established vertical selection method (*ReDDE*, discussed in Si and Callan (2003)) with our entity based representation.

The pseudo-code for the CPS algorithm is shown in Algorithm 3. In the initial

Algorithm 2: pick_object

```
Input: pivot $\omega$, set of items $I$, a cost function $f$ that checks vertical
diversity and composite object size constraints, a threshold $\beta$ on the
number of items in a composite object
Output: composite object $s$
s = $\{\omega\}$; active ← $I \setminus \omega$; $finish$ = false
while not $finish$ do
    $i \leftarrow argmax_{\{i \in active\}} s(i, \omega)$
    if $f(s \cup \{i\})$ then
        $s \leftarrow s + i;$
    else
        $finish = true$
    $active \leftarrow active \setminus \{i\}$
return $s$
```

phase, vertical selection is performed using the ReDDE methodology to select the *central vertical* and *satellite verticals*. The top-1 vertical in the ReDDE vertical ranking is treated as the *central vertical*. A set of items from this central vertical forms the central item set I. Similarly, a set of *satellite verticals* is created (top-n verticals except the *central vertical*) and a set of satellite items coming from those verticals form the satellite item set S. In our experiments, we fix the central vertical to *"General Web"* as we find this to minimise noise introduced by vertical selection, and set a threshold of $n = 2$ on the number of satellite verticals selected.

In the second phase, BOBO is used to generate and select a set *Cand* of candidate objects, using only items that originate from the central vertical. This provides us with a set of coherent composite objects to which we can attach items from the satellite verticals.

In the third phase, satellite items are attached to the objects in *Cand* only if a set of constraints (e.g., coherence) are satisfied. The object-item similarity $s(b, i)$ is based on item-item similarity $s(u, i)$. We simply assume that a composite object b is represented by the elements common to all items within that composite object. In our experiments, after the object-item similarity estimation, we only add items to an object if the items contain at least a certain threshold (i.e., 30%) of entities from that composite object.

6.4.2 Choose Objects

Our approach of choosing composite objects is different from the PAC approach discussed in Amer-Yahia et al. (2013), to the extent that we aim to select the ob-

Algorithm 3: Produce Objects: CPS

```
Input: A central item set $I$, a set $S$ of satellite items, a cost function $f$ that checks diversity, coherence and object size constraints, a threshold $\beta$ on the number of items in an object, minimum object score $\mu$, number of objects $c$, the number of objects to select $k$
Output: A set $s$ of valid objects
$Cand \leftarrow BOBO(I, \alpha, f, \beta, \mu, c)$
$Cand \leftarrow ChooseObjects(k, Cand)$
for $b$ in $Cand$ do
    $i \leftarrow argmax_{\{i \in S\}} s(b, i)$
    while $f(b \cup \{i\})$ do
        $b \leftarrow b \cup \{i\}$
        $S \leftarrow S \setminus i$
        $i \leftarrow argmax_{\{i \in S\}} s(b, i)$
return $Cand$
```

Algorithm 4: Choose Objects

```
Input: number of composite objects $k$, a set of candidate objects $Cand$, $\omega(S \in Cand) = \sum_{u,v \in S} s(u, v)$
Output: a set $\Omega$ of valid composite objects
$\Omega \leftarrow \emptyset$
while $|\Omega| < k$ do
    $u \leftarrow argmax_{\{v \in Cand\}} \omega(v)$
    $Cand \leftarrow Cand \setminus u$
    $\Omega \leftarrow \Omega \cup u$
return $\Omega$
```

jects that have the highest degree of relevance and coherence. Given the number of required objects, k, a set *Cand* of candidate objects, and a similarity function between items, Algorithm 4 selects the top k most cohesive composite objects in the candidate set (determined by the function w). Relevance is considered using different relevance estimation approaches (as discussed in section 6.4.3.2).

Since we aim to return objects that are not only relevant and coherent, but also topically diverse, we apply a post-diversification (section 6.4.2.1) on the objects we choose at this stage. We do not directly consider topical diversity when generating composite objects since doing so degrades the relevance of the objects.

6.4.2.1 Post-Diversification

We propose two different diversification strategies to the set of composite objects we select, based on how we estimate topical distance between objects.

DT Diversification The first diversification strategy we consider is similar to Maximal Marginal Relevance (MMR) ranking strategy (Carbonell and Goldstein, 1998). We denote this approach as *DT Diversity* and a suffix of DT is attached to the approach employed by this strategy (e.g., BOBO-DT). This approach is based on applying MMR diversification to objects (i.e., rather than documents) and the methodology to determine distance d between two composite objects S_i, S_j is defined as follows:

$$d(S_i, S_j) = 1 - argmax_{u \in S_i, v \in S_j} s(u, v)$$

The distance is computed as the maximum similarity between any two items in the two objects. At each step, we select the object that is most cohesive and relevant, but at the same time the most dissimilar from the previously selected object.

DE Diversification We propose another diversification strategy that is based on explicitly diversifying query intents using entities (denoted as *DE Diversity*). The basic idea is that we estimate different subtopics (intents) of the query q by a set of query-specific entities q_e. We consider each entity $e \in q_e$ as being a subtopic of q, and we compute the probability of *aboutness* of an entity e to a document by simply using the frequency of the entities $freq(e)$ appearing in document d:

$$P(e|d_e) = \frac{freq(e)}{\sum_{e \in d_e} freq(e)}$$

After the estimation of entity-document *aboutness* as above, for every object $S \in Cand$ and every entity $e \in q_e$, we define an object-entity *aboutness* score that is calculated as an average of the entity-document *aboutness* score of all the documents in the object:

$$aboutness(S, e) = \frac{\sum_{d_e \in S, e \in q_e} P(e|d_e)}{\sum_{d \in S}}$$

Therefore, to diversify, for each entity e (treated as a subtopic or intent), we select the object that has the highest *aboutness* score to e and we assume that the object corresponds to the subtopic e of query q.

6.4.3 Evidence and Estimation

We now describe our approach for estimating similarity between items and estimating item relevance.

6.4.3.1 Estimating Similarity

Before attempting to produce coherent composite objects, we must define a measure of similarity between items (i.e., results or documents). There are two challenges when estimating the similarity of documents within a multi-vertical environment. Firstly, the documents are heterogeneous (general web documents vs. multimedia documents) and the different term distributions across verticals make the estimation difficult. Secondly, the retrieval ranking functions of documents from multiple verticals can vary. Therefore, we propose to use named entities as a bridge between verticals in assessing document similarity. We first assume that Wikipedia entries are representative of all the entities present in the documents. Then we used a state-of-the-art annotation tool that maps textual spots (i.e., terms) to Wikipedia entries (Ferragina and Scaiella, 2010) on our documents, such that every document d in our collection has a corresponding entity representation d_e. To further select the most representative entities from a document, we sort the entities in d_e using a traditional $tf \times idf$ measurement (entity frequency in the document multiplied by the inverse of its frequency across the collection), and select the top 100 entities for every document; we represent each document d by a 100-dimensional entity vector $d_e = \{e_1, e_2, ..., e_{100}\}$. Finally, we use the Jaccard coefficient of their entity sets to compute the similarity $s(u, v)$ of two documents u, v.

$$s(u, v) = \frac{|u \cap v|}{|u \cup v|}$$

As mentioned previously, we estimate the object-item similarity $s(b, i)$ by assuming that an object b is represented by the entities that appear in all the items (i.e., results) contained within that object. We then also use the Jaccard coefficient between entity sets of the object b and the item i to compute the similarity $s(b, i)$.

6.4.3.2 Estimating Relevance

To estimate the relevance of a document to a given query based on its entity representations, we annotate queries with entities by using a pseudo-relevance feedback technique. For a given query q, from the highest top 10 ranked documents in a BM25 ranking, from each document we extract and sort entities as previously described. From the set of entities extracted from these documents, we select the top 10 most frequently occurring entities as the entity representation of the query $q_e = \{e_1, e_2, ..., e_{10}\}$.

We describe the methods used to estimate document relevance further. We have three sources of evidence: ***V***: the estimated query-vertical similarity based on *ReDDE* resource selection approach; ***T***: the estimated query-document similarity based on term-based BM25 ranking; and ***E***: the estimated query-document similarity based on entity representations of the queries and documents.

For ***V***, we compute a probability $P(v|q)$ of a query's (q) orientation to vertical (v) using the *ReDDE* (Si and Callan, 2003) approach. *ReDDE* scores a target vertical based on its expected number of documents relevant to the query. It derives this expectation from a ranking using a central index that combines documents sampled from every target vertical. Given this ranking, *ReDDE* accumulates a vertical score $ReDDE_q(v_i)$ from its document scores $P(q|\theta_d)$, taking into account the difference between the size of the original vertical N^{v_i} and a sample size N^{samp}.

$$ReDDE_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d \in topm} I(d \in v_i) P(q|\theta_d)$$

where $I(.)$ is a indicator function. To be consistent with Si and Callan (2003), we choose $m = 1000$ in our experiments. For ***T***, we compute $P(d|q)$ as follow:

$$P(d|q) = \frac{bm25(d,q)}{\sum_d bm25(d,q)}$$

where the $bm25(d,q)$ is the BM25 scoring function. For ***E***, we compute the similarity of a document to a given query based on the entity representation $P(d_e|q_e)$:

$$P(d_e|q_e) = \frac{\sum_{e \in q_e, e \in d_e} P(e|q_e) \cdot P(d_e|e)}{\sum_{e \in d_e, e \in q_e}}$$

where $P(e|q_e)$ is estimated as the probability of generating entity e from the entity representations of the top-10 BM25 retrieved documents to the initial query q.

We can incorporate any of these three relevance estimates into our objective function (i.e., $w(v)$ in Algorithm 4) when choosing objects by simply including them as a factor in the initial objective function $(w(v))$ and we study their effectiveness on choosing objects. We add a prefix of a given relevance estimation method to a composition approach identifier when reporting our results if we incorporate it into the choose objects stage. For example, BOBO-VT means the

Vertical	*Description*	*Example websites crawled*
Image	Online images	Photobucket
Video	Online videos	Hulu, YouTube
Jobs	Job description pages	LinkedIn Jobs, Simply Hired
News	News articles	Google News, ESPN
Blog	Blog articles	Google Blogs, WordPress
Q&A	Answers to questions	Yahoo Answers, Answers.com
Shopping	Product shopping page	Amazon, eBay
Academic	Research papers or reports	Nature, CiteSeerX, SpringerLink
Encyclopedia	Encyclopedic entries	Wikipedia
Books	Book search pages	Google Books
Social	Social interaction services	Facebook, Tumblr, Twitter
General web	Standard web pages	Google, Yahoo, AOL, Bing

Table 6.1: Verticals used to assemble the federated search test collection used in our study, as described in Nguyen et al. (2012). The collection contains results from 108 real Web search engines, of varying sizes, covering a diverse set of media types and domains.

BOBO approach that incorporates both vertical orientation relevance estimation $\boldsymbol{V}$ and term-based relevance estimation $\boldsymbol{T}$, whereas BOBO means the original approach without incorporating relevance estimations in any way.

6.5 Experimental Setup

In this section, we describe the test collection, as well as the evaluation metrics and baseline systems used in our work.

6.5.1 Data

We used a federated search test collection (Nguyen et al., 2012) as our test collection. This is a public dataset used in the TREC FedWeb track 2013.[1] The collection contains search result pages from 108 web search engines covering a variety of information sources, ranging from *"general web search engine"* (e.g., Google, Yahoo!), to vertical search engines that focus on specific media or genres (e.g., YouTube and Wikipedia). Examples of verticals are listed in table 6.1.

To provide a representation of each vertical search engine, several query-based samplings were provided for the vertical selection. For items (textual or multimedia documents) returned by each search engine, the authors collected relevance judgements by judging both the snippet created by the engine, and the

[1] https://sites.google.com/site/trecfedweb/

actual document content. The TREC Web Track 2010 queries were reused to collect documents. This test collection is well suited for the study of heterogeneous web search problems.

6.5.2 Evaluation

As mentioned before, there are several factors that can affect user perceived usefulness of a composite page. First, as assumed in the Cranfield paradigm setting, the **topical relevance** of items significantly contributes to page utility. In addition, we aim to form composite objects with each object reflecting a coherent aspect of a given topic, given that the **coherence** of items within an object (i.e., whether the items contained within the object focus on the same aspect) can negatively impact the utility of the page. Indeed, as shown in Dumais et al. (2001), presenting results incoherently in terms of topicality resulted in lower user satisfaction. The result set **diversity** (i.e., the set of composite objects) may also have an effect on user satisfaction. Sanderson et al. (2010) show that that there is a preference amongst users for systems that are measured to have more topical diversity, as determined by α-$nDCG$ (Clarke et al., 2008), for faceted or ambiguous informational needs. Finally, Arguello (2017) report that vertical diversity plays an important role in user satisfaction in a heterogeneous web search setting. All the above mentioned factors are important in evaluating the performance of our result composition framework. In this study, we measure performance by *topical relevance* (i.e., retrieval precision as determined using expert-assigned relevance labels), *coherence*, *topical diversity* and *vertical diversity*. We report performance results for each metric separately, allowing us to obtain a broader understanding regarding the effectiveness of our proposed approaches.

Since we are mostly concerned with results returned at the high rank of the page, we report our **relevance** performance based on precision metrics ($P@5$, $P@10$, $P@30$) and a set of top-heavy rank-biased metrics ($nDCG@5$, $nDCG@10$, $nDCG@30$). As discussed in section 2.2.2.1, these metrics evaluate how effective our proposed result composition algorithms are at placing relevant content higher in the results ranking, compared to traditional approaches, such as general web only ranking, or aggregated search. In order to compute precision and gain metrics on composite objects, we consider the composite result page as a linear ranking of items. Given that our algorithms output a ranked set of objects, with each object containing a ranked set of results, we construct a linear ranking of items by placing each result on the page according to the ranking of its corresponding object in the set of objects, and that of its position within the object itself. We use this approach because this allows us to accurately compare our approach to traditional ranking approaches (e.g., federated search), but also because the document collection on which we explored our result composition algorithms provides evaluation tools (i.e., relevance labels) that can be used reliably in such a setting. Methods for evaluating more complex approaches to ranking results extracted from multiple sources within a unified page are an active area of current information retrieval research (Zhou et al., 2012).

For *coherence* and *topical diversity*, we report normalised cohesion metric $tcoh(P)$ and normalised diversity metric $tdiv(P)$, respectively (described in section 6.3). These metrics relate to cluster quality (i.e., how similar the items contained within composite objects are to each other, or how dissimilar composite objects are from one another on the results page). Similar metrics, related to cluster quality, are used extensively in prior work on result composition (Amer-Yahia et al., 2014), to assess the quality of composite object and the effectiveness of composition algorithms. One of the limitations of using these types of metrics in evaluating composite objects is that they can be employed only in comparing different result composition or clustering approaches (i.e., they are estimates of cluster quality) rather than comparing result composition to other result ranking approaches, such as aggregated search. As such, our work extends prior work on result composition by evaluating composition approaches using traditional information retrieval metrics as well. For **vertical diversity**, we use the expected intent-recall $vdiv(P)$, which is the fraction of verticals (intents) with relevant items retrieved, that are present on the composite page P. Expected intent recall is used extensively in the evaluation of aggregated search pages (Arguello, 2017).

The evaluation approach used throughout this chapter is system-centred, in that it is intended to examine the effectiveness of result composition algorithms based on certain assumptions regarding user search behaviour (e.g., that users scan results linearly from the top to the bottom of a results page, that items within a composite object can be presented as a list of items to users). This evaluation approach is common in research on information retrieval models, as discussed in detail in section 2.2.2. Although chapters 4 and 5 provide a perspective on user-centred evaluation approaches for result composition, further work is required to adapt the algorithmic approaches discussed in this chapter to the user-centred design of modern search systems. In particular, understanding the display and presentation of composite objects in modern web search is a crucial aspect for the development of modern search systems and is one of the directions for future work that we discuss in depth in chapter 8.

6.5.3 Baseline Systems

We compare our result composition approaches with three baselines: ***(i)*** general web search engine only (denoted GW); ***(ii)*** traditional federated search systems (denoted FS); and ***(iii)*** aggregated search systems (denoted AS). General web search engines form the core of web search today and we use this as our primary baseline. Traditional federated search systems, rather than aggregating results, aim to merge rankings of different search engines within one single ranking. Aggregated search is the most similar approach to ours. However, different from our result composition work, in aggregated search systems, result presentation is based on a block paradigm (i.e., a block of homogeneous results, extracted from a single vertical, inserted within a ranking of general web results) and does not highlight topical connections between results originating from different sources. Note that comparing our approach with *general web* is not *"fair"*, as our approaches make use of additional information (i.e., results from other verticals). However, we include this to demonstrate the effectiveness of our approach in combining information from heterogeneous vertical sources on the web.

For the general web baseline (GW), we index the general web collection only and use BM25 as our ranking function. For the federated search baseline (FS), we use the state-of-the-art *ReDDE* (Si and Callan, 2003) resource selection approach to rank relevant verticals and *CORI* (Callan et al., 1995) result merging approach to merge results from different resources. We have another federated search baseline (i.e., Federated Search Central, abbreviated FSC). FSC is obtained by mixing items from all verticals into a central index and all items are ranked by a traditional ranking function (BM25).

For our aggregated search baseline, we rank verticals based on *ReDDE* (same as federated search) while we use a simple fixed-threshold approach to select relevant verticals (e.g., always selecting top-3 verticals as relevant for the query). For embedding the selected vertical results into the general web results ranking, we use a simple approach. Following previous work by Zhou et al. (2013), there are three possible embedding positions: top of the page (ToP), middle of the page (MoP) and bottom of the page (BoP). We simply embed our first (i.e., most relevant) vertical on ToP, second vertical on MoP and third vertical on BoP. Although we have not developed state-of-the-art aggregated search systems based on more advanced vertical ranking and selection methods (discussed in detail in Arguello (2017)), as this is not our main focus, we assume our baselines to be sufficient for illustrating and comparing our approaches to other established methods for heterogeneous web search.

6.6 Experimental Results

We report our experimental results in a homogeneous environment in section 6.6.1, followed by results in a heterogeneous environment in section 6.6.2. In the latter section, we study the importance of different sources of evidence in estimating relevance for our two approaches as well as the effectiveness of using entity representations of heterogeneous items in our framework. Finally, we study the robustness of result composition approaches. We aim to answer the following research questions — elaborations of our three main research questions:

(RQ1) In the homogeneous space, can result composition improve performance (in terms of Cranfield-style relevance metrics) compared to a general web baseline?

(RQ2) In the heterogeneous space, can we generate composite pages that are more relevant (in terms of Cranfield-style relevance metrics) than aggregated search or federated search rankings?

(RQ3) Which methods of estimating item and object relevance improve result composition performance, in terms of Cranfield-style relevance metrics?

(RQ4) Can we extract query-related entities as an anchor to bridge the semantic gap between items retrieved across sources of information for heterogeneous result composition?

(RQ5) How robust are the different result composition approaches to the quality of the initial rankings?

We use the following settings for our result composition algorithms. We set the constraint of the maximum number of items allowed within composite objects to be 3 to parallel similar web search settings (e.g., aggregated search blocks displaying three images). For a composite page, we assume 10 objects are presented.

6.6.1 Homogeneous Composite Retrieval

To answer ***(RQ1)***, we conducted our experiments on the general web only search engine. The baseline is GW (traditional BM25 ranking function with default parameters). Table 6.3 reports the retrieval performance of various result composition approaches (BOBO, BOBO-DT, BOBO-DE) in the homogeneous, general web only, environment. We do not include the CPS approach since it does not apply to a homogeneous environment (i.e., no satellite items can be attached). Note

	GW	*BOBO*	*BOBO-DT*	*BOBO-DE*
P@5	0.540	0.624	0.440	0.636
P@10	0.562	0.564	0.416 ▼	0.586
P@30	0.577	0.343 ▼	0.343 ▼	0.343 ▼
nDCG@5	0.333	0.479	0.350 △	0.461
nDCG@10	0.373	0.453 △	0.342 ▼	0.452 △
nDCG@30	0.428	0.352 ▼	0.314 ▼	0.349 ▼

Table 6.3: Performance of various result composition approaches in a homogeneous, general web (GW) only environment. Significance is determined using pairwise t-tests, values marked with △ $(p - value < 0.05)$, ▲ $(p - value < 0.01)$ and ▽ $(p - value < 0.05)$, ▼ $(p - value < 0.01)$ indicate respectively significant improvement or deterioration over the GW baseline.

that as in typical web search settings, generally, top ranking performance (e.g., top-10 or top-5) is a major evaluation concern. Pairwise t-test significance test is conducted to identify significant improvement or deterioration over the general web baseline. We identify several trends from table 6.3:

- In general, result composition approaches perform comparatively as well as the baseline in the top rankings. Overall, the BOBO-DE approach performs the best in this setting and can improve relevance performance over the traditional, general web only baseline, especially in the top ranking (indicated by both precision and nDCG metrics).

- All BOBO approaches perform better in the top ranking (e.g., top 5 results) but worse in the latter ranking (e.g., top 30 results). This is due to more irrelevant items being introduced by this approach from the lower ranking (i.e., higher probability to pick irrelevant items as a pivot for objects lower in the ranking).

- Comparing BOBO-DT and BOBO-DE, we can observe that promoting inter-object topical diversity after ranking objects can either deteriorate (*BOBO-DT*) or not affect (*BOBO-DE*) performance. This suggests that our approach to promote diversity is not effective to boost relevance in the homogeneous environment. This is not surprising since relevance and diversity have been empirically demonstrated to act against each other in homogeneous web search settings (Clarke et al., 2008).

Note that result composition in a homogeneous setting is similar to cluster-based retrieval. It is not surprising that our approach performs well in this setting since, in the context of cluster-based information retrieval, it has been shown that positioning documents of query-specific clusters at the top of the result list can improve retrieval performance (with respect to traditional relevance metrics) as compared to ranking documents directly.

Returning to ***(RQ1)***, we conclude that the result composition approaches proposed here can improve retrieval performance in the top ranking of a homogeneous web search environment.

6.6.2 Heterogeneous Composite Retrieval

In this section, we aim to test the effectiveness and robustness of our proposed result composition approaches.

6.6.2.1 Effectiveness

To answer ***(RQ2)*** and ***(RQ3)***, we conduct our experiments on the heterogeneous test collection. Part (a) of table 6.4 reports the performance of various result composition approaches (*BOBO, BOBO-DT, BOBO-DE, CPS, CPS-DT, CPS-DE*) in the heterogeneous web environment with federated search baseline (*FSC*). The federated search baseline (*FSC*) is obtained by mixing items from all different verticals into a central index and ranking items using a traditional ranking function (BM25). This system is generally assumed to be an oracle system (the upper-bound system performance) in federated search area. Part (b) of table 6.4 presents results related to different relevance estimation in BOBO and aims to report performance changes. In addition to relevance, results are compared based on a set of other criteria defined for evaluating composite retrieval (discussed in section 6.3): topical coherence (*tcoh*), topical diversity (*tdiv*) and vertical diversity (*vdiv*). These evaluation criteria can only be applied to compare the composite retrieval approaches and therefore the baseline system FSC is excluded in these comparisons. Pairwise t-tests are conducted to indicate significant improvement and deterioration of composite retrieval approaches over the federated search central (*FSC*) baseline. Our findings can be summarised as follows:

- Federating heterogeneous information is a challenging problem. Even the central federated approach (*FSC*) performs significantly worse than the homogeneous general-web only ranking (*GW*). For example, with respect to *nDCG*@10 in tables 6.4 and 6.5, we observe that *FSC*(0.287) < *GW*(0.373).
- Most of the result composition approaches perform better than the baseline (*FSC*) in the early rankings (top 10). *BOBO* is the worst performing ap-

	(a) Result Composition Approaches							(b) Adding Relevance Estimation			
	FSC	***BOBO***	***BOBO-DT***	***BOBO-DE***	***CPS***	***CPS-DT***	***CPS-DE***	***BOBO-VT***	***BOBO-VE***	***BOBO-E***	***BOBO-T***
P@5	0.448	0.500	0.568 △	0.536	0.536 ▲	0.604 ▲	0.560 △	0.568	0.508	0.572 △	0.600
P@10	0.460	0.510	0.540	0.534 △	0.568 ▲	0.588 ▲	0.568 ▲	0.514	0.520	0.538 △	0.568
P@30	0.505	0.472	0.472	0.472	0.296 ▼	0.296 ▼	0.296 ▼	0.436 ▼	0.456 ▽	0.466	0.464
nDCG@5	0.260	0.327	0.401 ▲	0.364 ▲	0.331 ▲	0.395 ▲	0.380 ▲	0.427 ▲	0.355 △	0.393 ▲	0.413
nDCG@10	0.287	0.345	0.400 ▲	0.367 ▲	0.373 ▲	0.412 ▲	0.398 ▲	0.404 △	0.367	0.388 △	0.415
nDCG@30	0.351	0.362	0.383	0.371	0.291 ▽	0.308	0.306	0.368	0.349	0.369	0.382
tcoh	-	0.301	0.301	0.301	0.676	0.676	0.676	0.289	0.262	0.271	0.297
tdiv	-	0.180	0.180	0.180	0.268	0.268	0.268	0.174	0.159	0.161	0.167
vdiv	-	0.260	0.260	0.260	0.115	0.115	0.115	0.241	0.255	0.266	0.273

Table 6.4: Performance of various result composition approaches in a heterogeneous web environment based on *Federated Search Central* (FSC): rankings based on a central index containing all verticals. Significance is determined using pairwise t-tests, values marked with △($p - value < 0.05$), ▲($p - value < 0.01$) and ▽($p - value < 0.05$), ▼($p - value < 0.01$) indicate respectively significant improvement or deterioration over the baseline (FSC).

	FSC	***FS***	***AS***	***OVS-FS***	***OVS-AS***	***GW***	***BOBO-DT***	***CPS-DT***
P@5	0.448 ▽▼	0.352 ▽▼	0.388 ▽▼	0.532	0.444 ▽▼	0.540	0.568	0.604
P@10	0.460 ▼	0.368 ▽▼	0.346 ▽▼	0.568	0.510 ▼	0.562	0.540	0.588
P@30	0.505 ▲	0.363 ▽	0.323 ▽	0.576 △▲	0.563 △▲	0.577 △▲	0.472	0.296
nDCG@5	0.260 ▽▼	0.192 ▽▼	0.229 ▽▼	0.303 ▽▼	0.241 ▽▼	0.333 ▽▼	0.401	0.395
nDCG@10	0.287 ▽▼	0.206 ▽▼	0.219 ▽▼	0.350	0.303 ▽▼	0.373 ▽▼	0.400	0.412
nDCG@30	0.351	0.229 ▽	0.219 ▽▼	0.409 ▲	0.388 ▲	0.428 △▲	0.383	0.308

Table 6.5: Comparison of best-performing result composition approaches in a heterogeneous web environment against all baselines. Values marked with △, ▲ indicate, respectively, significant improvements over BOBO-DT and CPS-DT (in this order). Similar convention with ▽, ▼ indicates values below BOBO-DT and CPS-DT. Significance is determined using pairwise t-tests with $p - value < 0.05$ in all cases (i.e., improvement or deterioration over the baseline).

proach whereas *CPS-DT* approach performs the best. Compared with *BOBO*, we can observe that promoting inter-object topical diversity after ranking the objects (*BOBO-DT*, *BOBO-DE*) can improve the performance of result composition.

- An interesting fact is that *CPS-DT* performs better than both *FSC* and *GW* in the early rankings. This is partly because of the conservative nature of the approach. It uses *GW* as the anchor and therefore it is more careful when selecting items from verticals.
- When comparing coherence and topical diversity, we observe that the *CPS* based approaches generally produce composite objects that are more coherent and topically diverse. However, compared with the *BOBO* based approaches, they are less vertically diverse. This can be explained by the fact that *CPS* is more conservative in the sense that it favours coherence and topical diversity when forming composite objects and only adds vertical items when it is sufficiently confident. On the other hand, the *BOBO* based approaches favour vertical diversity.
- When adding relevance estimation of items the result composition process, the *BOBO-E* approach incorporating the entity-based relevance $P(d_e|q_e)$ performs best and it improves over the baseline. Indicated by $nDCG@5$ and $nDCG@10$ from *BOBO-VT* and *BOBO-VE*, we observe that adding vertical orientation estimation $P(v|q)$ can also boost retrieval performance. *BOBO-T*, which incorporates only term-based relevance estimation $P(d|q)$, performs the worst and does not significantly improve over the baseline. This demonstrates the effectiveness of using entity representation for relevance estimation across heterogeneous verticals.

In table 6.5, we compare the best performing approaches (*BOBO-DT* and *CPS-DT*) to a set of baselines and demonstrate their effectiveness. We aim to find whether the best performing result composition approaches in a heterogeneous web environment can outperform other baselines (different search paradigms): general web search only (*GW*), federated search (*FS*) and aggregated search (*AS*). Table 6.5 compares each baseline against *BOBO-DT* and *CPS-DT* respectively and shows whether each baseline performs worse. Since we found in our experiments that vertical selection greatly affects retrieval performance, we also add two artificial systems (*OVS-FS*, *OVS-AS*) that use the oracle vertical selection (using *OVS* as prefix, indicating the upper bound of vertical selection performance). To obtain *OVS*, assuming that we have relevance assessments for the items, we rank all verticals based on the recall of relevant items and set a simple cut-off threshold (verticals with fraction of relevant items less than 10% are not selected). Those two systems aim to reflect the oracle performance of *FS* and *AS*. Several trends can be observed in table 6.5:

- When comparing different baseline search paradigms, we observe that *FS* and *AS* are similar, and *AS* outperforms slightly at top ranks, indicated by *nDCG@*5 and *P@*5. When the oracle vertical selection is added, the performance of each search paradigm increases and *OVS-FS* outperforms *OVS-AS*. As demonstrated before, *GW* performs well and the performance is similar to *OVS-FS*.

- Result composition approaches generally outperform all other search paradigms (*GW*, *FS* and *AS*) in the top rankings (top 5 or top 10). One interesting fact is that they outperform *GW* in the top ranking, which suggests that incorporating results from other vertical can improve retrieval performance. This is different from the conclusions we obtain when comparing *FS* and *AS* against *GW* where we found that heterogeneous federation degrades retrieval performance whereas *FS* and *AS* performed worse than *GW*.

- Another interesting aspect is that result composition, especially *CPS-DT*, can outperform other search paradigms where the oracle vertical selection is applied. This might be due to the fact that the proposed result composition is more conservative when adding vertical results as only results that are coherent and, thus, related to top-ranked and potentially relevant documents are added.

Going back to ***(RQ2)***, the result composition paradigm we propose can outperform both aggregated search and federated search ranking on relevance performance (in terms of Cranfield-style relevance metrics) in a heterogeneous environment. Returning to ***(RQ3)***, relevance estimation is useful for improving result composition approaches whereas both entity-based item relevance estimation and vertical orientation can improve performance. Since we observe that by using entities all our result composition approaches (*BOBO* and *CPS* based) perform comparatively well in a heterogeneous environment, with respect to ***(RQ4)***, we conclude that entities can be used as a bridge to link heterogeneous items.

6.6.2.2 Robustness

To answer ***(RQ5)***, we varied the initial rankings that result composition approaches are based on and investigate the robustness of our different approaches; results

	Original	*BOBO*	*BOBO-DT*	*BOBO-DE*
FS	0.206	0.221	0.243	0.234
AS	0.219	0.211	0.220	0.211
FSC	0.287	0.345	0.400 ▲	0.364 ▲
OVS-AS	0.303	0.308	0.377 ▲	0.308
OVS-FS	0.350	0.307	0.371	0.370
GW	0.373	0.453 △	0.342 ▼	0.451 △

Table 6.6: Robustness of various result composition approaches to initial ranking quality. All systems are measured by nDCG@10. Significance is determined using pairwise t-tests, values marked with △ ($p - value < 0.05$), ▲ ($p - value < 0.01$) and ▽ ($p - value < 0.05$), ▼ ($p - value < 0.01$) indicate respectively significant improvement or deterioration over the original ranking.

are reported in table 6.6. Robustness in this context refers to the extent to which a result composition approach can still perform well when the initial ranking is degraded in terms of relevance performance. Therefore, it is compared against the original ranking where the first column in table 6.6 specifies the original ranking (i.e., ranked by the relevance performance *nDCG*@10 in a ascending order). The column headed *"Original"* specifies *nDCG*@10 of the original ranking where the intersection of a given column and row specifies the *nDCG*@10 score of a given result composition approach that is based on the corresponding original ranking. Note that, because *CPS* constructs "central" objects based on general web search documents only, to which it then attaches documents from satellite verticals, it is not influenced by the quality of various initial rankings and it is not included in our robustness analysis. We can observe the following trends:

- When the performance (in terms of Cranfield-style relevance metrics) of the original ranking is low (e.g., *nDCG*@10 lower than 0.25), the result composition *BOBO* approach suffers from the large number of irrelevant items within the ranking and therefore does not produce composite objects that contain relevant items.

- There is a general trend that when relevance performance of the original ranking improves, the performance of *BOBO* approaches increases.

Returning to ***(RQ5)***, we conclude that, in general, result composition is robust with regard to initial ranking quality.

6.7 Discussion and Conclusions

Our objective was to investigate whether result composition can promote *relevance*, *coherence* and *diversity* in a heterogeneous multi-vertical web search environment. Our results indicate that result composition can improve performance (in terms of Cranfield-style relevance metrics) over various current search paradigms, such as traditional general web only, federated search or aggregated search ranking.

Through our study, we showed that result composition can improve retrieval performance in both homogeneous and heterogeneous web search environments ***(RQ1)***. In particular, in the heterogeneous environment, our proposed result composition approach *CPS-DT* outperformed current state-of-the-art search paradigms (i.e., general web search, federated search and aggregated search ranking). We also demonstrated that incorporating our proposed entity-based relevance estimation of items and vertical orientation estimation (based on the state-of-the-art resource selection approach ReDDE) improves result composition approaches compared to those that disregard them ***(RQ2)***. Finally, we found that our proposed result composition approaches can be robust with respect to the quality of initial ranking, in a heterogeneous web environment ***(RQ3)***.

Our results have implications for work in heterogeneous information access and diversity in information retrieval and web search. The result composition search paradigm we discuss in our study aims to promote a diverse information space for users to explore. In our case, diversity is promoted in two dimensions, topical and vertical. For a multi-faceted search task, rather than issuing multiple queries with respect to different aspects of an information need, to several vertical search engines, result composition can provide a unified page that consists of *relevant*, *coherent* and *diverse* composite objects that can help users better navigate the query-result-evaluation cycle that is typical in web search. However, promoting both topical diversity and multi-vertical information aggregation is challenging (Arguello, 2017; Clarke et al., 2008) and our work is an incipient effort towards understanding how both topical and vertical diversity can be promoted in web search. We show that without affecting relevance, we can effectively promote diversity in both dimensions. Our work opens a fruitful research avenue as heterogeneous information access is becoming more and more present in everyday web search interactions.

6.8 Chapter Summary

Retrieving results from heterogeneous sources and presenting them in a unified interface is a difficult problem. Aggregated (web) search has become the most prevalent method for selecting and displaying results from different sources on a single page. Even though widely adopted by modern search engines, aggregated search is limited in that it merges blocks of homogeneous content within a heterogeneous ranking. To date, prior work has not investigated in detail *which results* from a particular vertical, rather than just the top few, to display within aggregated search blocks on the results page (Arguello, 2017). This is an important aspect of merging heterogeneous content within a unified page as prior studies have shown that results from one source can influence user engagement with results from other sources (Arguello and Capra, 2016). As such, understanding the interactions between heterogeneous items can be informative for selecting which result to extract from verticals, rather than returning the top few. Secondly, limited effort has been put into understanding methods for the display of heterogeneous content on the results page. It remains unclear, for instance, why, post-retrieval, heterogeneous content needs to be aggregated by type (e.g., a block of image results or a block of video results), rather than topic or other features, or structured in any other way.

The work presented in this chapter is an attempt at moving beyond merging blocks of results from different verticals within a singular ranking, as in the case of aggregated search, and constructing *composite objects*, focused on specific topics of a searcher's query and containing results from multiple verticals, as a way of giving users access to heterogeneous content within a unified results page. Our work investigates multiple algorithmic approaches for constructing composite objects and explores how composite object properties can be manipulated and assessed under constraints.

Our experiments show that result composition can improve retrieval performance in both homogeneous and heterogeneous web search environments; in a heterogeneous environment, our result composition approach outperformed current state-of-the-art search paradigms, such as general web search ranking only, federated search and aggregated search. In particular, our *central-plus-satellite* result composition approach provides further evidence that constructing composite objects around a ***central set*** of relevant documents leads to highest performance, in terms of Cranfield-style evaluation metrics, relative to other approaches, further supporting the statement of this thesis.

One of the main challenges in constructing composite objects is the disparity of term distributions across verticals (Santos et al., 2011), which in turn makes similarity assessment for heterogeneous documents problematic. The next chapter of this thesis tackles this challenge by attempting to create a unified representation space for heterogeneous items. We explore the application of a click-graph representation learning mechanism to a heterogeneous web search environment and show that graph-learnt representations can lead to cross-vertical document aggregates that are more relevant.

Chapter 7

Click Graph Representation Learning for Heterogeneous Documents

Click-through logs have been used for various tasks in information retrieval: from query suggestion or direct result ranking to representation learning. Recent work has shown that query and document representations learnt through term propagation in click-graphs can be used effectively for search result ranking.

In this chapter, we present an experiment investigating the application of graph-learnt term representations to the retrieval and aggregation of heterogeneous content. To this end, we construct a click-graph using log data from a commercial web search engine containing tens of millions of search interactions, generated by over a million unique searchers, in a real web search setting. We then describe the structure of the click-graph, emphasising connectivity properties of heterogeneous documents within the graph, and highlighting potential limitations that graph structure introduces on representation learning for heterogeneous documents. In addition, we explore simple methods of manipulating graph structure in order to overcome these structural biases. Our findings show that graph-learnt representations can be used effectively in retrieving and aggregating documents across verticals, and that modifying graph structure can lead to improvements in both ranking and aggregation performance, but that this improvement is dependent on how the propagation algorithm is configured and what data is used to initialise it.

7.1 Introduction

Modern web search engines incorporate implicit user feedback (such as clicks on search results) into a wide range of techniques aimed at improving search experience. To achieve this, traces left by users' interactions on the search page — which typically include queries issued to the system, mouse movements, and clicks on the displayed results — are monitored and recorded at scale, into what is commonly known as a *click* or *query log*. Popular search systems collect millions of query-document click pairs daily, with each click being considered a weak indicator of document relevance to a given query. Even though clicks are not equivalent to explicit relevance judgements, there is extensive evidence suggesting that they are useful for direct ranking of documents (Agichtein, Brill and Dumais, 2006; Agichtein, Brill, Dumais and Ragno, 2006); indeed, Jiang et al. (2016) report that click-based features are used as one of the *primary* signals to improve ranking quality for popular queries in modern web search engines.

One of the problems with using historical click data to improve ranking is that there is limited coverage in past click data with respect to all possible query-document pairs on the web. Indeed, many relevant documents might not be present in historical click logs because they are not retrieved by search engines in the first place (e.g., a document that is very recent and has not been indexed). This means that click-based features can be computed for a small set of documents, given a user query (i.e., documents that have been shown to searchers previously), a problem that is more prominent for less frequent queries. Even more, searcher preferences with respect to search results typically change over time. Documents in a historical click-log might reflect past user preferences rather than current ones. And in addition to sparse coverage, clicks can be noisy, generated by users exploring search result rankings in order to locate relevant items (e.g., a user clicking multiple results in order to find at least one that is relevant). Together, noise, latency and sparsity affect the quality of click-based features and can be detrimental to ranking methods that employ them.

Prior work has shown that using graphs constructed from click-through data (i.e., *click-graphs*) to learn vector representations for both documents and queries in the same feature space is an effective way of dealing with the problems of using click log information directly. A range of methods that learn vector representations in latent spaces through graph-based approaches have been proposed and shown to be effective (Gao et al., 2010; Shen et al., 2014; Wu et al., 2013). More recently, Jiang et al. (2016) proposed a unified framework through which both click and content information is used to learn term-based vector representations for queries and documents, in a common feature space, through the propagation of term vectors in a large-scale click graph. Their method generates representations that bridge the vocabulary gap (Müller and Gurevych, 2009) between queries and documents, are human-readable and generalise to unseen queries or documents.

The advantages of learning query and document representations in a common feature space are manifold. In addition to bridging the semantic gap between queries and documents, documents of different *types* are represented in the same feature space as well. Although graph-based representations have been shown to improve retrieval performance in general, whether they are effective across verticals of heterogeneous content (e.g., *images*, *video*, *news*) remains an understudied problem. Inherent biases in the ***(i)*** initial term representations used in the representation learning algorithm for different types of documents (e.g., image documents have less informative textual content compared to Wikipedia documents) and ***(ii)*** the distribution of heterogeneous documents across graph components can render graph-based representation learning ineffective in the case of documents originating from different sources. In addition, a common representation space across verticals lends itself to assessing the similarity of heterogeneous documents. Constructing cross-vertical document aggregates (e.g., aggregated search pages or composite objects) is an interesting yet equally understudied application of graph-learnt document representations.

Graph structure and connectedness are the underlying basis of graph-based representation learning algorithms. The inter-play between graph structure and propagation algorithms is apparent: graph elements (e.g., vertices or sub-graphs) that are more connected can benefit from their immediate neighbourhood and converge on more expressive vector representations. In contrast, graph elements that are isolated or disconnected converge on trivial representations that are not discriminative (discussed in more detail in section 7.3.3). As such, understanding how click graphs are structured in terms of their constituent components can be useful to reinforce the benefits (and adjust for the limitations) of graph-based representation learning algorithms.

Informed by an understanding of the structural properties of click-graphs, manipulating graph structure can potentially enhance the quality of representations learnt through term propagation. Typical click-graphs are disconnected (i.e., fragmented into multiple sub-graphs), with a majority of vertices incorporated in very small sub-graphs in which propagation mechanisms are less effective. These small components are typically pruned when constructing large-scale click-graphs, effectively discarding a large proportion of click information collected in logs. Increasing graph connectivity can not only link isolated sub-graphs and thus potentially improve the effectiveness of propagation algorithms for queries and documents located in fragmented components, but also prevents excessive pruning, keeping more click information in the graph. Even more, additional edges in the graph can be beneficial within connected components as well, by linking related items (e.g., queries) through direct paths, rather than intermediate paths, thus potentially improving the quality of vectors learnt through propagation. However, modifying graph structure with the goal of enhancing term propagation algorithms is not trivial. Adding arbitrary edges in the click-graph can increase noise in the document and query vectors, and at the same time increase computational costs associated with graph-based algorithms.

Our contributions in this chapter consist of an in-depth analysis of click-graph structure, with emphasis on the distribution of heterogeneous documents across sub-graphs. We show that not only are direct rankings derived from graph-learnt representations effective across verticals of heterogeneous content, but that document representations can also be used to bridge the cross-vertical gap and construct more relevant cross-vertical document aggregates. Finally, we show that using simple methods to increase the number of edges in a click-graph can lead to more informative document and query representations learnt through term vector propagation. Specifically, in our study, we aim to answer the following research questions:

(RQ1) *(1.1)* How connected are graphs constructed from click-through log data? *(1.2)* How are queries and documents from different verticals distributed across graph components?

(RQ2) *(2.1)* How effective are term representations learnt through term vector propagation in retrieval, across verticals of heterogeneous content? *(2.2)* How can these representations be used to create relevant cross-vertical document aggregates (e.g., aggregate search pages)?

(RQ3) *(3.1)* How do simple modifications of graph structure influence graph connectedness? *(3.2)* How do these changes to graph structure affect term vectors learnt through vector propagation, and in turn, derived rankings?

7.2 Prior Work

Central to our work is the graph-based representation learning algorithm proposed by Jiang et al. (2016). In their work, Jiang et al. (2016) put forward a unified framework through which both click and content information is used to learn term-based vector representations for queries and documents. We extend their contribution by evaluating the representation learning algorithm in the context of retrieving and aggregating heterogeneous documents. Our work brings together two threads of prior effort in this space: the use of click graphs in information retrieval and methods for improving aggregated search quality.

7.2.1 Click Graphs

Click graphs have been used extensively in information retrieval, in various application areas, such as query to document matching (Craswell and Szummer, 2007) or query suggestion (Beeferman and Berger, 2000; Wen et al., 2001). More related to our work, click graphs have been used to generate query and document representations in a common space (Gao et al., 2010; Jiang et al., 2016; Shen et al., 2014; Wu et al., 2013; Xue et al., 2004). In contrast to prior work, our study focuses on the evaluation of retrieval and aggregation of heterogeneous documents specifically, leveraging representations learn through term propagation in a large-scale click graph.

7.2.2 Aggregated Search

The aim of aggregated search is to provide search access to documents originating from a wide range of heterogeneous sources of information (i.e., *verticals*) from a unified interface. Our work is orthogonal to prior research in this space by investigating the use of underlying document representations in constructing cross-vertical aggregates. We review prior work related to aggregated search in more detail in sections 2.2.3.2 and 3.2.3

7.3 Data and Methods

Search engines collect information about queries issued and documents clicked on by searchers. If a document is clicked by a user for a given query, the query and document form a co-clicked pair. Click events are typically enhanced with meta-data about search activity, such as event timestamp, and records of click events form click-through logs or click logs. In our work, we make use of a large-scale click log, sourced from a modern web search engine, containing real-world search interactions. We describe this click log in more detail next.

7.3.1 Click Log

We make use of click-through data sourced from a large-scale, commercial web search engine. To sample click events, from a pool of desktop-only search users in the United States, we first identify a subset of users who issued at least one query to the system on the 26th of November, 2016, and track their search activity for exactly one week. Although not strictly a uniform random sample, this type of sampling has been employed in previous work (Silverstein et al., 1999) and we consider our sample to be representative of search behaviour in general. To this log, we applied several layers of filtering and further sampling. Firstly, we removed all queries that start or end with typical URL components (e.g., *"https://"*, *"www."*, *".com"*, *".org"*); secondly, we removed all queries that contain only English stopwords or have length less than three characters. In total, our log contained tens of millions of search interactions generated by over one million unique searchers of the web.

In addition to click meta-data (e.g., click timestamp), our log contained the query issued by the user, the clicked search result URL, title and snippet associated with the result, and, for aggregated search pages containing results from different verticals or for queries issued directly to vertical search engines, a label indicating result type: *"general web"* (abbreviated throughout this chapter as *"gw"*) which refers to textually rich web resources, such as Wikipedia pages or news articles, *"image"* or *"video"*. This additional labelling for heterogeneous documents allows us to explore the effectiveness of graph-based representation learning for documents of different types. Of the unique documents we observed in our log, 1.2% were labelled as images and 0.8% as videos.

7.3.2 Click Graphs

In this section, we formally describe our click-graph variants. Figure 7.1 shows an example click-graph, in which queries are represented by dark circles, and labelled with their corresponding terms, and documents are represented with light circles, labelled with corresponding URL and vertical label (i.e., *web*, *image* or *video*). The graph in figure 7.1 also contains three unconnected components: at the top, the largest component in our example, containing q1 and q2 and their adjacent document vertices; in the middle, the component containing query q3 and at the bottom, the component containing query q4 and its adjacent document vertices. We begin by introducing formal notation for an unmodified click-graph, and then describe graph structural modifications based on textual or temporal proximity of user queries. Our work closely parallels that of Jiang et al. (2016) and, in consequence, we use similar notation.

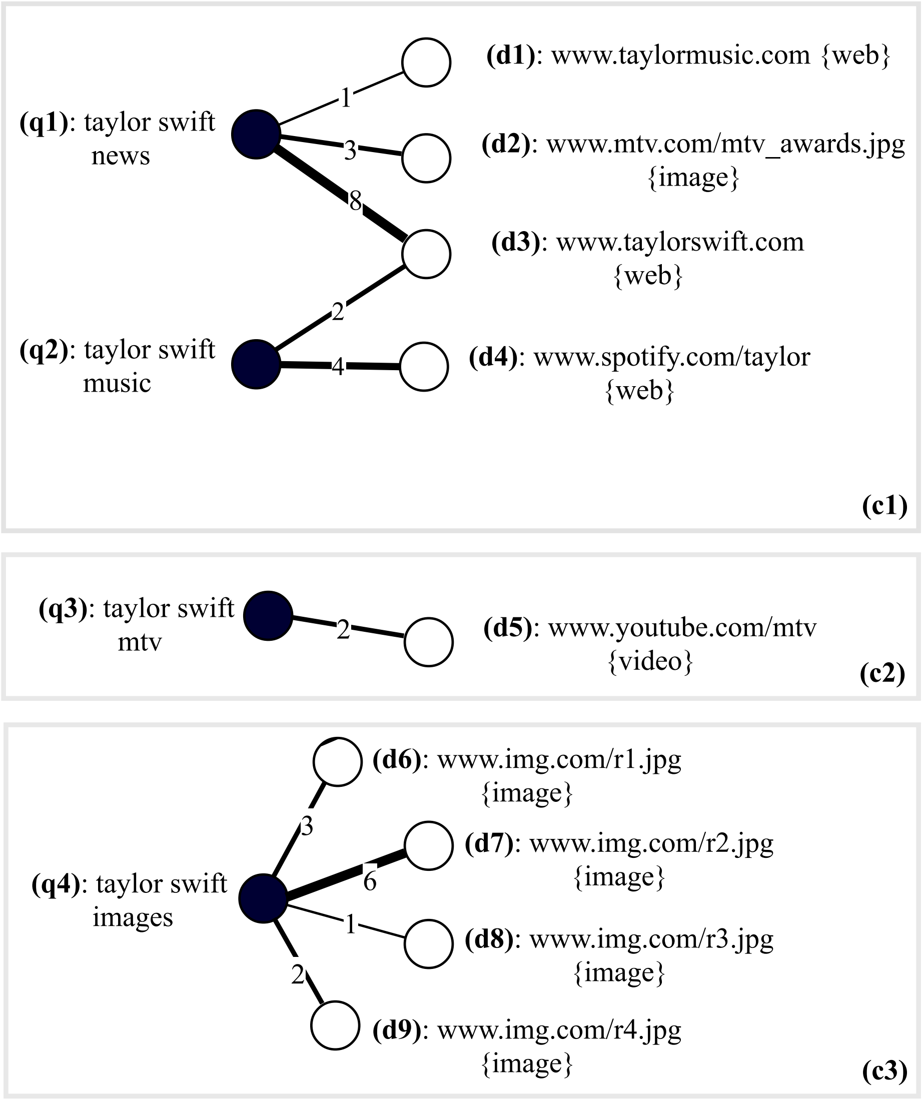

(d1): www.taylormusic.com {web}
1
(q1): taylor swift news
3
(d2): www.mtv.com/mtv_awards.jpg {image}
8
(d3): www.taylorswift.com {web}
2
(q2): taylor swift music
4
(d4): www.spotify.com/taylor {web}
(c1)
(q3): taylor swift mtv
2
(d5): www.youtube.com/mtv {video}
(c2)
(d6): www.img.com/r1.jpg {image}
3
(d7): www.img.com/r2.jpg {image}
6
(q4): taylor swift images
1
(d8): www.img.com/r3.jpg {image}
2
(d9): www.img.com/r4.jpg {image}
(c3)

Figure 7.1: Example bipartite click-graph that contains three disconnected components (***c1***, ***c2*** and ***c3***). Queries are represented by dark circles, and labelled with their corresponding terms, whereas documents are represented with light circles, labelled with corresponding URL and vertical label (i.e., *web*, *image* or *video*).

7.3.2.1 Original Graph

Let *Query* be the set of all queries in the log and *Doc* the set of all documents in the log. For a query $q \in Query$ and a document $d \in Doc$, if there is a co-click between them, an undirected edge is added between their corresponding graph vertices. The weight of the edge is determined by the number of co-clicks between the query and document, aggregated across users. We denote the set of edges in our graph as *Edge*.

To construct the graph, we represent both *Doc* and *Query*, and the adjacency matrix of the graph — denoted by C — in vector form. The entry in the i^{th} row and j^{th} column of C equals the number of co-clicks, aggregated across users, between query q_i and document d_j. We denote the vector representation of *Query* as Q, a matrix in which the i^{th} row (Q_i) is the vector representation of query q_i. Given a vocabulary size of V, the size of Q is $|Query| \times V$, and element q_{ij} of Q represents the weight of term j in the representation of query i. Analogously, we denote the vector representation of *Doc* as D, a matrix in which the i^{th} row corresponds to the vector representation of document d_i, and is of size $|Doc| \times V$.

7.3.2.2 Augmented Graphs

In our attempts to modify graph structure, we create pseudo document vertices to connect query vertices that are related: given two query vertices, q_i and q_j, a pseudo-vertex ps is added to *Doc*, such that both q_i and q_j are connected to ps and the weight of each of their connecting edges is w. Pseudo-vertices rather than direct edges are used to connect related queries in the augmented graphs because this preserves the bipartite nature of the click graph and allows for the representation learning algorithm to remain unchanged. Figure 7.2 shows an example bipartite click graph augmented through the use of pseudo-vertices. We explore two methods of manipulating graph structure next.

Text Augmented Graph. In our first attempt at modifying graph structure, we create pseudo-vertices in the click graph to connect query vertices that are lexically similar. Lexical similarity is determined using the normalized Levenshtein edit distance (Yujian and Bo, 2007) and additional connections are added in the case of query vertices for which lexical similarity is greater than or equal to an arbitrary threshold of 0.8, which we assume indicates strong lexical similarity. Creating graph links based on lexical similarity is motivated by the fact that similar queries can have non-intersecting co-clicked document sets (either due to

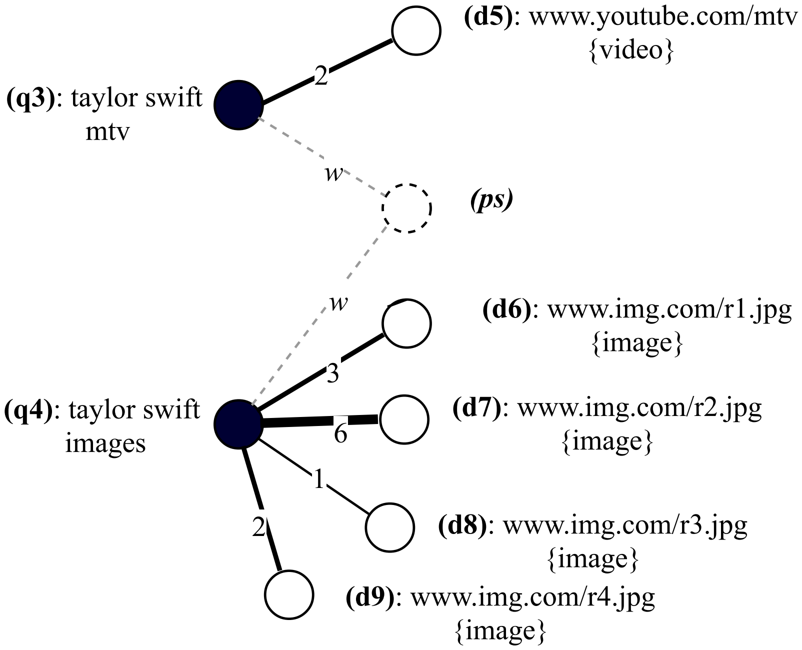


Figure 7.2: Example bipartite click graph containing two disconnected components connected through a pseudo document vertex ***(ps)*** with edge weights w.

click sparsity, user preferences or log sampling). Thus, connecting similar queries through direct paths can potentially benefit retrieval performance.

Time Augmented Graph. In addition to lexical similarity, we investigate the use of temporal proximity as an indicator of query relatedness, and add connections between queries which, across users, are frequently issued in close succession. Adding temporal edges to the graph is motivate by prior work (Odijk et al., 2015) which has shown that users typically issue related queries within individual search sessions. As such, creating direct paths in the graph between queries which are determined to be temporally related, across searchers, might lead to more discriminative representations for queries and their relevant documents in the term vector propagation process.

Formally, given that our approach to manipulating graph structure does not change its bipartite nature, the notation for both lexical and temporal augmented graphs remains unchanged. The weight of edges connecting related query vertices to their common pseudo-vertex is determined by the similarity function used to quantify relatedness. In the case of lexical similarity, the weight is set to the normalized Levenshtein distance if greater than 0.8 or 0 otherwise. In the case of temporal similarity, we use an exponential decay function to determine whether two queries are related. Specifically, if a user issues two queries within 30 minutes of each other, an edge is added between the two queries in the graph, weighted by the function: $w(t) = e^{-\lambda t}$, where t is the time in minutes between issued queries. Edge weights are then aggregated across users, thus if multiple users issue the same queries in succession of each other (regardless of query order) the weight between them is accumulated (summed); temporal edges with weight less than 0.1 are then pruned. As in the case of lexical similarity, both the parameter of the decaying function ($\lambda = 0.07$), the 30 minute similarity window, and the pruning threshold are based on our empirical assumptions of similarity.

7.3.3 Propagation Algorithm

We use the term vector propagation algorithm described in Jiang et al. (2016) as our starting point. The algorithm they propose is similar to the score propagation in hyperlink-induced topic search (HITS) algorithm (Kleinberg, 1999). The goal of the algorithm is to learn representations for queries and documents in the same feature space by propagating term vectors, derived from query or document text content, through the graph structure. Given a bipartite click graph, consisting of queries, documents and their co-clicks, the algorithm starts with content information initialized as vectors on either side of the graph, and propagates the term vectors to the connected nodes on the opposite side of the click-graph. During the propagation steps, the vectors are weighted by the number of co-clicks between queries and documents such that more frequently co-clicked terms gain higher weights, whereas less informative terms are gradually phased out. Formally, given a document d_j, at the n^{th} iteration of the algorithm, term vectors are updated using:

$$D_j^{(n)} = \frac{1}{\left\| \sum_i^{|Query|} C_{i,j} \cdot Q_i^{(n-1)} \right\|_2} \cdot \sum_i^{|Query|} C_{i,j} \cdot Q_i^{(n-1)}$$

where the $\mathcal{L}2$ norm is used to normalise term weights. Similarly, query term vectors are updated using:

$$Q_i^{(n+1)} = \frac{1}{\left\| \sum_j^{|Doc|} C_{i,j} \cdot D_j^{(n)} \right\|_2} \cdot \sum_j^{|Doc|} C_{i,j} \cdot D_j^{(n)}$$

The propagation algorithm can be initialised with either document *or* query terms, and depending on the initialisation step, the propagation mechanisms starts on the document or the query side of the graph. In-depth details and evaluation of the algorithm are provided in Jiang et al. (2016).

Through the propagation process, term vectors grounded in the same semantic space (e.g., query or document vocabulary space) are learnt for both queries and documents. The benefits of this approach are manifold:

(i) ***Vocabulary mismatch between queries and documents***: prior work studying the relatedness of query and document terms (Müller and Gurevych, 2009) has shown that 13.5% of relevant documents do not contain *any* of the query terms and therefore cannot be retrieved by traditional term-matching algorithms. [1] Using the term-propagation algorithm, a representation grounded in the same feature space is learned for both queries and documents, effectively bridging the vocabulary mismatch.

(ii) ***Feature interpretation and debugging***: term representations learnt through vector propagation are directly interpretable, unlike recent methods that learn representations in latent spaces, such as word embeddings. In comparison to latent-space methods, term-vector propagation generates representations in an interpretable, *"human-designed"* space (i.e., words).

(iii) ***Cross-vertical document representation***: the vocabulary mismatch is present in the case of heterogeneous Web documents as well. For example, the mismatch between initial term representations for image documents, which typically have poorer term representations, compared to Wikipedia documents, which contain rich textual information. In our log, images and videos have shorter snippets that, although can reflect query terms, are not discriminative with respect to similar images (e.g., images retrieved for a query such as *"celebrity images"*).

Graph Structure and Propagation. Graph structure and connectedness is the underlying basis of graph-based representation learning algorithms. In graph theory, a connected component of an undirected graph is a sub-graph[2] in which any two vertices are connected to each other by paths, and which is not connected to additional vertices in the super-graph. Figure 7.1 shows an example click graph containing three unconnected components.

The interplay between graph structure and term vector propagation becomes apparent if we consider the following example: in figure 7.1, even though the queries represented in the graph appear to be related (e.g., q_1 *"taylor swift news"*

[1] Experiments run using data from the HARD track at the TREC 2003 conference.

[2] We use the terms *sub-graph* and graph *component* interchangeably.

and q_4 *"taylor swift images"*), they are not directly connected in the graph because their co-clicked document sets do not intersect. Although artificial, our example helps illustrate some of the issues that click-graph structure can raise with respect to the graph-based representation learning algorithm we explore in this work. We highlight certain types of sub-graphs that can generate poor representations due to their implicit structure:

(i) ***Single-query components***: in disconnected components, such as the graph component containing a single query vertex q_4 and document vertices d_6, d_7, d_8, d_8 in figure 7.1, the propagation algorithm encodes no additional information besides query-document co-click. Even though term re-weighting does occur, the propagation learnt vectors for all documents in this structure will be identical, because all document vertices are connected to a single query vertex. In such cases, term-vector propagation is less useful.

(ii) ***Single-vertical components***: in the graph component containing query vertex q_3 and document vertices d_5, d_6 d_7, d_8, because of sparse initial term vector representation for the co-clicked documents (e.g., all documents being represented by vectors containing the terms *"celebrity"* and *"images"*), the vectors output by the propagation algorithm will not benefit from neighbouring vertices with richer textual representations;

(iii) ***Poor local connectivity***: although we draw attention to sub-graph structures in which propagation algorithms can be less effective, we note that these types of structures can exist within connected components as well (i.e., poor local connectivity) and therefore have impact on the representation learning mechanism, regardless of connectivity across components.

We highlight the effects of graph structure and connectedness on term vector propagation algorithms as they inform our experimental design, which is discussed in the following section.

7.3.4 Experimental Design

Using log data, we construct the click-graph and its variants and iterate the representation learning algorithm until convergence. The learnt term vector representations are then used to compute query-document similarity scores and directly rank documents for a given query. We evaluate ranking performance using traditional metrics: precision (P@k, MAP) and cumulative gain (nDCG@k).

For our evaluation, we sample queries from the graph based on the connectivity properties of their corresponding vertices in the unmodified click-graph. Therefore, we randomly sample 30 *head* queries, which occur more than 1000 times in our log, and are part of the largest component of the unmodified click-graph, and 30 *tail* queries, which appear less than 10 time in our data, and are part of small, disconnected components. For each of the queries we generate results rankings containing ten documents, using representations learnt under different graph and algorithm configurations.

Given that our log data is sourced from real-world search interactions, we do not have relevance judgements associated with query-document pairs, thus we use crowd-sourcing to annotate individual results with relevance labels.

Crowd-sourcing relevance labels. We collect relevance labels using the Amazon Mechanical Turk (MTurk) platform. Each Human Intelligence Task (HIT) corresponded to a single relevance assessment. Workers were shown one query and one search result (i.e., document surrogate containing title and snippet for *general web* documents, or the image itself in the case of images) and asked to rate its relevance to the query on a three-point scale: not relevant, somewhat relevant, and very relevant. We used this type of assessment for both *general web* and *image* results. Each hit was priced at $0.01 USD; HITs had a mean duration of 11.66 seconds. For each query-document pair, we collected three redundant relevance labels, and, in total, we collected 13468 relevance labels for general web results and 7626 for image results.

For quality control, we allowed only workers with approval rating higher than 98% and more than 1000 approved HITs, from English speaking countries[3] to attempt our tasks. In addition, we allowed individual workers to complete at most 100 labelling tasks. In total, 517 unique workers completed our tasks. For *general web* labels, Fleiss $kappa = 0.553$, and for *image* labels Fleiss $kappa = 0.504$, which both indicate fair to good inter-rater agreement. For our ranking evaluation, we use the median assessment value across annotators as our final relevance label.

Algorithm variants. We compare rankings derived from various underlying query and document representations. Firstly, we hypothesize that graph-learnt representations can be used effectively in the retrieval and aggregation of heterogeneous content. Secondly, we hypothesize that modified graph structures generate more discriminative representations and therefore, higher quality rankings.

[3] USA, UK, IRL, AU, NZ, CAN

Thus, we compare rankings derived from representations learnt in the following graph configurations:

- ***(i) Unmodified***: refers to rankings derived from representations learnt using the original propagation algorithm proposed by Jiang et al. (2016), using an unmodified click-graph. In our results, we refer to this variant of the algorithm as *VP*.

- ***(ii) Text augmented***: refers to representations and rankings derived from the text augmented graph, using additional connections between query vertices based on lexical similarity, as described in section 7.3.2.

- ***(iii) Time augmented***: refers to representations and rankings derived from the time augmented graph, using additional connections between query vertices based on temporal proximity, as described in section 7.3.2.

The term vector propagation algorithm can be initialized using either query or document terms (i.e., representations can be learnt in either the query or document vocabulary space). We therefore append the suffixes *"_query"* or *"_doc"* to the results we report in the following sections in order to indicate which terms were used when initialising the representation learning mechanism.

Baselines. To gain additional insight into how well graph-based representations capture query-document relevance, we compare rankings derived from term vector propagation to multiple baselines:

- ***(i) Click-through rate (CTR)***: documents are ranked using click-through rate for a given query, as observed in our log sample.

- ***(ii) BM25***: given that we use document title and snippet in our propagation algorithms, we use the traditional BM25 ranking function (Robertson et al., 1995), instead of its variants, over these fields as a baseline.

- ***(iii) Word Mover's Distance (WMD)***: using the word embedding vector set trained on Google News (Mikolov et al., 2013), we measure query-document relevance using the framework proposed by Kusner et al. (2015), and rank documents accordingly.

- ***(iv) Graph neighbourhood***: click graph structure allows for applying simple algorithms to determine query-document similarity, beyond just co-click information. Previous work has shown that applying agglomerative clustering to the vertices of a bipartite click-graph is an effective way to identify related queries and documents Beeferman and Berger (2000). Given two vertices, we can determine their similarity by computing the overlap (Jaccard coefficient) between their respective neighbourhoods, where neighbourhood is defined as the set of vertices adjacent to a given vertex.

7.4 Experimental Results

Click graphs are defined by queries and their respective co-clicked document sets. In turn, intersecting document sets across queries determine the structure of the click-graph. To understand the interactions between graph connectedness and term propagation algorithms, we first report on the structure of our click-graph.

7.4.1 Graph Connectedness

In our data, we observe relative sparseness regarding the intersection between queries and their co-clicked document sets. As expected, this sparseness is more pronounced for torso and tail queries, for which co-clicked document sets are typically smaller than for head queries. In consequence, the click graph constructed from our log is disconnected, consisting of multiple sub-graphs of various sizes. Roughly 38% of all documents and 43% of all queries are contained within the largest component of our graph (by number of vertices), with the remaining documents being distributed across relatively small components — the second largest component of our graph containing less than 1% of all vertices.

To understand whether the structure we observe is typical of click graphs in general or an artefact of our sampling method, we conduct a simple experiment in which we simulate different graph structures by randomly sampling from our initial click log. At each iteration, we extract sub-samples of increasing size (from 10% up to 90% of our initial log) and construct click graphs with each sub-sample. Figure 7.3 shows the relationship between sub-sample size (as proportion of our initial log), and size of largest sub-graph (as proportion of total vertices contained) in each of the graphs we constructed. Each marker represents a sub-sample of our log and its corresponding graph, while the dashed lines show simple quadratic approximations of the relationship between click-log sample size and the proportion of total vertices contained in the largest sub-graph. It is interesting to note that all graphs we constructed were disconnected, and that, consistently, the largest components of the graphs incorporate only a *minority* of query (28% - 43%) and document (22% - 38%) vertices, with the majority of the graph being fragmented into much smaller, disconnected components. Also

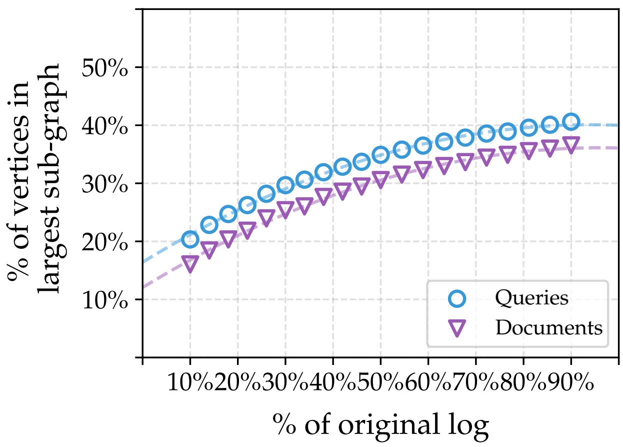


Figure 7.3: Simulating graph structures of increasing size by randomly sampling higher proportions of our initial click log. With each sample we construct a click graph and measure the size of its largest component (by total number of query and document vertices contained).

interesting to note is that increasing the amount of click data used to construct graphs does increase connectivity, but with diminishing effect. We estimate that the graph we observe in our analysis is representative of large-scale click graphs in general and we proceed by describing its structure in more detail.

The interplay between graph structure and propagation algorithms is apparent: as discussed in section 7.3, sub-graphs in which, for instance, either few vertices exist or all document vertices are connected to a single query vertex are susceptible to generating uninformative term vectors through propagation, compared to more connected counterparts. Given that our click-graph is disconnected, and that more than 50% of query and document vertices are distributed across small, fragmented sub-graphs, we set out to quantify graph components in which term propagation is less effective.

Figure 7.4 (left) shows the distribution of document and query vertices across sub-graphs, partitioned by the number of document vertices they contain. Even though approximately 40% of query and document vertices are contained within our graph's largest component, roughly 30% of all document vertices and 40% of all query vertices are located in small sub-graphs that contain fewer than 10 document vertices. Furthermore, figure 7.4 (right) shows the distribution of query and document vertices across sub-graphs, partitioned by the number of query vertices they contain. More than 30% of queries and documents are contained within single query components, which, as discussed in section 7.3, is a structure that can be detrimental to the quality of term representations learned through term vector propagation.

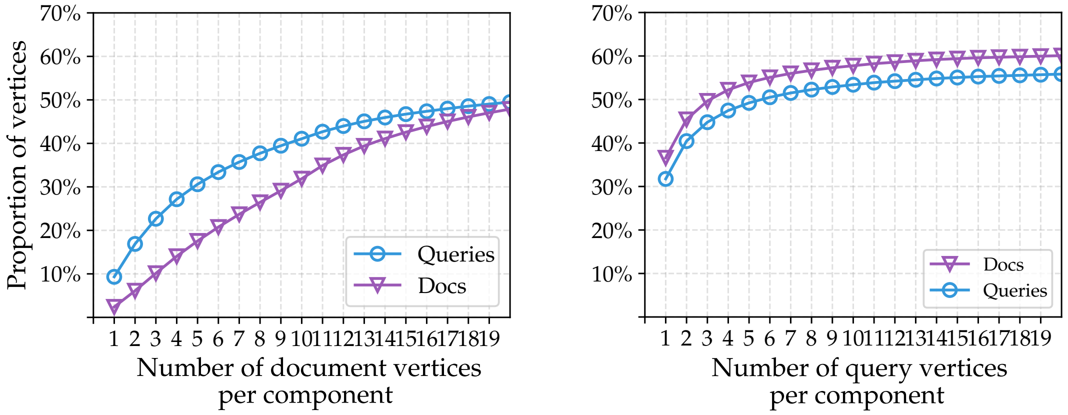


Figure 7.4: Distribution of query and document vertices across small sub-graphs in our unmodified graph variant. The majority of queries and documents in our click log are distributed across small graph components.

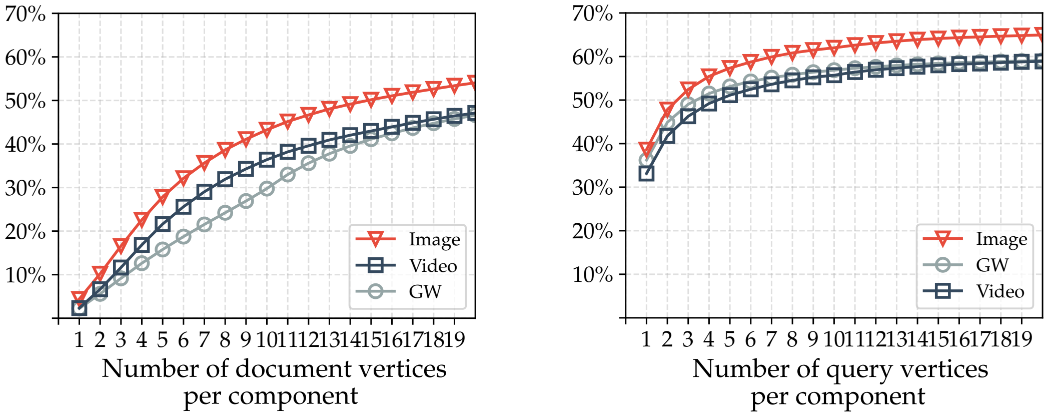


Figure 7.5: Distribution of heterogeneous document vertices across small sub-graphs in our unmodified graph variant. There is a tendency for heterogeneous documents to be distributed across small, disconnected components.

In addition to graph fragmentation, we analyse the distribution of heterogeneous documents (images and videos) across graph components. Figure 7.5 shows the distribution of vertical documents across sub-graphs partitioned by number of document vertices (7.5 left) or number of query vertices (7.5 right). It is interesting to note that, relative to general web documents, image and video vertices tend to be integrated into small sub-graphs (by number of total document vertices) at a higher proportion. Furthermore, roughly 6% of all images and 3% of all video documents are part of single-vertical sub-graphs. In our context, both graph structure and initial textual content — of queries or documents — are integrated into the representation learning mechanism. As such, initialising the propagation algorithm using document terms can be detrimental to documents which are disconnected and have poor initial textual representation. Our findings suggest that not only do heterogeneous documents have poorer initial term representations, but they are also frequently integrated into small, disconnected sub-graphs and together, these factors can amplify some of the limitations that arise in the use of term vector propagation for heterogeneous documents.

We conclude that typical click-graphs are disconnected, with a majority of vertices located in small sub-graphs, whose structure can be detrimental to graph-based representation learning mechanisms. Even more, our analysis suggests that heterogeneous documents are more frequently located in disconnected components and, as such, can develop uninformative representations. Typically, disconnected sub-graphs are pruned when constructing large-scale click graphs, which, as we have shown here, involves discarding the *majority* of information collected in click logs. Therefore, methods for augmenting graph connectivity through linking isolated sub-graphs and increasing within sub-graph local connectivity can potentially increase coverage and quality of representations learnt through term vector propagation in click graphs.

7.4.2 Retrieval Effectiveness

In this section we report on the effectiveness of generating rankings using query and document representations learnt through term propagation in the click-graph. Given the initial textual content and connectivity differences between documents of different types, and how these properties interact in the representation learning mechanism, we set out to understand whether graph learnt representations are effective in direct ranking of documents, across verticals. Throughout our analysis, we distinguish between head and tail queries. Head and tail queries not only occur with different frequencies in our click log, but their corresponding vertices also have different connectivity properties. We note that, in our query sample, head queries originate from the largest sub-graph, whereas tail queries are located within smaller, disconnected components.

7.4.2.1 General Web Rankings

Table 7.1 shows our evaluation of general web rankings derived from different underlying query and document representations. With respect to general web rankings, for head queries, term vector propagation leads to the highest performing rankings in term of precision. Click-through rate performs higher in terms of gain, which is not surprising, given the frequency and number of clicks observed for head queries. For tail queries, however, representations learnt through term propagation generate the highest performing rankings. Although terms used

	Head queries					**Tail queries**				
	P@5	*P@10*	*nDCG@5*	*nDCG@10*	*MAP*	*P@5*	*P@10*	*nDCG@5*	*nDCG@10*	*MAP*
	General web rankings									
CTR	0.840(±0.04)	0.733(±0.06)	**0.749(±0.04)**	**0.638(±0.04)**	**0.765(±0.03)**	0.469(±0.05)	0.259(±0.04)	0.559(±0.04)	0.441(±0.04)	0.394(±0.06)
BM25	0.913(±0.03)	0.880(±0.03)	0.656(±0.04)	0.623(±0.03)	0.727(±0.04)	0.552(±0.06)	0.466(±0.05)	0.512(±0.05)	0.470(±0.04)	0.299(±0.05)
WMD	0.771(±0.06)	0.718(±0.06)	0.604(±0.06)	0.576(±0.05)	0.551(±0.06)	0.538(±0.08)	0.481(±0.07)	0.387(±0.06)	0.372(±0.05)	0.343(±0.07)
Neighbours	0.687(±0.06)	0.643(±0.06)	0.489(±0.05)	0.464(±0.05)	0.476(±0.07)	0.640(±0.06)	0.507(±0.06)	0.641(±0.04)	0.561(±0.03)	0.431(±0.06)
VP_query	0.920(±0.02)	**0.900(±0.02)**	0.636(±0.03)	0.602(±0.03)	0.743(±0.03)	**0.700(±0.06)**	0.553(±0.06)	**0.677(±0.04)**	**0.606(±0.03)**	0.406(±0.06)
VP_doc	**0.933(±0.02)**	0.897(±0.02)	0.545(±0.03)	0.560(±0.03)	0.741(±0.03)	0.693(±0.06)	**0.557(±0.06)**	0.663(±0.04)	0.598(±0.04)	**0.419(±0.06)**
	Image rankings									
CTR	0.612(±0.08)	0.382(±0.06)	0.350(±0.07)	0.259(±0.05)	0.405(±0.08)	0.143(±0.06)	0.071(±0.03)	0.059(±0.03)	0.047(±0.02)	0.071(±0.07)
BM25	0.352(±0.05)	0.341(±0.05)	0.266(±0.05)	0.271(±0.04)	0.145(±0.03)	0.259(±0.05)	0.218(±0.04)	0.153(±0.03)	0.161(±0.03)	0.075(±0.03)
WMD	0.286(±0.05)	0.246(±0.04)	0.205(±0.05)	0.191(±0.04)	0.094(±0.03)	0.231(±0.05)	0.173(±0.04)	0.146(±0.03)	0.142(±0.03)	0.055(±0.02)
Neighbours	0.427(±0.06)	0.390(±0.05)	0.376(±0.06)	0.367(±0.05)	0.208(±0.04)	0.362(±0.06)	0.200(±0.04)	0.217(±0.05)	0.184(±0.05)	**0.251(±0.07)**
VP_query	0.680(±0.04)	0.580(±0.04)	0.714(±0.04)	0.666(±0.03)	0.381(±0.04)	**0.448(±0.07)**	**0.376(±0.06)**	**0.369(±0.06)**	**0.377(±0.06)**	0.231(±0.05)
VP_doc	**0.707(±0.04)**	**0.610(±0.04)**	**0.764(±0.04)**	**0.710(±0.03)**	**0.417(±0.04)**	0.414(±0.05)	0.352(±0.04)	0.351(±0.05)	0.368(±0.05)	0.160(±0.04)
	Aggregated search rankings									
Baseline	0.680(±0.04)	0.717(±0.03)	0.349(±0.06)	0.287(±0.06)	0.489(±0.03)	0.517(±0.04)	0.438(±0.04)	0.252(±0.06)	0.223(±0.06)	0.261(±0.04)
WMD	0.667(±0.04)	0.737(±0.03)	0.396(±0.06)	0.338(±0.06)	0.517(±0.04)	0.503(±0.05)	0.434(±0.04)	0.407(±0.05)	0.310(±0.05)	0.257(±0.04)
Neighbours	0.720(±0.04)	0.737(±0.03)	0.432(±0.06)	0.310(±0.06)	0.521(±0.04)	0.531(±0.05)	0.448(±0.04)	0.285(±0.06)	0.250(±0.06)	0.271(±0.04)
VP_query	**0.767(±0.04)**	0.763(±0.03)	**0.520(±0.05)**	0.457(±0.05)	0.560(±0.03)	**0.538(±0.05)**	**0.452(±0.04)**	**0.460(±0.05)**	**0.386(±0.05)**	0.267(±0.04)
VP_doc	**0.767(±0.04)**	**0.783(±0.03)**	0.518(±0.05)	**0.467(±0.05)**	**0.585(±0.04)**	**0.538(±0.05)**	0.448(±0.04)	0.446(±0.05)	0.324(±0.06)	**0.277(±0.04)**

Table 7.1: Evaluation of direct ranking based on graph-learnt representations (mean and standard error). For each ranking type, maximum value in each column is highlighted.

to initialise the propagation algorithm (i.e., query or document terms) affect retrieval performance, these differences are not consistent across evaluation metrics and query types, which suggests that in the case of documents with rich textual content, such as general web documents, the feature space in which representations are learnt has less impact on the quality of derived rankings. Overall, our results are consistent with Jiang et al. (2016), where graph-based representations outperform similar baselines in direct ranking.

Ranking methods that leverage graph structure alone by computing vertex neighbourhood overlap, on average perform under our baselines for head queries, but above baselines in the case of tail queries. In particular, we note that, for tail queries, ranking based on graph neighbourhood (*Neighbours*) is the highest performing ranking method that does not make use of the term propagation algorithms, across metrics. This suggests that, even though disconnected, small components typically incorporate closely related documents. Given that representation learning is computationally expensive, our results show that approximate methods that leverage only graph structure can be used effectively, in the case of tail queries, for ranking general web documents.

7.4.2.2 Image Rankings

As we have shown, images not only have poor initial textual representation, but are also more frequently integrated into small graph components, and together, these aspects can be detrimental to graph-based representation learning and associated image rankings.

Table 7.1 shows our evaluation of image rankings derived from the vector propagation algorithm. Firstly, we point out that graph-based representations consistently lead to the highest performing rankings, across head and tail queries. Our work is, therefore, a first confirmation that graph-learnt representations can be effectively used in the retrieval of heterogeneous documents. Secondly, we point out that using query terms rather than document terms to initialise the propagation algorithm is more effective in the case of tail queries. This can be explained by the fact that image documents have poorer initial term representations (e.g., *"celebrity images"*) and query terms (e.g., *"taylor swift images"*) can be beneficial in disambiguating their content.

7.4.2.3 Aggregate Search Rankings

Graph-based representation learning bridges the query-document vocabulary mismatch, and also the representation gap across verticals of heterogeneous content.

Learning representations for different types of documents (e.g., images and Wikipedia pages) in the same feature space enables the assessment of cross-vertical document similarity. As such, we investigate whether graph-learnt representations can be used to construct aggregated search pages by leveraging cross-vertical document similarity scoring.

We use our BM25 rankings as the basis from which to construct aggregated search pages. This is achieved by inserting image blocks from an image ranking into a general web ranking. Given the BM25 general web and image rankings, we use document representations learnt through term propagation to assess the similarity of images to the top-K documents in the general web ranking, as measured by cosine similarity between associated vectors. Specifically, given a set s of n general web documents, we merge and normalise their learnt term vectors, and then select images that are most similar, based on cosine similarity, to the merged vector. We select the top-M most similar images and insert them into the general web ranking, at rank three. For our baseline, the image block contains the top-M highest ranking images in the BM25 image ranking (in our evaluation, we arbitrarily choose $K = M = 3$).

Table 7.1 shows our results across head and tail queries. Using representations learnt through term propagation to aggregate search results leads to the highest performing rankings. We conclude that graph based representations can be used to construct more relevant cross-vertical aggregates – and in the absence of graph based representations, graph structure and document neighbourhood can be used effectively as well.

7.4.3 Modifying Graph Structure

Our approaches to modifying click-graph structure involve creating additional connections between query vertices, with the intent of learning more informative vectors. Given that the majority of documents (and queries) in our log are located in small graph components, we hypothesise that constructing additional connections between sub-graphs is an effective way of improving representation learning through vector propagation. We report our analysis of retrieval effectiveness under different graph configurations in the following section, but first report how graph augmentations modify graph structure.

7.4.3.1 Changes to Graph Connectivity

To modify graph structure, we increase the number of edges in the graph by linking query vertices based on their lexical similarity (*text augmented graph*) or

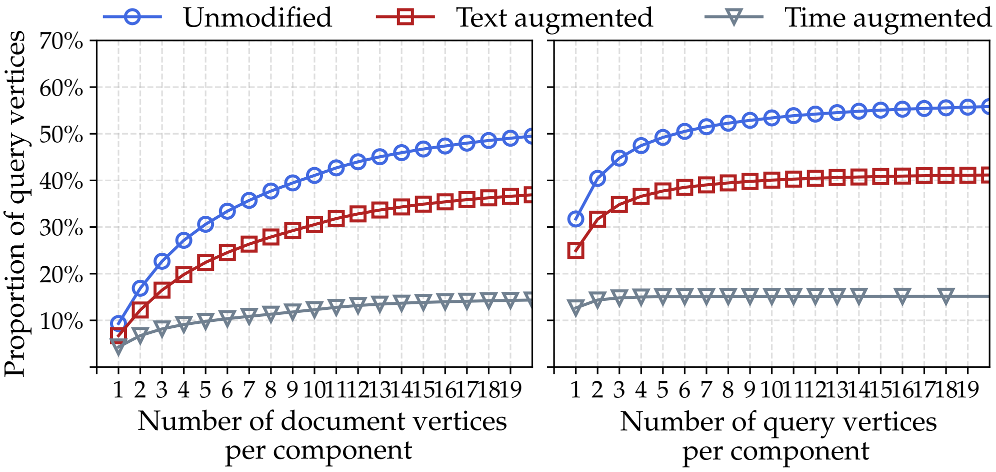


Figure 7.6: Distribution of query vertices across sub-graphs in different configurations of the click graph (*original*, *text augmented* or *time augmented*).

temporal proximity (*time augmented graph*). Figure 7.6 shows the distribution of query vertices across small components in the graph variants we explored. In the case of our text augmented graph, we increase the number of edges in our graph by 51%, thus connecting 66% of query and 68% of document vertices within one graph component. Reducing the proportion of documents within single-query components from 38% to 29%. In the case of our time augmented graph, we increase the number of edges in our graph by 79%, connecting 82% of queries and 85% of documents within one component.

We now analyse how effective modifying graph structure is in generating more informative representations for queries and documents, and how this is reflected in ranking quality. We note that both our graph modifications (*text* or *time* based) integrate our sampled tail queries within the largest sub-graph of their respective graph variants.

7.4.3.2 Retrieval Effectiveness

Table 7.2 shows our evaluation of rankings derived from representation learnt in graphs with varying structure. We compare these rankings to their equivalents derived from representations learnt in the unmodified click-graph (e.g., *Text_VP_query* to *VP_query*). Our results suggest that augmentations of the click-graph can lead to improvement in general web ranking performance, however this improvement is dependent on the textual content used to initialise the representation learning algorithm. In the case of initialising the algorithm with query terms, both types of graph augmentation (text-based and time-based) deteriorate performance with respect to the unmodified graph, as reflected in our evaluation

	Head queries					**Tail queries**				
	P@5	*P@10*	*nDCG@5*	*nDCG@10*	*MAP*	*P@5*	*P@10*	*nDCG@5*	*nDCG@10*	*MAP*
					General web rankings					
VP_query	0.920(±0.02)	0.900(±0.02)	0.636(±0.03)	0.602(±0.03)	0.743(±0.03)	0.700(±0.06)	0.553(±0.06)	0.677(±0.04)	0.606(±0.03)	0.406(±0.06)
Text_VP_query	0.760(±0.05)	0.810(±0.04)	0.482(±0.04)	0.510(±0.04)	0.608(±0.05)	0.693(±0.06)	0.553(±0.06)	0.661(±0.05)	0.594(±0.04)	0.404(±0.06)
Time_VP_query	0.540(±0.08)	0.530(±0.07)	0.370(±0.06)	0.372(±0.06)	0.401(±0.07)	0.593(±0.07)	0.473(±0.06)	0.542(±0.06)	0.481(±0.05)	0.327(±0.06)
VP_doc	0.933(±0.02)	0.897(±0.02)	0.545(±0.03)	0.560(±0.03)	0.741(±0.03)	0.693(±0.06)	0.557(±0.06)	0.663(±0.04)	0.598(±0.04)	0.419(±0.06)
Text_VP_doc	**0.967(±0.02)**	0.953(±0.01)	0.666(±0.03)	0.673(±0.03)	**0.831(±0.02)**	**0.713(±0.06)**	**0.603(±0.06)**	**0.691(±0.04)**	**0.628(±0.03)**	**0.450(±0.06)**
Time_VP_doc	0.953(±0.02)	**0.957(±0.01)**	**0.699(±0.03)**	**0.700(±0.02)**	0.830(±0.02)	0.667(±0.06)	0.560(±0.06)	0.671(±0.04)	0.607(±0.03)	0.407(±0.06)
					Image rankings					
VP_query	0.680(±0.04)	0.580(±0.04)	0.714(±0.04)	0.666(±0.03)	0.381(±0.04)	**0.448(±0.07)**	**0.376(±0.06)**	**0.369(±0.06)**	**0.377(±0.06)**	**0.231(±0.05)**
Text_VP_query	0.693(±0.05)	0.580(±0.05)	0.720(±0.04)	0.671(±0.03)	0.398(±0.05)	0.393(±0.05)	0.352(±0.04)	0.341(±0.05)	0.363(±0.04)	0.161(±0.04)
Time_VP_query	0.453(±0.06)	0.373(±0.05)	0.373(±0.06)	0.354(±0.05)	0.202(±0.05)	0.367(±0.05)	0.310(±0.04)	0.328(±0.05)	0.343(±0.05)	0.146(±0.03)
VP_doc	**0.707(±0.04)**	**0.610(±0.04)**	**0.764(±0.04)**	**0.710(±0.03)**	**0.417(±0.04)**	0.414(±0.05)	0.352(±0.04)	0.351(±0.05)	0.368(±0.05)	0.160(±0.04)
Text_VP_doc	0.647(±0.04)	0.590(±0.05)	0.659(±0.04)	0.639(±0.03)	0.376(±0.05)	0.386(±0.06)	0.310(±0.05)	0.331(±0.06)	0.327(±0.05)	0.165(±0.04)
Time_VP_doc	0.653(±0.05)	0.553(±0.05)	0.661(±0.05)	0.602(±0.04)	0.371(±0.05)	0.380(±0.06)	0.287(±0.05)	0.329(±0.06)	0.309(±0.05)	0.157(±0.04)
					Aggregated search rankings					
VP_query	**0.767(±0.04)**	0.763(±0.03)	0.520(±0.05)	0.457(±0.05)	0.560(±0.03)	**0.538(±0.05)**	0.452(±0.04)	0.460(±0.05)	**0.386(±0.05)**	0.267(±0.04)
Text_VP_query	0.733(±0.04)	0.750(±0.03)	**0.525(±0.05)**	0.429(±0.05)	0.537(±0.03)	**0.538(±0.05)**	0.452(±0.04)	0.433(±0.06)	0.378(±0.05)	0.271(±0.04)
Time_VP_query	0.747(±0.04)	0.763(±0.03)	0.509(±0.05)	**0.475(±0.05)**	0.559(±0.04)	**0.538(±0.05)**	**0.459(±0.04)**	0.420(±0.06)	0.368(±0.05)	0.277(±0.04)
VP_doc	**0.767(±0.04)**	**0.783(±0.03)**	0.518(±0.05)	0.467(±0.05)	**0.585(±0.04)**	**0.538(±0.05)**	0.448(±0.04)	0.446(±0.05)	0.324(±0.06)	0.277(±0.04)
Text_VP_doc	0.753(±0.04)	0.760(±0.03)	0.479(±0.06)	0.404(±0.06)	0.554(±0.03)	**0.538(±0.05)**	0.452(±0.04)	**0.465(±0.05)**	0.333(±0.06)	0.274(±0.04)
Time_VP_doc	0.720(±0.04)	0.747(±0.03)	0.448(±0.06)	0.422(±0.05)	0.534(±0.03)	0.531(±0.05)	0.452(±0.04)	0.405(±0.06)	0.331(±0.06)	**0.279(±0.04)**

Table 7.2: Evaluation of changes to graph structure and their impact on retrieval performance (mean and standard error). For each ranking type, maximum value in each column is highlighted.

metrics. On the other hand, graph augmentations increase retrieval performance when initialising the algorithm using document terms, the best performing general web rankings, overall, being derived from representations learnt from document terms in a text-augmented graph (*Text_VP_doc*). The interactions between terms used to initialise the propagation algorithm and structural changes to the click-graph are more prevalent in the case of head queries and are perhaps best explained through an example.

Table 7.3 shows the top ten terms and their associated weights learnt under different algorithm and graph configurations for the head query *"taylor swift"*. Initialising the propagation algorithm using query terms, it is not surprising to notice that the terms *"taylor"* and *"swift"* are the top weighted terms across graph configurations. However, the effects of modifying graph structure are apparent in representations learnt under the text and time augmented graphs. In the case of the text-augmented graph, top weighted terms derived from lexically related queries (e.g., *"talor"*, *"tayler"*) are included by the additional query-to-query edges. Similarly, in the case of our time-augmented graph, terms derived from temporally related queries are included (e.g., *"lady"*, *"gaga"*). This example illustrates how additional edges modify the content of learnt vector representations when initialising the algorithm using query terms, and shows that potentially unrelated terms are introduced into the query representations when modifying graph structure, which in turn leads to the drop in performance when initialising the algorithm using query terms. We also note that term weights are more uniformly distributed in the case of augmented graph representations than in the case of the unmodified graph, which in turn gives higher discrimination power to terms that are perhaps unrelated to the intended query.

Initialising the propagation algorithm using document terms, however, benefits from changes to graph structure. It is apparent from our example that representations learnt under different graph configurations are very similar with respect to their content, and share a high proportion of their top-weighted terms. Given that representation learnt in the text-augmented click-graph overall produce the highest performing general web rankings, our results suggest that additional edges, in coordination with rich initial term representation, are useful in term re-weighting and can help learn higher weights for relevant terms.

In the case of image rankings, our results suggest that modifying graph structure is less effective and leads to poorer retrieval performance. In contrast to general web rankings, where we have shown that augmenting graph structure, in combination with using document terms in the propagation mechanism, can be beneficial for retrieval, in the case of image rankings, poor initial textual repres-

Query terms	***Unmodified***	swift	0.4923	taylor	0.4915	instagram	0.0038	shake	0.0025	twitter	0.0016	2016	0.0007	account	0.0007	songs	0.0006	album	0.0005
	Text	swift	0.4149	talor	0.1602	tayler	0.1602	taylorswift	0.1602	taylor	0.0956	instagram	0.0019	shake	0.0018	hill	0.0011	twitter	0.0005
	Time	taylor	0.0487	swift	0.0475	bikini	0.0402	jenner	0.0332	coats	0.025	kendall	0.025	conway	0.0236	gaga	0.0235	lady	0.021
Doc terms	***Unmodified***	taylor	0.2271	swift	0.2262	images	0.1757	results	0.0925	image	0.0875	concert	0.0263	video	0.0112	videos	0.0106	rush	0.0075
	Text	taylor	0.2248	swift	0.2239	images	0.1731	results	0.0912	image	0.0862	concert	0.0265	video	0.0111	videos	0.0104	rush	0.0074
	Time	images	0.1637	taylor	0.1444	swift	0.1437	results	0.0843	image	0.0795	concert	0.0236	video	0.0134	videos	0.0089	official	0.008

Table 7.3: Top 9 terms learnt under different algorithm and graph configurations for the head query *"taylor swift"*.

entations and perhaps noisy additional connections (as introduced by our graph manipulations) deteriorate the quality of learnt term representations. Similarly, with respect to creating cross-vertical aggregates, our results suggest that improvements are less clear when modifying graph structure, and are highly dependent on how the propagation algorithm is initialised.

7.5 Chapter Summary

One of the main challenges in constructing composite objects is the disparity of term distributions across verticals (Santos et al., 2011), which in turn makes similarity assessment for heterogeneous documents problematic. The work presented in this chapter tackles this problem by leveraging a click graph structure to learn representations for heterogeneous items in a unified feature space.

The experiment in this chapter firstly provides an in-depth description of a typical click-graph with respect to its constituent components and the distribution of queries and documents of different types (i.e., from different verticals) across sub-graphs. Our results show that typical click-graphs are disconnected, with a majority of query and document vertices incorporated into small graph components. This finding has implications for graph-based representation learning algorithms applied in heterogeneous environments, as there is tendency for heterogeneous documents to be located in isolated, vertically uniform sub-graphs. Informed by our analysis of graph structure, we evaluate the use of a recently proposed representation learning algorithm in the retrieval and aggregation of documents of different types. Our results show that graph-learnt representations can be used effectively for the retrieval of relevant documents, across verticals. In addition, our results show that learnt representations can also be used to construct relevant, cross-vertical document aggregates.

Secondly, to address some of the structural bias related to document distribution over constituent sub-graphs, we explore ways of manipulating the click-graph to enhance the discriminative power of feature vectors learnt through term propagation. We introduce pseudo-vertices in the click-graph to increase local connectivity of graph components and link disconnected sub-graphs through related query vertices. We show that improving retrieval performance is possible through manipulations of click graph structure, but improvement is dependent on how the propagation algorithm is configured and initialised. Structural manipulation of click graphs aimed at improving retrieval performance remains an under-explored application area and we provide a first perspective in this space.

Although due to limitations in gathering relevance labels for a web-scale document collection we are unable to fully apply the result composition framework presented in chapter 6 to the graph-learnt feature space, our experiments constructing cross-vertical document aggregates using only two verticals (i.e., general web and images) show that using a unified representation space can lead to an increase in performance, with respect to laboratory-based evaluation metrics.

Chapter 8

Discussion and Future Work

In this chapter, we provide an in-depth discussion regarding the research outcomes of our work. In addition, we discuss broad directions for future work in the space of result composition. We begin this chapter by framing the research findings discussed throughout previous chapters within the system-centric result composition process introduced in chapter 1.

8.1 Discussion

Search engines are complex software architectures which typically integrate many different specialised components, from crawlers to search interfaces. Croft et al. (2009) identify two major high-level functions that all search engines support: the *indexing* process and the *query* process. In their definition, the indexing process builds the structures that enable searching (i.e., the search index) and the query process uses those structures together with a user query to produce a results page.

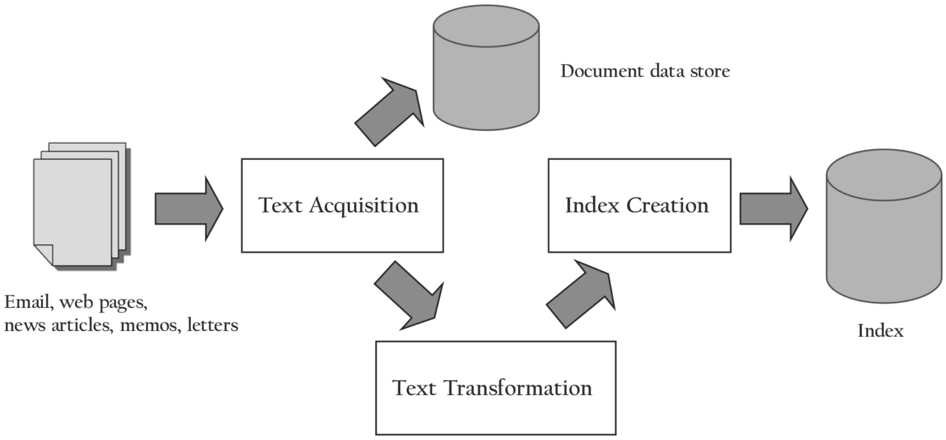


Figure 8.1: High-level representation of a search engine's indexing process, reproduced from Croft et al. (2009, p. 15).

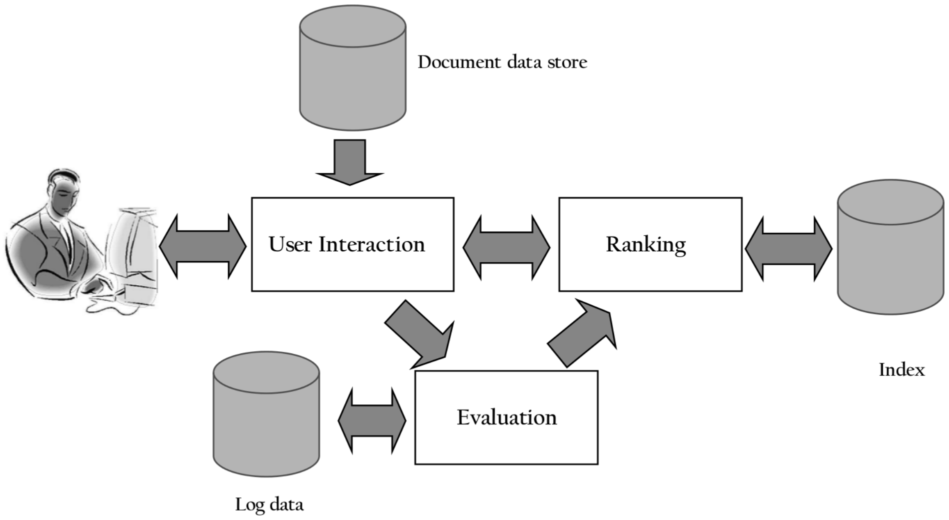


Figure 8.2: High-level representation of a search engine's query process, reproduced from Croft et al. (2009, p. 16).

Figure 8.1 shows the high-level components of the indexing process (i.e., text acquisition, text transformation, and index creation) whereas figure 8.2 shows the components of the query process (i.e., user interaction, ranking, evaluation). This representation of a search engine's software architecture is not a code-level description of its functionality, but rather a high-level representation that enables researchers and engineers to communicate about its structure.

A search engine employing result composition techniques would have to integrate these techniques within its indexing and querying processes. As such, the following section discusses how our research outcomes can be integrated within a search engine's high-level structure and processes, and makes specific recommendations for the implementation of such a search system.

The composite search engine. In chapter 1, we briefly introduced a high-level representation of a composite search engine. Figure 8.3 enhances this representation, by illustrating some of the challenges associated with each component of such a search system, and by highlighting the challenges addressed throughout this thesis. Given the vast complexity of search, many challenges remain unaddressed by our work, and these challenges are discussed later in this chapter.

Following the system representation shown in figure 8.3, as part of the indexing process of search, our work addresses the task of creating a common feature space for heterogeneous documents. Specifically, chapter 7 uses both text and click information to generate a feature space in which the cross-vertical representation gap is bridged. A search system that employs composite objects needs to represent results within such a space to determine the similarity of items across

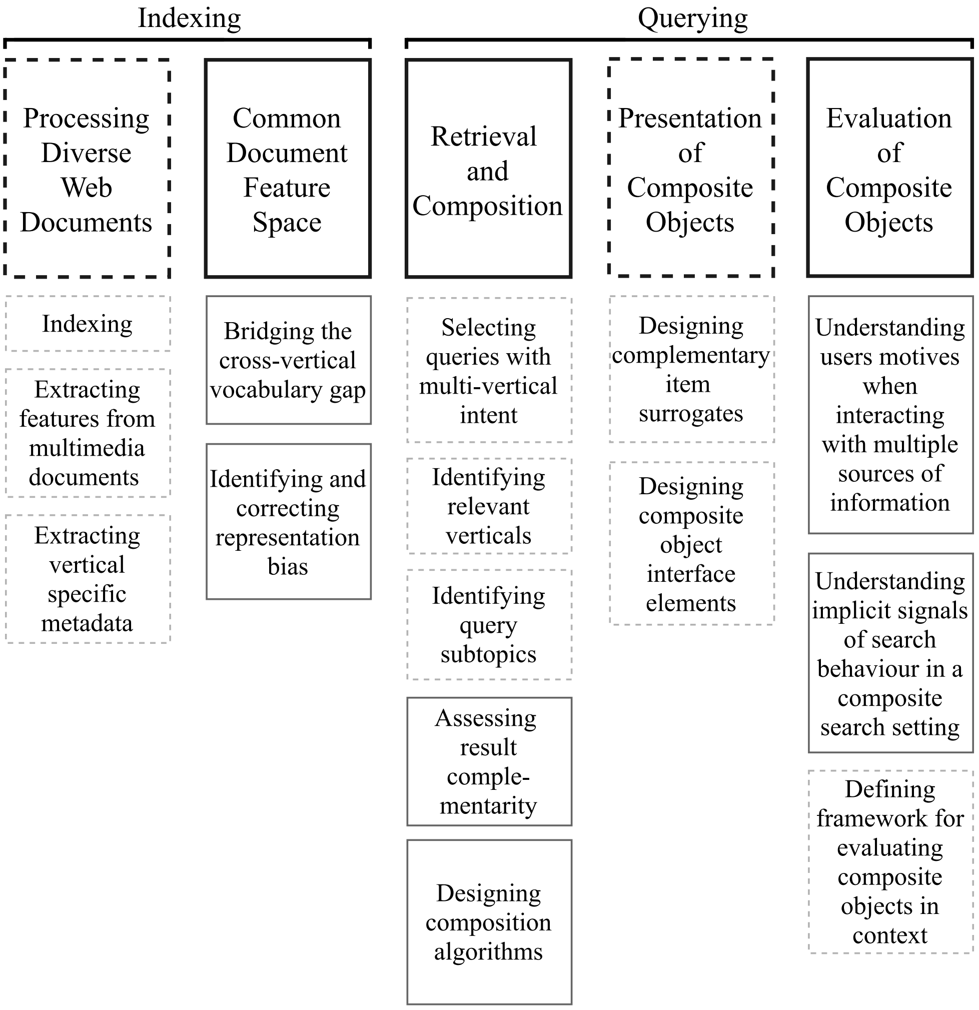

Indexing
Querying
Processing Diverse Web Documents
Common Document Feature Space
Retrieval and Composition
Presentation of Composite Objects
Evaluation of Composite Objects
Indexing
Extracting features from multimedia documents
Extracting vertical specific metadata
Bridging the cross-vertical vocabulary gap
Identifying and correcting representation bias
Selecting queries with multi-vertical intent
Identifying relevant verticals
Identifying query subtopics
Assessing result comple-mentarity
Designing composition algorithms
Designing complementary item surrogates
Designing composite object interface elements
Understanding users motives when interacting with multiple sources of information
Understanding implicit signals of search behaviour in a composite search setting
Defining framework for evaluating composite objects in context

Figure 8.3: High-level representation of a composite search engine. Top row illustrates the components of a composite search engine, whereas a subset of challenges associated with each component are illustrated below. Components and associated challenges that are explicitly addressed through work presented in this thesis are presented within a solid contour.

verticals, but also to enable assessing object properties (e.g., its overall coherence). Our work provides search practitioners with a foundation on which to develop a unified document representation space, but also highlights the many challenges of creating such a feature space using click-graphs (e.g., disconnected graph structure) and studies the use of simple techniques aimed at overcoming these challenges.

Much of the work presented in this thesis can be conceptually placed within the querying process of a search engine. Integrating the research outcomes presented in chapters 6 and 7 within a unified (conceptual) search engine, the retrieval and composition algorithms presented in chapter 6 can be enhanced by a common representation space for heterogeneous documents, generated using the techniques discussed in chapter 7. Although in our proposed algorithmic framework for constructing composite objects, the document representation space is derived from document textual data only, enhancing document representation by using both text and click information, as described in chapter 7, can not only effectively bridge the cross-vertical vocabulary gap, but also potentially improve algorithmic result composition (as shown briefly in section 7.4.2.3). Together, findings from chapters 6 and 7 provide the basis on which search practitioners can develop effective composite search systems.

The research findings supported by chapters 4 and 5 can be conceptually placed within the querying process of a search engine, specifically within the evaluation component of the querying process. The task of search engine evaluation is to measure and monitor effectiveness and efficiency (Croft et al., 2009). However, in the case of composite objects, it is difficult in a practical setting to decide what to measure (i.e., what metrics to monitor) and also unclear how to interpret measurements (e.g., know whether a measurement indicates user satisfaction or not) given that composite objects are novel search elements — challenges that are present in other areas of evaluating modern web search, such as aggregated search (Arguello, 2017; Zhou et al., 2012). As such, our work in chapter 4 studies the way searchers construct composite objects manually, and also how they explicitly assess their properties. Our findings are informative for algorithm design (e.g., result composition algorithms can aim to replicate human-designed composite objects) and also for developing a comprehensive evaluation framework for heterogeneous information access in modern web search. In particular, our results suggest a complex interplay between composite object properties, showing that topical relevance is not the only important characteristic to consider when designing evaluation measures for a composite search engine.

Chapter 5 provides additional support for the claim that composite object usefulness is determined by a complex interplay of object properties, and the research findings described in chapter 5 are informative for the development of evaluation methodologies for complex result aggregates that are based on implicit search behaviour signals. In practical terms, we show how mouse movements and clicks can be used as implicit signals of perceived task effort in a composite search setting. We also show that object relevance and diversity have stronger effects than object coherence on users' search behaviour. Overall, chapters 4 and 5 are a solid foundation for the future development of evaluation frameworks for composite results.

One of the core components of a search engine is its user-facing interface. How results — whether composite or traditional — are displayed on the results page has broad impact on users' perception of the search system overall. The presentation of composite objects is a crucial aspect of developing composite search engines, but is, perhaps, an aspect that is understudied in our work. As such, we consider it further in the following section, where we discuss areas for the future development of composite web search.

8.2 Future Work

The fast pace with which technology advances means that much of the research behind its advance becomes obsolete from one day to the next. The study of web search is no exception — neither is the work presented in this thesis. As such, in this section, we present several directions for the future study of result composition, focusing on broad aspects of heterogeneous information access that we believe will remain of consequence even as web search technologies advance.

The display of composite objects. One key aspect of result composition that is, perhaps, understudied in this thesis relates to the display and presentation of composite objects on a unified results page.

Typically, items within a web result ranking consist of document title and brief summary of the document (i.e., the result *snippet* or *abstract*). This representation of a document is sometimes referred to as the *document surrogate*. Prior work has focused extensively on how features of documents surrogates affect searchers. Query-biased document summaries (also known as *keyword-in-context*) have been shown to improve click-through rate (Clarke et al., 2007), improve performance in terms of precision, recall and time taken to find relevant information (Tombros and Sanderson, 1998; White et al., 2003). With respect to snippet length, Cutrell and Guan (2007) show that longer snippets (6-7 lines) improve performance for information tasks, but degrade performance for navigational task, a finding further supported by Kaisser et al. (2008). Furthermore, Yue et al. (2010) found a click-through bias in favour of textual snippets that simply displayed bolded versus non-bolded query-terms. More related to the work explored in this thesis, prior work suggests that, for example, surrogates augmented with images pulled from the underlying document can help users make more accurate and faster relevance judgements (Capra et al., 2013; Teevan et al., 2009).

In the case of aggregated search, different types of documents are typically associated with different surrogate representations (i.e., images are usually displayed in a grid of thumbnails, videos can allow auto-play directly on the results page, tweets can be shown with author and full text information, etc.). Arguello (2017) states that "[f]or aesthetic reasons and to better convey how the vertical may have relevant content [...] vertical results are typically grouped together (either stacked horizontally or vertically)" and that "the goal of vertical presentation is to display the most relevant verticals in a more salient way". Indeed, prior studies have shown that searchers tend to click on results that are more visually salient (Sushmita, Joho, Lalmas and Villa, 2010). Overall, Arguello (2017) suggest a complex interplay between the relevance, position, and presentation of aggregated search results, all three elements influencing user engagement with the aggregated results, measured in terms of either clicks or eye fixations.

Prior efforts on result presentation listed above illustrate how various aspects of not only what information is presented on the results page, but how it is presented affects searchers' impressions of what is relevant and their effectiveness in accomplishing search-mediate tasks. So far, limited effort has been dedicated to understanding effective document surrogates for heterogeneous content. It is possible that, as in the case of traditional results, creating query-biased surrogates for diverse content (e.g., highlighting query terms in tweets) can help users better navigate different types of content. Moving beyond single item surrogates, merging results from various different sources, each associated with different types of surrogates, within composite objects that are useful requires an investigation of complementary surrogates that highlight various connections between the items within the object, the query and the overall page of results.

Entity cards are, when discussing results presentation, a useful visual metaphor (and an instance) of composite objects. Even though adopted by most popular web search engines and used widely, to date, limited effort has been dedicated to understanding presentation factors that influence their usefulness. Our work presented in chapter 5 is a first step in this direction, as existing prior work focuses on the effects these interface elements have on gaze patterns or search interactions (Lagun et al., 2014; Navalpakkam et al., 2013) or the algorithmic assembly of these objects (Hasibi et al., 2017). Furthermore entity cards (and other novel search interface elements, such as entity carousels or enhanced aggregated search blocks that allow searchers to scroll through content) have recently become interactive, allowing users to expand various elements of the card or execute actions (e.g., play music) directly on the search interface. Given entity card popularity in modern web search, understanding not only how to assemble information within such complex objects but also, now, how to effectively enable user (transactional) interactions with such objects is likely a high impact direction for future research.

Composite objects for different types of user needs. The work presented in this thesis makes the assumption that composite results can be returned in response to ambiguous (Sanderson, 2008), multi-faceted (Kong and Allan, 2013) or task-oriented queries that have vertical intent (i.e., queries for which multiple sources of information are relevant). Indeed, in our work, we use the same queries (and test collections, in chapter 6) as previous work on aggregated search, without exploring what types of queries might benefit from the presence of composite objects on their corresponding result pages. Even though the display of entity cards has become very frequent (Bota et al., 2016; Enge, 2017), and a majority of queries exhibit some form of vertical intent (Arguello et al., 2009), it is likely that not all searches benefit from the display of composite objects on the results page. A closer look at how query-intent (Broder, 2002; Rose and Levinson, 2004) interacts with the usefulness of composite objects is a necessary next step in the study of result composition.

In addition to query-intent, users typically engage with search systems to satisfy a wide range of possible information needs (Ingwersen and Järvelin, 2005, provide a comprehensive review of information seeking and retrieval needs). The studies presented in chapters 4 and 5 simulate simple and well-specified search scenarios, without investigating different types of information needs or the role task complexity plays in either the manual composition of results (chapter 4) or in the effect entity cards have on search behaviour (chapter 5). With regard to task complexity, prior work studying information seeking behaviour has found that more complex tasks are typically associated with greater levels of search interaction, as determined by a greater number of queries, clicks, bookmarks, longer dwell-times and task completion times (Aula, Khan and Guan, 2010; Liu, Cole, Liu, Bierig, Gwizdka, Belkin, Zhang and Zhang, 2010; Liu et al., 2012; Liu, Liu, Gwizdka and Belkin, 2010; Wu et al., 2012). In addition, searchers tend to engage with results originating from a wider range of sources when performing more complex tasks (Jansen et al., 2009). Given this broad body of work indicating interactions between information need, task complexity and search behaviour, it is likely that composite result usefulness is affected by features regarding searchers' information need and the complexity of their underlying, search-mediated task.

Finally, with respect to aggregated search, Turpin et al. (2016) studied the effect of perceptual speed — the *"speed in comparing figures and symbols [...] or carrying out other simple tasks involving visual perception"* (Ekstrom et al., 1979) — on search performance and user behaviour, finding that users with low perceptual speed took longer to complete their tasks when using an aggregated search interface (i.e., an interface containing merged blocks of heterogeneous results). This suggests that, much like aggregated search, composite objects (e.g., entity cards) are not a "one size fits all" solution, and different user abilities play a strong role in determining their usefulness. Understanding the interaction between information need, task complexity, user abilities and the usefulness of composite results in web search remains an understudied, yet essential aspect of developing modern web search interfaces.

The ethics of result composition. Although a topic not discussed widely in this thesis, the content search engines display on the results page (and the way they choose to display it) has a wider range of consequences on people's perception of information, beyond immediate search behaviour or clicks on ads. Pariser (2011) and O'Neil (2016) discuss how algorithmic decisions have detrimental effect on society either through explicit discrimination (e.g., voluntary or involuntary biases encoded within algorithms that discriminate certain communities) or excessive personalisation (e.g., search results only reflecting personal beliefs, to the extent that searchers become unaware of other perspectives on the same issue). Indeed, many aspects regarding the ethics of algorithms (Mittelstadt et al., 2016), or the ethics of web search (Tavani, 2016), are becoming central in the public discourse around our use of information technology — and have been considered, in various forms, over at least half a century (Wiener, 1950). It is, then, important to consider wider implications of aggregating content from multiple sources within information objects that satisfy searchers' needs directly on the results page.

In a recent article, Ford and Graham (2016) discuss how the digital representation of entities — in this case, urban communities or cities — affects their wider perception. They focus specifically on the rise of the *"semantic web"* (Berners-Lee et al., 2001) and the use of entity cards by search engines as mechanisms through which the public perception of entities is shaped. They discuss an example (i.e., the entity card for the city of Jerusalem, displayed in response to the query *"jerusalem"* on Google's results page) illustrating how in the process of linking data across databases and when assembling entity cards, elements of their underlying data representation become skewed or obscured. Specifically, they show how many of the nuanced and complex political issues regarding the city of Jerusalem (e.g., Jerusalem being the capital city of Israel and, at the same time, the claimed capital city of Palestine) that are discussed on Wikipedia, one of the sources from which entity card information is aggregated, are lost or abruptly settled when presenting these cards on the results page (e.g., the entity card for the query *"jerusalem"* displays the title *"Jerusalem, capital of Israel"*, regardless of geographic personalisation of results). Secondly, they show how the provenance of data underlying the entity card is obscured through aggregation, and how the factoids shown in the entity card can be misrepresenting. In their example, one of the encyclopaedic facts shown on the entity card for Jerusalem (i.e., the population of the city), sourced from UN Data[1], displays the population for West Jerusalem (i.e., the capital of Israel), rather than that of the whole city, excluding the population of East Jerusalem (i.e., the claimed capital of Palestine). The authors also trace the data trail from UN Data to Google's entity card, showing how complex the process of finding where entity card data actually comes from is, as data sources are not indicated on the cards themselves. Finally, they discuss how the agency of users affected by information displayed in entity cards is diminished by these novel displays. The authors discuss how, on Wikipedia, factual statements about controversial topics are debated and corrected by a community of editors and can be discussed, directly, by the people affected by misrepresentation. Entity cards make use of data collected from Wikipedia, through a linking database called WikiData, data that is finally cached in Google's own Knowledge Graph. Challenging or even discussing misrepresentations within entity cards, then, becomes difficult or impossible, as there is no support for that type of community interaction around entity cards and, furthermore, changes to information on Wikipedia do not necessarily propagate (or propagate slowly) through the data pipeline to Google's entity cards. Even though entity cards typically display a very subtle *"Provide feedback"* button on the card, no information is available about what happens to users' feedback once they provide it and the feedback given by searchers is not displayed publicly; it is unclear whether searcher feedback has any effect at all. In addition to the politically controversial entity of Jerusalem, public reports have discussed similar issues regarding celebrities and other entities (Dewey, 2016; Mathews, 2015).

Of course, the fact that information on the web is often incorrect or voluntarily misleading is not a novelty. The problem with entity cards, however, is that, being such prominent elements of the results page, integrated within the somewhat trusted gateway to the entire web (i.e., Google's results page), is that the information they contain can be accepted as an unequivocal truth, "as unsourced and absolute as if handed down by God" (Dewey, 2016), which ultimately "undermines people's ability to verify information and [...] develop well-informed opinions" (Dewey, 2016).

To sum up, the problems of complex information objects that aggregate data from multiple sources and are directed at satisfying information needs directly on

[1] http://data.un.org/

the results page (e.g., entity cards) are that ***(i)*** they remove nuance and context, abruptly settling complex issues, ***(ii)*** they obscure the provenance of the data they contain, and ***(iii)*** diminish people's ability to challenge misrepresentations that affect them directly. Addressing these problems is an important future direction for the study of modern web search.

Underlying discussions around the semantic web, linked data, and much of information technology research are narratives (implied or explicit) about apolitical and purely technical processes of structuring information. However, the example of Jerusalem's entity card on Google shows how even apolitical decisions about aggregating information across various sources of data has deep political implications. These implications exist whether or not they are discussed by researchers and developers. Achieving a better understanding of these implications in relation to the display of composite objects on the results page is a salient matter for future research.

8.3 Chapter Summary

In this chapter we discuss the conceptual structure of a search engine, and how the research findings presented throughout this thesis can be integrated within such a structure. In addition, we consider broad directions for future work in the space of result composition. The following chapter ends this thesis by summarising the key contribution and conclusions derived from our work.

Chapter 9

Conclusion

This thesis investigates the aggregation of web search results retrieved from various document sources (e.g., images, tweets, Wiki pages) within information *"objects"* to be integrated in the results page assembled in response to user queries. We use the terms *"composite objects"* or *"composite results"* to refer to such objects, and overall use the terminology of Composite Web Search (i.e., *composite objects*, *result composition*) to distinguish our approach from other methods of aggregating heterogeneous content within a unified results page (i.e., Aggregated Web Search). In our definition, the aspects that differentiate composite objects from aggregated search blocks are that composite objects ***(i)*** contain results from multiple sources of information, ***(ii)*** are specific to a common topic or facet of a topic, and ***(iii)*** are not a uniform ranking of homogeneous results ordered only by their topical relevance to a query.

Modern web search engines now deploy a variety of such "non-traditional" elements on the results page, many of which contain information from multiple source. Assembled around the long-established ranked list of blue links — and more recent aggregated search blocks — are rich format ads (Lagun et al., 2016), in-line answers (Chilton and Teevan, 2011) or entity cards (Bota et al., 2016; Hasibi et al., 2017; Navalpakkam et al., 2013). As these information objects become more and more prevalent, understanding their role, properties and influence on searchers is an essential aspect of modern web search science.

The work presented throughout this thesis attempts the task of studying composite objects by exploring users' perspectives on accessing and aggregating heterogeneous content, by analysing the effect composite objects have on search behaviour and perceived workload, and by investigating different approaches to constructing such objects from heterogeneous results. Overall, our experimental findings support our assertion that central documents (i.e., *pivots*) within composite objects are decisive in determining their usefulness, and that the overall properties of composite objects (i.e., object *relevance*, *diversity* and *coherence*) play a combined role in influencing search behaviour and also in the algorithmic result composition process. The following sections summarise the main contributions and main conclusions derived from our experiments.

9.1 Summary of Contributions

Broadly, the main contributions of this thesis are:

An exploration of user-constructed composite objects. In chapter 4, we present the outcomes of an exploratory user-study, conducted with 40 participants, aimed at understanding how users interact with heterogeneous content in a search scenario, and how they manually construct composite objects that satisfy their information needs. In section 4.4.1 we analyse the contents of user-generated composite objects, in order to understand how composite objects are structured by users. In section 4.4.2 we take a look at users' assessments of composite object quality, in order to determine a hierarchy of object properties with respect to their importance to users. Our work represents a first study of users' perspectives on the aggregation of heterogeneous web content, rather than simply users' preferences with respect to finding or accessing heterogeneous information, and has implications for the design of novel search interface elements that aggregate information from multiple sources.

An analysis of composite object influence on user search behaviour. In chapter 5, we present the outcomes of a large-scale, crowd-sourced user study, with more than 500 unique participants, investigating the effect various types of entity cards have on searchers' behaviour and their perceived task workload in a traditional web search environment. In our definition, entity cards are instances of composite objects, and as such, we assume our findings generalise to other types of composite objects integrated into result pages in similar way to entity cards. In section 5.4.2 we study how entity card relevance influences search behaviour, and in section 5.4.3 we manipulate card *diversity* and card *coherence* in order to understand how these properties of composite objects impact users' behaviour and perceived workload. Our work represents a first study investigating the interplay between entity card properties, search behaviour and user perceived workload and has implications for the design and application of these novel interface elements in web search systems.

An algorithmic framework for constructing composite objects under constraints. In chapter 6, we adapt a general *composite retrieval* framework to the task of constructing composite objects from heterogeneous web search results. The work presented in chapter 6 is an attempt at moving beyond merging blocks of results from different verticals within a singular ranking, as in the case of aggregated search, and constructing composite objects, focused on specific topics of a searcher's query and containing results from multiple verticals, as a way of giving users access to heterogeneous content within a unified results page. In section 6.3 we propose multiple algorithmic approaches for constructing composite objects. We evaluate these approaches in section 6.6 in order to understand how composite object properties can be manipulated effectively and assessed under constraints. Our main contribution is an algorithmic framework that can be used to construct composite objects, from highly heterogeneous results, while maintaining constraints on composite object properties.

A method for representing heterogeneous documents within a unified feature space. In chapter 7, we investigate the application of graph-learnt representations to the retrieval and aggregation of heterogeneous content. In section 7.4 we provide an in-depth analysis of click-graph structures with respect to graph connectedness and with respect to the distribution of heterogeneous content across graph components. In section 7.4.3 we propose methods of manipulating click graph structure in order to limit biases introduced by the distribution of heterogeneous content across sub-graphs, and limit their effect on graph-based representation learning algorithms. Our work is a first analysis of ***(i)*** the distribution of heterogeneous documents across the components of a web-scale click-graph and ***(ii)*** the use of graph-learnt representations in bridging the cross-vertical gap.

9.2 Summary of Conclusions

We now summarise the main conclusions drawn from our experimental results.

On user-generate composite objects. In chapter 4, we asked participants to our study to construct composite objects and assess their quality. We observe a trend for composite objects to contain central documents, or *pivots*, that are more relevant and reflect the object's topical focus. These documents tend to originate from verticals with higher semantic load (such as General Web or Wiki). Furthermore, ornament documents, which tend to be less relevant than pivots and more vertically diverse, are also included within user-generated composite objects. In our case, image, video and Q&A verticals are popular origins of ornament documents. Our results suggest that, even though relevance is crucial, less relevant documents are explicitly attached to composite objects by participants as they can provide value by complementing *pivots* and by providing diversity. This suggests that one effective strategy for result composition is to first select a small subset of key pivot documents, and then explicitly attach other documents that complement the pivots, in order to enhance coverage, complementarity and vertical diversity of objects.

With respect to user assessments of composite object quality, although our results do not establish a clear hierarchy of object properties, we make similar findings as prior work (Bailey et al., 2010*a*,*b*) and determine that relevance, coherence and diversity are important to participants, but are difficult to assess independently. Corroborated with the above-mentioned insights on vertical diversity, this implies that, although explicit relevance is crucial to users, composition of diverse results can generate additional value.

On the effect composite objects have with respect to users' search behaviour and their perceived task workload. In chapter 5, we analysed the effect different manipulations of entity cards have on searcher behaviour and perceived task workload. Firstly, with respect to card relevance, our results suggest that the presence of entity cards on the results page can lead to increased engagement with general web results, irrespective of entity card relevance. Secondly, our findings suggest that non-relevant (i.e., off-topic) entity cards tend to increase perceived task workload. Thirdly, our analysis reveals that card coherence does not have as strong an effect on user engagement with the results page or their perceived workload, compared to relevance or diversity manipulations, suggesting that cards with "imperfect" content or intent prediction (e.g., 70% relevant on-topic content) do not necessarily have a negative impact on user experience. In contrast, our experiment reveals strong effects of card diversity on perceived task workload. Our findings have salient practical implications, on one hand with respect to user modelling and evaluation approaches for modern web search, and on the other hand, with respect to the design and application of modern search interface elements.

On composite retrieval applied to heterogeneous web results. In chapter 6, we study the application of a general *composite retrieval* framework to heterogeneous web search. Through our experiments, we show that result composition can improve retrieval performance in both homogeneous and heterogeneous web search environments. In particular, in the heterogeneous environment, our proposed result composition approach, *central-plus-satellite*, outperformed, in terms of traditional evaluation metrics, current search paradigms (i.e., general web search, federated search and aggregated search). Our results also indicate that incorporating our proposed entity-based item relevance estimation and vertical relevance estimation improves result composition approaches. Finally, we find that our proposed result composition approaches can be robust with respect to quality of the initial ranking from which they are derived, in a heterogeneous web environment. The work presented in chapter 6 is an attempt at showing that more complex methods of aggregating heterogeneous results than simply merging blocks of results from different verticals within a singular ranking, as in the case of aggregated search, can be effective.

On a unified representation space for heterogeneous web results. In chapter 7, we investigate the use of a web-scale click-graph to learn representations for heterogeneous documents within a unified feature space. We first analyse graph structure, showing that typical click graphs are disconnected, with a majority of query and document vertices incorporated into small graph components. We also show a tendency for heterogeneous documents to be located in isolated, vertically uniform sub-graphs. This is problematic for graph-based representation learning algorithms because disconnected components tend to be pruned when initiating the learning mechanism, which not only removes a majority of click information contained in the click log, but also removes disproportionately more documents originating from diverse verticals (in our case, images and videos). We show that graph-based representations can be used effectively in both retrieval, which had been studied previously in a homogeneous environment only (Jiang et al., 2016), and aggregation of heterogeneous documents, and that manipulating graph structure, in order to enhance the discriminative power of feature vectors learnt through the representation learning mechanism we explored, can be effective as well, but improvement in this case is dependent on how the propagation algorithm is configured and initialised.

On pivot documents and their impact on composite objects. One of the claims that this thesis makes is that composite object usefulness is constrained by a document or set of documents that play a central role within the object – we refer to these documents as *pivots*. We provide evidence from multiple perspectives to support this assertion. With respect to user-generate composite objects, we observe a tendency for users to structure the objects they create around a central document, or more rarely, a central set of documents, that are explicitly assessed by users as being more relevant than other, more diverse, documents within the composite object (i.e., *ornaments*), and are likely to represent the object's topical focus or influence the assignment of object title, as discussed in section 4.4.1.

With respect to the role *pivot* documents play on search behaviour and workload, we observe that explicitly deteriorating the quality of *ornament* documents (in our case, the image, Wikipedia facts and related entities components of entity cards) has a weaker effect on searchers' behaviour and their perceived task workload (discussed in section 5.5), compared to relevance or diversity manipulations, suggesting that as long as the central components of composite objects are relevant (or on-topic), searchers can extract useful information from these objects.

When constructing composite objects algorithmically, we apply the strategy of selecting relevant *pivot* documents to which we attach diverse content in our *central-plus-satellite* approach, as discussed in section 6.4.1.2. Our results show that this is the most effective result composition strategy we explored, further supporting our assertion that *pivot* documents play a critical role in determining composite object usefulness.

9.3 Final Remarks

Our work provides a novel perspective on complex information objects integrated within search engine result pages. Even as the web becomes more diverse, and web search transitions from its dated *"ten blue links"* paradigm towards accessing information in the *"semantic web"*, we expect our findings to remain informative and provide a foundation for novel hypotheses about heterogeneous information access in web search.

Bibliography

Agichtein, E., Brill, E. and Dumais, S. (2006), Improving Web Search Ranking by Incorporating User Behavior Information, *in Proceedings of SIGIR '06*, ACM, New York, NY, USA, pp. 19–26. [60, 177]

Agichtein, E., Brill, E., Dumais, S. and Ragno, R. (2006), Learning User Interaction Models for Predicting Web Search Result Preferences, *in Proceedings of SIGIR '06*, ACM, New York, NY, USA, pp. 3–10. [60, 177]

Agrawal, R., Gollapudi, S., Halverson, A. and Ieong, S. (2009), Diversifying Search Results, *in Proceedings of WSDM '09*, ACM, New York, NY, USA, pp. 5–14. [152]

Alonso, O. and Lease, M. (2011), Crowdsourcing for Information Retrieval: Principles, Methods, and Applications, *in Proceedings of SIGIR '11*, ACM, New York, NY, USA, pp. 1299–1300. [58]

Amati, G. and Van Rijsbergen, C. J. (2002), Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness, *ACM Trans. Inf. Syst.* 20(4), 357–389. [38]

Amer-Yahia, S., Bonchi, F., Castillo, C., Feuerstein, E., Méndez-Díaz, I. and Zabala, P. (2013), Complexity and Algorithms for Composite Retrieval, *in Proceedings of WWW '13*, ACM, New York, NY, USA, pp. 79–80. [52, 78, 79, 145, 153, 154, 156]

Amer-Yahia, S., Bonchi, F., Castillo, C., Feuerstein, E., Mendez-Diaz, I. and Zabala, P. (2014), Composite Retrieval of Diverse and Complementary Bundles, *IEEE Transactions on Knowledge and Data Engineering* 26(11), 2662–2675. [52, 53, 54, 78, 79, 89, 144, 145, 146, 147, 148, 149, 150, 152, 153, 154, 163]

André, P., Teevan, J. and Dumais, S. T. (2009), From X-rays to Silly Putty via Uranus: Serendipity and Its Role in Web Search, *in Proceedings of CHI '09*, ACM, New York, NY, USA, pp. 2033–2036. [12, 29]

Angel, A., Chaudhuri, S., Das, G. and Koudas, N. (2009), Ranking Objects Based on Relationships and Fixed Associations, *in Proceedings of EDBT '09*, ACM, New York, NY, USA, pp. 910–921. [79]

Arguello, J. (2015), Improving Aggregated Search Coherence, *in Advances in Information Retrieval*, Springer International Publishing, Cham, pp. 25–36. [73]

Arguello, J. (2017), *Aggregated Search*, Now Publishers Inc., Hanover, MA, USA. [6, 49, 50, 70, 71, 72, 97, 162, 163, 164, 173, 174, 206, 208]

Arguello, J. and Capra, R. (2012), The Effect of Aggregated Search Coherence on Search Behavior, *in Proceedings of CIKM '12*, ACM, New York, NY, USA, pp. 1293–1302. [71, 72, 73, 79, 80, 105, 107, 109, 110]

Arguello, J. and Capra, R. (2014), The Effects of Vertical Rank and Border on Aggregated Search Coherence and Search Behavior, *in Proceedings of CIKM '14*, ACM, New York, NY, USA, pp. 539–548. [6, 58, 73, 99, 101, 105]

Arguello, J. and Capra, R. (2016), The Effects of Aggregated Search Coherence on Search Behavior, *ACM Trans. Inf. Syst.* 35(1), 2:1–2:30. [6, 58, 72, 73, 79, 80, 99, 100, 105, 120, 141, 174]

Arguello, J., Capra, R. and Wu, W.-C. (2013), Factors Affecting Aggregated Search Coherence and Search Behavior, *in Proceedings of CIKM '13*, ACM, New York, NY, USA, pp. 1989–1998. [73, 105, 110]

Arguello, J., Diaz, F., Callan, J. and Crespo, J.-F. (2009), Sources of Evidence for Vertical Selection, *in Proceedings of the SIGIR '09*, ACM, New York, NY, USA, pp. 315–322. [5, 209]

Arguello, J., Wu, W.-C., Kelly, D. and Edwards, A. (2012), Task Complexity, Vertical Display and User Interaction in Aggregated Search, *in Proceedings of SIGIR '12*, ACM, New York, NY, USA, pp. 435–444. [70, 72, 100]

Ashkan, A., Clarke, C. L., Agichtein, E. and Guo, Q. (2009), *in* M. Boughanem, C. Berrut, J. Mothe and C. Soule-Dupuy, eds, *Advances in Information Retrieval*, Vol. 5478 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Germany, pp. 578–586. [107]

Aula, A., Khan, R. M. and Guan, Z. (2010), How Does Search Behavior Change As Search Becomes More Difficult?, *in Proceedings of CHI '10*, ACM, New York, NY, USA, pp. 35–44. [209]

Aula, A., Khan, R. M., Guan, Z., Fontes, P. and Hong, P. (2010), A Comparison of Visual and Textual Page Previews in Judging the Helpfulness of Web Pages, *in Proceedings of WWW '10*, ACM, New York, NY, USA, pp. 51–60. [65]

Azzopardi, L. (2011*a*), The Economics in Interactive Information Retrieval, *in Proceedings of SIGIR '11*, ACM, New York, NY, USA, pp. 15–24. [28]

Azzopardi, L. (2011*b*), Searching For Unlawful Carnal Knowledge, *in Proceedings of SIGIR '11 Workshop on Supporting Complex Search Tasks*, ACM, New York, NY, USA. [29]

Azzopardi, L. (2017), Building Cost-Benefit Models of Information Interactions, *in Proceedings of CHIIR '17*, ACM, New York, NY, USA, pp. 425–428. [28]

Azzopardi, L. and Zuccon, G. (2016), An Analysis of the Cost and Benefit of Search Interactions, *in Proceedings of ICTIR '16*, ACM, New York, NY, USA, pp. 59–68. [28]

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. [30, 31, 33, 37, 39, 40]

Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A. and Tahaghoghi, S. (2010*a*), Evaluating Whole-page Relevance, *in Proceedings of SIGIR '10*, ACM, New York, NY, USA, pp. 767–768. [82, 94, 216]

Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A. and Tahaghoghi, S. M. (2010*b*), Evaluating Search Systems Using Result Page Context, *in Proceedings of IIiX '10*, ACM, New York, NY, USA, pp. 105–114. [82, 94, 216]

Basu Roy, S., Amer-Yahia, S., Chawla, A., Das, G. and Yu, C. (2010), Constructing and Exploring Composite Items, *in Proceedings of SIGMOD '10*, ACM, New York, NY, USA, pp. 843–854. [52, 54, 155]

Bates, M. (1989), The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *On-line Review* 13. [27]

Baudisch, P., Lee, B. and Hanna, L. (2004), Fishnet, a Fisheye Web Browser with Search Term Popouts: A Comparative Evaluation with Overview and Linear View, *in Proceedings of the working conference on Advanced visual interfaces*, ACM, New York, NY, USA, pp. 133–140. [64]

Bayes, T. (1763), An essay towards solving a Problem in the Doctrine of Chances, *Philosophical Transactions* 53, 370–418. [34]

Beeferman, D. and Berger, A. (2000), Agglomerative Clustering of a Search Engine Query Log, *in Proceedings of KDD '00*, ACM, New York, NY, USA, pp. 407–416. [180, 190]

Belkin, N. (1978), Information Concepts for Information Science, 34, 55–85. [24]

Belkin, N. and Cool, C. (2002), Classification of Interactions with Information, *in Proceedings of CoLIS '02*, Libraries Unlimited, Greenwood Village, CO, USA, pp. 1–15. [24]

Belkin, N. J., Cole, M. and Bierig, R. (2008), Is Relevance the Right Criterion for Evaluating Interactive Information Retrieval?, *in Proceedings of SIGIR '08 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments*. [25]

Belkin, N. J. and Robertson, S. E. (1976), Information Science and the Phenomenon of Information, *J. Am. Soc. Inf. Sci. Technol.* 27(4), 197–204. [24]

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F. and Secret, A. (1994), The World-Wide Web, *Commun. ACM* 37(8), 76–82. [1]

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), The Semantic Web, *Scientific American* 284(5), 34–43. [210]

Bernstein, Y., Shokouhi, M. and Zobel, J. (2006), Compact Features for Detection of Near-duplicates in Distributed Retrieval, *in Proceedings of SPIRE '06*, Springer-Verlag, Berlin, Heidelberg, pp. 110–121. [48]

Bezdek, J. C., Ehrlich, R. and Full, W. (1984), FCM: The Fuzzy C-means Clustering Algorithm, *Computers & Geosciences* 10(2-3), 191–203. [54]

Bordino, I., Mejova, Y. and Lalmas, M. (2013), Penguins in Sweaters, or Serendipitous Entity Search on User-generated Content, *in Proceedings of CIKM '13*, ACM, New York, NY, USA, pp. 109–118. [12, 29]

Borlund, P. (2003*a*), The Concept of Relevance in IR, *J. Am. Soc. Inf. Sci. Technol.* 54(10), 913–925. [25]

Borlund, P. (2003*b*), The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems, *Information Research. An International Electronic Journal* 8(3). [44]

Bota, H. (2015), Heterogeneous Information Access Through Result Composition, *in Proceedings of the Symposium on Future Directions in Information Access*, BCS Learning & Development Ltd., Swindon, UK, pp. 20–24. [18]

Bota, H. (2016), Nonlinear Composite Search Results, *in Proceedings of CHIIR '16*, ACM, New York, NY, USA, pp. 345–347. [19]

Bota, H., Zhou, K. and Jose, J. M. (2015), *in* A. Hanbury, G. Kazai, A. Rauber and N. Fuhr, eds, *Advances in Information Retrieval*, Springer International Publishing, Cham, Switzerland, pp. 13–24. [18, 77]

Bota, H., Zhou, K. and Jose, J. M. (2016), Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload, *in Proceedings of CHIIR '16*, ACM, New York, NY, USA, pp. 131–140. [13, 14, 19, 74, 97, 209, 213]

Bota, H., Zhou, K., Jose, J. M. and Lalmas, M. (2014), Composite Retrieval of Heterogeneous Web Search, *in Proceedings of WWW '14*, ACM, New York, NY, USA, pp. 119–130. [18, 145]

Brenes, D. J., Gayo-Avello, D. and Pérez-González, K. (2009), Survey and Evaluation of Query Intent Detection Methods, *in Proceedings of WSCD '09*, ACM, New York, NY, USA, pp. 1–7. [107]

Brennan, K., Kelly, D. and Arguello, J. (2014), The Effect of Cognitive Abilities on Information Search for Tasks of Varying Levels of Complexity, *in Proceedings IIiX '14*, ACM, New York, NY, USA, pp. 165–174. [58, 100, 110]

Brin, S. and Page, L. (1998*a*), The Anatomy of a Large-scale Hypertextual Web Search Engine, *in Proceedings of WWW*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, pp. 107–117. [13]

Brin, S. and Page, L. (1998*b*), The Anatomy of a Large-scale Hypertextual Web Search Engine, *Comput. Netw. ISDN Syst.* 30(1-7), 107–117. [39]

Broder, A. (2002), A Taxonomy of Web Search, *SIGIR Forum* 36(2), 3–10. [22, 209]

Brodsky, A., Morgan Henshaw, S. and Whittle, J. (2008), CARD: A Decision-guidance Framework and Application for Recommending Composite Alternatives, *in Proceedings of RecSys '08*, ACM, New York, NY, USA, pp. 171–178. [54]

Bron, M., van Gorp, J., Nack, F., Baltussen, L. B. and de Rijke, M. (2013), Aggregated Search Interface Preferences in Multi-session Search Tasks, *in Proceedings of SIGIR '13*, ACM, New York, NY, USA, pp. 123–132. [71, 79, 80]

Buscher, G., Dumais, S. T. and Cutrell, E. (2010), The Good, the Bad, and the Random: An Eye-tracking Study of Ad Quality in Web Search, *in Proceedings of SIGIR '10*, ACM, New York, NY, USA, pp. 42–49. [62]

Callan, J. P., Lu, Z. and Croft, W. B. (1995), Searching Distributed Collections with Inference Networks, *in Proceedings of SIGIR '95*, ACM, New York, NY, USA, pp. 21–28. [48, 164]

Capra, R., Arguello, J. and Scholer, F. (2013), Augmenting Web Search Surrogates with Images, *in Proceedings of CIKM '13*, ACM, New York, NY, USA, pp. 399–408. [65, 105, 207]

Carbonell, J. and Goldstein, J. (1998), The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries, *in Proceedings of SIGIR '98*, ACM, New York, NY, USA, pp. 335–336. [158]

Carterette, B. (2015), Bayesian Inference for Information Retrieval Evaluation, *in Proceedings of ICTIR '15*, ACM, New York, NY, USA, pp. 31–40. [113]

Chaudhuri, S. and Kaushik, R. (2009), Extending Autocompletion to Tolerate Errors, *in Proceedings of SIGMOD '09*, ACM, New York, NY, USA, pp. 707–718. [60]

Chen, Y., Liu, Y., Zhou, K., Wang, M., Zhang, M. and Ma, S. (2015), Does Vertical Bring More Satisfaction?: Predicting Search Satisfaction in a Heterogeneous Environment, *in Proceedings of CIKM '15*, ACM, New York, NY, USA, pp. 1581–1590. [71, 72]

Chilton, L. B. and Teevan, J. (2011), Addressing People's Information Needs Directly in a Web Search Result Page, *in Proceedings of WWW '11*, ACM, New York, NY, USA, pp. 27–36. [14, 16, 22, 74, 75, 213]

Chuklin, A. and Serdyukov, P. (2012), Good Abandonments in Factoid Queries, *in Proceedings of WWW '12*, ACM, New York, NY, USA, pp. 483–484. [15, 74]

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 SIGIR '07*, ACM, New York, NY, USA, pp. 135–142. [61, 64, 207]

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 SIGIR '08*, ACM, New York, NY, USA, pp. 659–666. [149, 152, 162, 166, 173]

Cleverdon, C. (1967), The Cranfield Tests on Index Language Devices, *in Proceedings of Aslib '67*, Vol. 19, pp. 173–194. [25]

Cleverdon, C. (1970), Evaluation tests of information retrieval systems, *Journal of Documentation* 26(1), 55–67. [40]

Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences (Revised Edition)*, Academic Press. [116]

Cohen, J. (1992), Statistical power analysis, *Current directions in psychological science* 1(3), 98–101. [116, 126, 134, 139]

Cole, M., Liu, J., Belkin, N. J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J. and Zhang, X. (2009), Usefulness as the Criterion for Evaluation of Interactive Information Retrieval, *in Proceedings of HCIR '09*, pp. 1–4. [25, 26]

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. (2017), Algorithmic Decision Making and the Cost of Fairness, *in Proceedings of KDD '17*, ACM, New York, NY, USA, pp. 797–806. [13]

Corley, C. and Mihalcea, R. (2005), Measuring the Semantic Similarity of Texts, *in Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Association for Computational Linguistics, pp. 13–18. [85]

Craswell, N. and Szummer, M. (2007), Random Walks on the Click Graph, *in Proceedings of SIGIR '07*, ACM, New York, NY, USA, pp. 239–246. [180]

Craswell, N., Zoeter, O., Taylor, M. and Ramsey, B. (2008), An Experimental Comparison of Click Position-bias Models, *in Proceedings of WSDM '08*, ACM, New York, NY, USA, pp. 87–94. [61]

Croft, B., Metzler, D. and Strohman, T. (2009), *Search Engines: Information Retrieval in Practice*, 1st edn, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. Updated online in 2015. [xi, 30, 32, 33, 34, 37, 40, 41, 203, 204, 206]

Cucerzan, S. and Brill, E. (2004), Spelling correction as an iterative process that exploits the collective knowledge of web users, *in Proceedings of EMNLP '04*, Association for Computational Linguistics. [60]

Cutrell, E. and Guan, Z. (2007), What Are You Looking for?: An Eye-tracking Study of Information Usage in Web Search, *in Proceedings of CHI '07*, ACM, New York, NY, USA, pp. 407–416. [64, 207]

Cutting, D. R., Karger, D. R., Pedersen, J. O. and Tukey, J. W. (1992), Scatter-/Gather: A Cluster-based Approach to Browsing Large Document Collections, *in Proceedings of SIGIR '92*, ACM, New York, NY, USA, pp. 318–329. [69]

Davidson, I. and Ravi, S. S. (2007), Intractability and Clustering with Constraints, *in Proceedings of ICML '07*, ACM, New York, NY, USA, pp. 201–208. [54]

Davidson, I., Ravi, S. S. and Ester, M. (2007), Efficient Incremental Constrained Clustering, *in Proceedings of KDD '07*, ACM, New York, NY, USA, pp. 240–249. [54]

De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R. and Yu, C. (2010), Automatic Construction of Travel Itineraries Using Social Breadcrumbs, *in Proceedings of HyperText '10*, ACM, New York, NY, USA, pp. 35–44. [54]

Demeester, T., Trieschnigg, D., Nguyen, D. and Hiemstra, D. (2013), Overview of the TREC 2013 federated web search track, *in Proceedings of the Text Retrieval Conference*, pp. 1–11. [80, 84]

Deng, T., Fan, W. and Geerts, F. (2012), On the Complexity of Package Recommendation Problems, *in Proceedings of the Symposium on Principles of Database Systems*, ACM, New York, NY, USA, pp. 261–272. [79]

Dewey, C. (2016), You probably haven't even noticed Google's sketchy quest to control the world's knowledge, Washington Post. Accessed 2018-03-24.
URL: *https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge* [211]

Diaz, F., Lalmas, M. and Shokouhi, M. (2010), From Federated to Aggregated Search, *in Proceeding of SIGIR '10*, pp. 910–910. [77]

Diaz, F., White, R., Buscher, G. and Liebling, D. (2013), Robust Models of Mouse Movement on Dynamic Web Search Results Pages, *in Proceedings of CIKM '13*, ACM, New York, NY, USA, pp. 1451–1460. [100]

Divoli, A., Wooldridge, M. A. and Hearst, M. A. (2010), Full Text and Figure Display Improves Bioscience Literature Search, *PLOS ONE* 5(4), 1–15. [65]

Dumais, S., Cutrell, E. and Chen, H. (2001), Optimizing Search by Showing Results in Context, *in Proceedings of CHI '01*, ACM, New York, NY, USA, pp. 277–284. [x, 67, 71, 162]

Dumais, S. T., Buscher, G. and Cutrell, E. (2010), Individual Differences in Gaze Patterns for Web Search, *in Proceedings of IIiX '10*, ACM, New York, NY, USA, pp. 185–194. [62]

Edmonds, A., White, R. W., Morris, D. and Drucker, S. M. (2007), Instrumenting the Dynamic Web, *J. Web Eng.* 6(3), 244–260. [61]

Ekstrom, R., French, J., Harman, H. and Dermen, D. (1979), *Kit of Factor-Referenced Cognitive Tests*, Educational Testing Service, Princeton, NJ, USA. [72, 209]

Elsweiler, D., Mandl, S. and Kirkegaard Lunn, B. (2010), Understanding Casual-leisure Information Needs: A Diary Study in the Context of Television Viewing, *in Proceedings of IIiX '10*, ACM, New York, NY, USA, pp. 25–34. [29]

Elsweiler, D., Wilson, M. and Kirkegaard Lunn, B. (2011), Understanding Casual-Leisure Information Behaviour, *in Library and Information Science* 1, 211–241. [29]

Enge, E. (2017), Featured Snippets: New Insights, New Opportunities. Accessed 2018-03-24.
URL: *https://www.stonetemple.com/featured-snippets-new-insights-new-opportunities/* [14, 74, 209]

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. (2015), Certifying and Removing Disparate Impact, *in Proceedings of KDD '15*, ACM, New York, NY, USA, pp. 259–268. [13]

Ferragina, P. and Scaiella, U. (2010), TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities), *in Proceedings of CIKM '10*, ACM, New York, NY, USA, pp. 1625–1628. [159]

Ford, H. and Graham, M. (2016), Semantic Cities: Coded Geopolitics and the Rise of the Semantic Web, *in Code and the City*, Routledge, London, UK. [14, 59, 210]

Fourney, A., Morris, M. R. and White, R. W. (2017), Web Search As a Linguistic Tool, *in Proceedings of WWW '17*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 549–557. [59, 76]

Fuhr, N. (1992), Probabilistic Models in Information Retrieval, *Comput. J.* 35(3), 243–255. [34]

Fuhr, N. (2018), Some Common Mistakes In IR Evaluation, And How They Can Be Avoided, *SIGIR Forum* 51(3), 32–41. [42]

Gao, J., He, X. and Nie, J.-Y. (2010), Clickthrough-based Translation Models for Web Search: From Word Models to Phrase Models, *in Proceedings of CIKM '10*, ACM, New York, NY, USA, pp. 1139–1148. [177, 180]

Gelman, A. and Tuerlinckx, F. (2000), Type S error rates for classical and Bayesian single and multiple comparison procedures, *Computational Statistics* 15(3), 373–390. [113, 116]

Goșa, A., Popa, S. and Coța, A. (2018), Survival Analysis of Post High-School Friendship in Eastern European Cultures., *Journal of Romanian Studies* 3(1), 1–218. [113]

Golbus, P. B., Zitouni, I., Kim, J. Y., Hassan, A. and Diaz, F. (2014), Contextual and Dimensional Relevance Judgments for Reusable SERP-level Evaluation, *in Proceedings of WWW '14*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 131–142. [82]

Goodman, S. (1999), Toward evidence-based medical statistics. 1: The p value fallacy, *Annals of Internal Medicine* 130(12), 995–1004. [113]

Guo, Q. and Agichtein, E. (2012), Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and Other Post-click Searcher Behavior, *in Proceedings of WWW '12*, ACM, New York, NY, USA, pp. 569–578. [62]

Guo, X. and Ishikawa, Y. (2011), Multi-objective Optimal Combination Queries, *in Proceedings of the International Conference on Database and Expert Systems Applications*, Springer-Verlag, Berlin, Heidelberg, Germany, pp. 47–61. [79]

Gwizdka, J. and Zhang, Y. (2015), Differences in Eye-Tracking Measures Between Visits and Revisits to Relevant and Irrelevant Web Pages, *in Proceedings of SIGIR '15*, ACM, New York, NY, USA, pp. 811–814. [62]

Hall, W. and O'Hara, K. (2009), *in* R. Meyers, ed., *Encyclopedia of Complexity and System Science*, Springer, New York, NY. [12]

Harman, D. (1992), Relevance Feedback Revisited, *in Proceedings of SIGIR '92*, ACM, New York, NY, USA, pp. 1–10. [33]

Hart, S. G. and Staveland, L. E. (1988), *in* P. A. Hancock and N. Meshkati, eds, *Human Mental Workload*, Vol. 52 of *Advances in Psychology*, North-Holland, pp. 139–183. [58, 100, 110]

Harvey, M., Wilson, M. and Church, K. (2014), Workshop on Searching for Fun 2014, *in Proceedings of IIiX '14*, ACM, New York, NY, USA, pp. 6–6. [29]

Hasibi, F., Balog, K. and Bratsberg, S. E. (2017), Dynamic Factual Summaries for Entity Cards, *in Proceedings of the SIGIR '17*, ACM, New York, NY, USA, pp. 773–782. [13, 75, 208, 213]

Hearst, M. (2009), *Search user interfaces*, Cambridge University Press. [3, 27, 62, 63, 64, 66, 67, 69]

Hearst, M. A. (2006), Clustering Versus Faceted Categories for Information Exploration, *Commun. ACM* 49(4), 59–61. [68]

Hearst, M. A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. A. and Ye, J. (2007), BioText Search Engine: beyond abstract search, *Bioinformatics* 23(16), 2196–2197. [65]

Hearst, M. A., Divoli, A., Ye, J. and Wooldridge, M. A. (2007), Exploring the Efficacy of Caption Search for Bioscience Journal Search Interfaces, *in Biological, translational, and clinical language processing, BioNLP@ACL*, pp. 73–80. [65]

Hersh, W. (2002), Text Retrieval Conference (TREC) Genomics Pre-track Workshop, *in Proceedings of the JCDL '02*, ACM, New York, NY, USA, pp. 428–428. [44]

Huang, J., White, R. W. and Dumais, S. (2011), No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search, *in Proceedings of CHI '11*, ACM, New York, NY, USA, pp. 1225–1234. [61]

Ingwersen, P. and Järvelin, K. (2005), *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer-Verlag, Secaucus, NJ, USA. [x, 20, 21, 22, 24, 26, 27, 209]

Introna, L. D. and Nissenbaum, H. (2000), Shaping the Web: Why the Politics of Search Engines Matters, *The Information Society* 16(3), 169–185. [13]

Jansen, B. J., Booth, D. and Smith, B. (2009), Using the Taxonomy of Cognitive Learning to Model Online Searching, *Inf. Process. Manage.* 45(6), 643–663. [209]

Jansen, B. J. and Pooch, U. (2001), A Review of Web Searching Studies and a Framework for Future Research, *J. Am. Soc. Inf. Sci. Technol.* 52(3), 235–246. [79]

Järvelin, K. and Ingwersen, P. (2004), Information Seeking Research Needs Extension Toward Tasks and Technology, *Information Research* 10(1). Paper 212. [x, 20, 21, 27]

Järvelin, K. and Kekäläinen, J. (2002), Cumulated Gain-based Evaluation of IR Techniques, *ACM Trans. Inf. Syst.* 20(4), 422–446. [43]

Jiang, J. J. and Conrath, D. W. (1997), Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *CoRR* cmp-lg/9709008. [85]

Jiang, S., Hu, Y., Kang, C., Daly, Jr., T., Yin, D., Chang, Y. and Zhai, C. (2016), Learning Query and Document Relevance from a Web-scale Click Graph, *in Proceedings of SIGIR '16*, ACM, New York, NY, USA, pp. 185–194. [60, 177, 179, 180, 181, 185, 186, 189, 195, 217]

Jiao, B., Yang, L., Xu, J. and Wu, F. (2010), Visual Summarization of Web Pages, *in Proceedings of SIGIR '10*, ACM, New York, NY, USA, pp. 499–506. [65]

Johnson, D. H. (1999), The insignificance of statistical significance testing, *Journal of Wildlife Management* 63(3), 763–772. [113]

Kaisser, M., Hearst, M. and Lowe, J. B. (2008), Improving Search Results Quality by Customizing Summary Lengths, *in Proceedings of ACL-HLT '08*, Association for Computational Linguistics, pp. 701–709. [64, 207]

Käki, M. (2005), Findex: Search Result Categories Help Users when Document Ranking Fails, *in Proceedings of CHI '05*, ACM, New York, NY, USA, pp. 131–140. [68]

Käki, M. and Aula, A. (2005), Findex: Improving Search Result Use Through Automatic Filtering Categories, *Interact. Comput.* 17(2), 187–206. [68]

Kekäläinen, J. and Järvelin, K. (2002), Evaluating Information Retrieval Systems Under The Challenges Of Interaction And Multidimensional Dynamic Relevance, *in Proceedings of CoLIS '02*, Libraries Unlimited, Greenwood Village, CO, USA, pp. 253–270. [20, 21]

Kelly, D. (2009), Methods for Evaluating Interactive Information Retrieval Systems with Users, *Found. Trends Inf. Retr.* 3(1&#8212;2), 1–224. [57, 58]

Kleinberg, J. M. (1999), Authoritative Sources in a Hyperlinked Environment, *J. ACM* 46(5), 604–632. [39, 185]

Koffka, K. (2013), *Principles of Gestalt psychology*, Vol. 44, Routledge. [71]

Kong, W. and Allan, J. (2013), Extracting Query Facets from Search Results, *in Proceedings of SIGIR '13*, ACM, New York, NY, USA, pp. 93–102. [209]

Kruschke, J. K. (2013), Bayesian estimation supersedes the t test., *Journal of Experimental Psychology: General* 142(2), 573. [113]

Kuhlthau, C. C. (1991), Inside the Search Process: Information Seeking from the User's Perspective, *J. Am. Soc. Inf. Sci. Technol.* 42(5), 361–371. [28]

Kules, B. and Shneiderman, B. (2008), Users Can Change Their Web Search Tactics: Design Guidelines for Categorized Overviews, *Inf. Process. Manage.* 44(2), 463–484. [68]

Kurland, O. (2008), The Opposite of Smoothing: A Language Model Approach to Ranking Query-specific Document Clusters, *in Proceedings of SIGIR '08*, ACM, New York, NY, USA, pp. 171–178. [51]

Kurland, O. and Domshlak, C. (2008), A Rank-aggregation Approach to Searching for Optimal Query-specific Clusters, *in Proceedings of SIGIR '08*, ACM, New York, NY, USA, pp. 547–554. [148]

Kusner, M. J., Sun, Y., Kolkin, N. I. and Weinberger, K. Q. (2015), From Word Embeddings to Document Distances, *in Proceedings of ICML '15*, JMLR.org, pp. 957–966. [189]

Lagun, D., Hsieh, C.-H., Webster, D. and Navalpakkam, V. (2014), Towards Better Measurement of Attention and Satisfaction in Mobile Search, *in Proceedings of SIGIR '14*, ACM, New York, NY, USA, pp. 113–122. [74, 100, 208]

Lagun, D., McMahon, D. and Navalpakkam, V. (2016), Understanding Mobile Searcher Attention with Rich Ad Formats, *in Proceedings of CIKM '16*, ACM, New York, NY, USA, pp. 599–608. [13, 213]

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* pp. 71–136. [64, 68]

Lease, M. and Yilmaz, E. (2013), Crowdsourcing for Information Retrieval: Introduction to the Special Issue, *Information Retrieval* 16(2), 91–100. [58]

Leroy, V., Amer-Yahia, S., Gaussier, E. and Mirisaee, H. (2015), Building Representative Composite Items, *in Proceedings of CIKM '15*, ACM, New York, NY, USA, pp. 1421–1430. [52, 54]

Li, J., Huffman, S. and Tokuda, A. (2009), Good Abandonment in Mobile and PC Internet Search, *in Proceedings of SIGIR '09*, ACM, New York, NY, USA, pp. 43–50. [75]

Li, Z., Shi, S. and Zhang, L. (2008), Improving Relevance Judgment of Web Search Results with Image Excerpts, *in Proceedings of WWW '08*, ACM, New York, NY, USA, pp. 21–30. [65]

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J. and Zhang, X. (2010), Search Behaviors in Different Task Types, *in Proceedings of JCDL '10*, ACM, New York, NY, USA, pp. 69–78. [209]

Liu, J., Liu, C., Cole, M., Belkin, N. J. and Zhang, X. (2012), Exploring and Predicting Search Task Difficulty, *in Proceedings of CIKM '12*, ACM, New York, NY, USA, pp. 1313–1322. [209]

Liu, J., Liu, C., Gwizdka, J. and Belkin, N. J. (2010), Can Search Systems Detect Users' Task Difficulty?: Some Behavioral Signals, *in Proceedings of SIGIR '10*, ACM, New York, NY, USA, pp. 845–846. [209]

Liu, T.-Y. (2009), Learning to Rank for Information Retrieval, *Found. Trends Inf. Retr.* 3(3), 225–331. [39]

Liu, X. and Croft, W. B. (2004), Cluster-based Retrieval Using Language Models, *in Proceedings of SIGIR '04*, ACM, New York, NY, USA, pp. 186–193. [51]

Liu, X. and Croft, W. B. (2008), Evaluating Text Representations for Retrieval of the Best Group of Documents, *in Proceedings of BCS ECIR '08*, Springer-Verlag, Berlin, Heidelberg, pp. 454–462. [51]

Liu, Z., Liu, Y., Zhou, K., Zhang, M. and Ma, S. (2015), Influence of Vertical Result in Web Search Examination, *in Proceedings of SIGIR '15*, ACM, New York, NY, USA, pp. 193–202. [72, 99]

Lundquist, C., Grossman, D. A. and Frieder, O. (1997), Improving Relevance Feedback in the Vector Space Model, *in Proceedings of CIKM '97*, ACM, New York, NY, USA, pp. 16–23. [33]

Lupu, M., Huang, J., Zhu, J. and Tait, J. (2009), TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC, *SIGIR Forum* 43(2), 63–70. [44]

Mager, A. (2018), Internet Governance as Joint Effort: (Re)ordering Search Engines at the Intersection of Global and Local Cultures, *New Media & Society* . [13]

Manning, C. D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA. [37, 38, 51]

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J. and Luo, H. (2016), When Does Relevance Mean Usefulness and User Satisfaction in Web Search?, *in Proceedings of SIGIR '16*, ACM, New York, NY, USA, pp. 463–472. [25, 26]

Marchionini, G. (1995), *Information Seeking in Electronic Environments*, Cambridge University Press, New York, NY, USA. [23]

Marchionini, G. (2006), Exploratory Search: From Finding to Understanding, *Commun. ACM* 49(4), 41–46. [12, 16]

Marchionini, G. and White, R. (2007), Find What You Need, Understand What You Find, *International Journal of Human-Computer Interaction* 23(3), 205–237. [23, 27]

Mathews, J. (2015), Is Hillary Clinton getting taller? Or is the Internet getting dumber?, Washington Post. Accessed 2018-03-24.
URL: *https://www.washingtonpost.com/lifestyle/style/is-hillary-clinton-getting-taller-or-is-the-internet-getting-dumber/2015/09/24/58af4dfa-5e33-11e5-9757-e49273f05f65_story.html* [211]

Maxwell, D., Azzopardi, L. and Moshfeghi, Y. (2017), A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience, *in Proceedings of SIGIR '17*, ACM, New York, NY, USA, pp. 135–144. [58, 64]

McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. (2017), Abandon Statistical Significance.
URL: *https://arxiv.org/abs/1709.07588* [113]

Megaw, T. (2005), *The Definition and Measurement of Mental Workload*, CRC Press, Boca Raton, Florida, USA. [110]

Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H. and Yilmaz, E. (2017), Auditing Search Engines for Differential Satisfaction Across Demographics, *in Proceedings of WWW '17*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 626–633. [13]

Meyers-Levy, J. and Sternthal, B. (1993), A Two-Factor Explanation of Assimilation and Contrast Effects, *Journal of Marketing Research* 30(3), 359–368. [99]

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), Efficient Estimation of Word Representations in Vector Space, *Computing Research Repository* 1301.3781. [189]

Millan-Cifuentes, J. D., Göker, A., Myrhaug, H. and MacFarlane, A. (2014), Curiosity Driven Search: When is Relevance Irrelevant?, *in Proceedings of IIiX '14*, ACM, New York, NY, USA, pp. 279–282. [25]

Miller, G. A. (1995), WordNet: A Lexical Database for English, *Commun. ACM* 38(11), 39–41. [85]

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016), The ethics of algorithms: Mapping the debate, *Big Data & Society* 3(2). [210]

Mizzaro, S. (1997), Relevance: The Whole History, *J. Am. Soc. Inf. Sci. Technol.* 48(9), 810–832. [25]

Morales, A. and Fitzsimons, G. (2007), Product contagion: Changing consumer evaluations through physical contact with "disgusting" products, *Journal of Marketing Research* 44(2), 272–283. [99]

Müller, C. and Gurevych, I. (2009), A Study on the Semantic Relatedness of Query and Document Terms in Information Retrieval, *in Proceedings of EMNLP '09*, ACL, Stroudsburg, PA, USA, pp. 1338–1347. [177, 186]

Murdock, V. and Lalmas, M. (2008), Workshop on Aggregated Search, *SIGIR Forum* 42(2), 80–83. [5]

Narum, S. R. (2006), Beyond Bonferroni: Less Conservative Analyses for Conservation Genetics, *Conservation Genetics* 7(5), 783–787. [113]

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 WWW '13*, ACM, New York, NY, USA, pp. 953–964. [13, 15, 74, 100, 208, 213]

Nguyen, D., Demeester, T., Trieschnigg, D. and Hiemstra, D. (2012), Federated Search in the Wild: The Combined Power of over a Hundred Search Engines, *in Proceedings of CIKM '12*, ACM, New York, NY, USA, pp. 1874–1878. [45, 161]

Norman, D. (1988), *The Psychology Of Everyday Things*, Ingram Publisher Services, New York, NY, United States. [27]

Norris, J. (1997), *Markov Chains*, Cambridge University Press, Cambridge. [39]

Odijk, D., White, R. W., Hassan Awadallah, A. and Dumais, S. T. (2015), Struggling and Success in Web Search, *in Proceedings of CIKM '15*, ACM, New York, NY, USA. [184]

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group, New York, NY, USA. [13, 210]

Palmer, S. E. (1992), Common region: A new principle of perceptual grouping, *Cognitive Psychology* 24(3), 436 – 447. [71]

Pariser, E. (2011), *The Filter Bubble: What The Internet Is Hiding From You*, Penguin Books Limited. [13, 210]

Pirolli, P. and Card, S. (2005), The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, *in Proceedings of International Conference on Intelligence Analysis '05*, pp. 2–8. [23]

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 CHI '96*, ACM, New York, NY, USA, pp. 213–220. [69]

Poisson, S. D. and Schnuse, C. H. (1841), *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*, Meyer. [114]

Rahman, A. and Wilson, M. L. (2015), Exploring Opportunities to Facilitate Serendipity in Search, *in Proceedings of SIGIR '15*, ACM, New York, NY, USA, pp. 939–942. [12, 29]

Robertson, S. E. (1997), *in* K. Spärck Jones and P. Willett, eds, *Readings in Information Retrieval*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 281–286. (Reprinted from *Journal of Documentation*, 1977, 33, pp. 294–304). [33]

Robertson, S. E. (2006), On GMAP: And Other Transformations, *in Proceedings of CIKM '06*, ACM, New York, NY, USA, pp. 78–83. [42]

Robertson, S. E., Walker, S., Spärck Jones, K., Hancock-Beaulieu, M. M. and Gatford, M. (1995), Okapi at TREC–3, *in Overview of the Third Text REtrieval Conference (TREC–3)*, NIST, Gaithersburg, Maryland, USA, pp. 109–126. [189]

Robertson, S. E., Zaragoza, H. and Taylor, M. (2004), Simple BM25 Extension to Multiple Weighted Fields, *in Proceedings CIKM '04*, ACM, New York, NY, USA, pp. 42–49. [33]

Rocchio, J. J. (1971), Relevance Feedback in Information Retrieval, *The SMART Retrieval System: Experiments in Automatic Document Processing* pp. 313–323. [33]

Rose, D. E. and Levinson, D. (2004), Understanding User Goals in Web Search, *in Proceedings of WWW '04*, ACM, New York, NY, USA, pp. 13–19. [22, 79, 209]

Russell, D. M., Stefik, M. J., Pirolli, P. and Card, S. K. (1993), The Cost Structure of Sensemaking, *in Proceedings of INTERACT '93 and CHI '93*, ACM, New York, NY, USA, pp. 269–276. [23]

Sakai, T. and Kando, N. (2008), On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, *Information Retrieval* 11(5), 447–470. [44]

Salton, G. (1968), *Automatic Information Organization and Retrieval.*, McGraw Hill Text, New York, NY, USA. [30]

Salton, G., Fox, E. A. and Wu, H. (1983), Extended Boolean Information Retrieval, *Commun. ACM* 26(11). [31]

Salton, G. and Lesk, M. E. (1968), Computer Evaluation of Indexing and Text Processing, *J. ACM* 15(1), 8–36. [31]

Sanderson, M. (2008), Ambiguous Queries: Test Collections Need More Sense, *in Proceedings of SIGIR '08*, ACM, New York, NY, USA, pp. 499–506. [209]

Sanderson, M. and Croft, W. B. (2012), The History of Information Retrieval Research, *Proceedings of the IEEE* 100(Special Centennial Issue), 1444–1451. [10, 99, 105]

Sanderson, M., Paramita, M. L., Clough, P. and Kanoulas, E. (2010), Do user preferences and evaluation measures line up?, *in Proceedings of SIGIR '10*, ACM, New York, NY, USA, pp. 555–562. [54, 149, 152, 162]

Santos, R. L. (2012), Explicit Web Search Result Diversification, *SIGIR Forum* 47(1), 67–68. [8, 149]

Santos, R. L. T., Macdonald, C. and Ounis, I. (2011), *in* G. Amati and F. Crestani, eds, *Advances in Information Retrieval Theory*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 250–261. [152, 175, 201]

Saracevic, T. (1975), Relevance: A review of the literature and a framework for thinking on the notion in information science., *J. Am. Soc. Inf. Sci. Technol.* 26(6), 321–343. [25, 26]

Saracevic, T. (1996), Relevance Reconsidered, *in Proceedings of CoLIS '96*, Libraries Unlimited, Greenwood Village, CO, USA. [25, 26]

Saracevic, T. (2007*a*), Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance, *J. Am. Soc. Inf. Sci. Technol.* 58(13), 1915–1933. [25, 26]

Saracevic, T. (2007*b*), Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance, *J. Am. Soc. Inf. Sci. Technol.* 58(13), 2126–2144. [25, 26]

Satterthwaite, F. E. (1946), An Approximate Distribution of Estimates of Variance Components, *Biometrics Bulletin* 2(6), 110–114. [89]

Sawilowsky, S. S. (2009), New effect size rules of thumb, *Journal of Modern Applied Statistical Methods* 8(2), 467–474. [126, 127, 134, 139, 140]

Schamber, L., Eisenberg, M. and Nilan, M. S. (1990), A Re-examination of Relevance: Toward a Dynamic, Situational Definition, *Inf. Process. Manage.* 26(6), 755–776. [25]

schraefel, m. c. (2009), Building Knowledge: What's Beyond Keyword Search?, *Computer* 42(3), 52–59. [12]

Sequiera, R. and Lin, J. (2017), Finally, a Downloadable Test Collection of Tweets, *in Proceedings of SIGIR '17*, ACM, New York, NY, USA, pp. 1225–1228. [44]

Shen, Y., He, X., Gao, J., Deng, L. and Mesnil, G. (2014), Learning Semantic Representations Using Convolutional Neural Networks for Web Search, *in Proceedings of WWW '14*, ACM, New York, NY, USA, pp. 373–374. [177, 180]

Shneiderman, B. (1997), *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd edn, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. [22, 23]

Shokouhi, M. (2013), Learning to Personalize Query Auto-completion, *in Proceedings of SIGIR '13*, ACM, New York, NY, USA, pp. 103–112. [60]

Shokouhi, M. and Si, L. (2011), Federated Search, *Found. Trends Inf. Retr.* 5(1), 1–102. [45, 46, 47, 48]

Shokouhi, M., Zobel, J. and Bernstein, Y. (2007), Distributed Text Retrieval from Overlapping Collections, *in Proceedings of the Australasian Database Conference '07*, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, pp. 141–150. [48]

Si, L. and Callan, J. (2002), Using Sampled Data and Regression to Merge Search Engine Results, *in Proceedings of SIGIR '02*, ACM, New York, NY, USA, pp. 19–26. [48]

Si, L. and Callan, J. (2003), Relevant Document Distribution Estimation Method for Resource Selection, *in Proceedings of SIGIR '03*, ACM, New York, NY, USA, pp. 298–305. [47, 48, 155, 160, 164]

Si, L. and Callan, J. (2004), Unified Utility Maximization Framework for Resource Selection, *in Proceedings of CIKM '04*, ACM, New York, NY, USA, pp. 32–41. [48]

Si, L. and Callan, J. (2006), CLEF 2005: Multilingual Retrieval by Combining Multiple Multilingual Ranked Lists, *in Proceedings of CLEF '05: Accessing Multilingual Information Repositories*, Springer-Verlag, Berlin, Heidelberg, pp. 121–130. [48]

Si, L., Callan, J., Cetintas, S. and Yuan, H. (2008), An Effective and Efficient Results Merging Strategy for Multilingual Information Retrieval in Federated Search Environments, *Inf. Retr.* 11(1), 1–24. [48]

Silverstein, C., Marais, H., Henzinger, M. and Moricz, M. (1999), Analysis of a Very Large Web Search Engine Query Log, *SIGIR Forum* 33(1), 6–12. [181]

Spärck Jones, K. (1972), A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of documentation* 28(1), 11–21. [33]

Spärck Jones, K. (1973), Index Term Weighting, *Information storage and retrieval* 9(11), 619–633. [33]

Spärck Jones, K. and van Rijsbergen, C. J. (1976), Information Retrieval Test Collections, *Journal of documentation* 32(1), 59–75. [44]

Spärck Jones, K., Walker, S. and Robertson, S. E. (2000), A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, *Inf. Process. Manage.* 36(6), 779–808. [34]

Speier, C. and Morris, M. G. (2003), The Influence of Query Interface Design on Decision-making Performance, *MIS Q.* 27(3), 397–423. [58, 110]

Spink, A., Jansen, B. J., Wolfram, D. and Saracevic, T. (2002), From E-Sex to E-Commerce: Web Search Changes, *Computer* 35(3), 107–109. [79]

Stasi, L. L. D., Antolí, A., Gea, M. and Cañas, J. J. (2011), A Neuroergonomic Approach to Evaluating Mental Workload in Hypermedia Interactions, *International Journal of Industrial Ergonomics* 41(3), 298–304. [58, 110]

Straw, A., Wiecki, T. and Fonnesbeck, C. (2015), Bayesian Estimation Supersedes the T-Test. Accessed January 2018.
URL: *https://docs.pymc.io/notebooks/BEST.html* [113, 114]

Sushmita, S., Joho, H. and Lalmas, M. (2009), A Task-Based Evaluation of an Aggregated Search Interface, *in Proceedings of SPIRE '09*, Springer-Verlag, Berlin, Heidelberg, pp. 322–333. [70]

Sushmita, S., Joho, H., Lalmas, M. and Villa, R. (2010), Factors Affecting Click-through Behavior in Aggregated Search Interfaces, *in Proceedings of CIKM '10*, ACM, New York, NY, USA, pp. 519–528. [71, 72, 87, 208]

Sushmita, S., Piwowarski, B. and Lalmas, M. (2010), Dynamics of Genre and Domain Intents, *in Information Retrieval Technology*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 399–409. [72]

Sutcliffe, A. and Ennis, M. (1998), Towards a cognitive theory of information retrieval, *Interacting with Computers* 10(3), 321 – 351. [27]

Tavani, H. (2016), *in* E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, fall 2016 edn, Metaphysics Research Lab, Stanford University. [210]

Teevan, J. (2015), Slow Search: Improving Information Retrieval Using Human Assistance, *in Proceedings of CIKM '15*, ACM, New York, NY, USA, pp. 1–1. [29]

Teevan, J., Collins-Thompson, K., White, R. W., Dumais, S. T. and Kim, Y. (2013), Slow Search: Information Retrieval Without Time Constraints, *in Proceedings of HCIR '13*, ACM, New York, NY, USA, pp. 1:1–1:10. [12, 29]

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 CHI '09*, ACM, New York, NY, USA, pp. 2023–2032. [65, 207]

Thomas, P. and Shokouhi, M. (2009), SUSHI: Scoring Scaled Samples for Server Selection, *in Proceedings of SIGIR '09*, ACM, New York, NY, USA, pp. 419–426. [48]

Tombros, A. and Sanderson, M. (1998), Advantages of Query Biased Summaries in Information Retrieval, *in Proceedings SIGIR '98*, ACM, New York, NY, USA, pp. 2–10. [64, 207]

Tombros, A., Villa, R. and Van Rijsbergen, C. J. (2002), The Effectiveness of Query-specific Hierarchic Clustering in Information Retrieval, *Inf. Process. Manage.* 38(4), 559–582. [148]

Turpin, L., Kelly, D. and Arguello, J. (2016), To Blend or Not to Blend?: Perceptual Speed, Visual Memory and Aggregated Search, *in Proceedings of SIGIR '16*, ACM, New York, NY, USA, pp. 1021–1024. [71, 72, 209]

Vakkari, P. (2000), Relevance and Contributing Information Types of Searched Documents in Task Performance, *in Proceedings of SIGIR '00*, ACM, New York, NY, USA, pp. 2–9. [28, 59]

van Rijsbergen, C. J. (1979), *Information Retrieval*, 2nd edn, Butterworth-Heinemann, Newton, MA, USA. [50, 148]

Varadarajan, R. and Hristidis, V. (2006), A System for Query-specific Document Summarization, *in Proceedings of CIKM '06*, ACM, New York, NY, USA, pp. 622–631. [64]

Voorhees, E. M. (1985), The Cluster Hypothesis Revisited, *in Proceedings of SIGIR '85*, ACM, New York, NY, USA, pp. 188–196. [51]

Voorhees, E. M. (2013), The TREC Medical Records Track, *in Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ACM, New York, NY, USA, pp. 239:239–239:246. [44]

Voorhees, E. M. and Harman, D. (1999), The Text REtrieval Conference (TREC): History and Plans for TREC-9, *SIGIR Forum* 33(2), 12–15. [44]

Wagstaff, K. and Cardie, C. (2000), Clustering with Instance-level Constraints, *in Proceedings of ICML '00*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1103–1110. [53]

Wagstaff, K., Cardie, C., Rogers, S. and Schrödl, S. (2001), Constrained K-means Clustering with Background Knowledge, *in Proceedings of ICML '01*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 577–584. [53]

Wen, J.-R., Nie, J.-Y. and Zhang, H.-J. (2001), Clustering User Queries of a Search Engine, *in Proceedings of WWW '01*, ACM, New York, NY, USA, pp. 162–168. [180]

White, R. W. (2016), *Interactions with Search Systems*, Cambridge University Press, Cambridge, UK. [23, 27, 28, 29, 57, 59, 62]

White, R. W., Bilenko, M. and Cucerzan, S. (2007), Studying the Use of Popular Destinations to Enhance Web Search Interaction, *in Proceedings of SIGIR '07*, ACM, New York, NY, USA, pp. 159–166. [66]

White, R. W. and Horvitz, E. (2013), Captions and Biases in Diagnostic Search, *ACM Trans. Web* 7(4), 23:1–23:28. [61]

White, R. W. and Huang, J. (2010), Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs, *in Proceedings of SIGIR '10*, ACM, New York, NY, USA, pp. 587–594. [23]

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, *Inf. Process. Manage.* 39(5), 707–733. [64, 207]

White, R. W. and Roth, R. A. (2009), *Exploratory Search: Beyond the Query-Response Paradigm*, Morgan & Claypool Publishers, San Rafael, CA, USA. [12, 29]

White, R. W. and Ruthven, I. (2006), A Study of Interface Support Mechanisms for Interactive Information Retrieval, *J. Am. Soc. Inf. Sci. Technol.* 57(7), 933–948. [23]

Wiener, N. (1950), *The Human Use of Human Beings: Cybernetics and Society*, Eyre & Spottiswoode, Ltd., London, United Kingdom. [210]

Williams, K., Kiseleva, J., Crook, A. C., Zitouni, I., Awadallah, A. H. and Khabsa, M. (2016*a*), Detecting Good Abandonment in Mobile Search, *in Proceedings of WWW '16*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 495–505. [15, 74, 76]

Williams, K., Kiseleva, J., Crook, A. C., Zitouni, I., Awadallah, A. H. and Khabsa, M. (2016*b*), Is This Your Final Answer?: Evaluating the Effect of Answers on Good Abandonment in Mobile Search, *in Proceedings of SIGIR '16*, ACM, New York, NY, USA, pp. 889–892. [76]

Wilson, M. L. (2017), The Tetris Model of Resolving Information Needs Within the Information Seeking Process, *in Proceedings of CHIIR '17*, ACM, New York, NY, USA, pp. 147–154. [28]

Wilson, M. L., Kules, B., m. c. schraefel 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. [2, 10, 12, 22, 23, 28]

Wilson, M. L., Ye, C., Twidale, M. B., Grasse, H., Rosenthal, J. and McKittrick, M. (2016), Search Literacy: Learning to Search to Learn, *in Second International Workshop on Search as Learning (SAL 2016)*. Published in: Proceedings of the Second International Workshop on Search as Learning / edited by Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, Noriko Kando. V. 1647,ISSN 1613-0073. [16]

Wilson, T. D. (1997), Information Behaviour: An Interdisciplinary Perspective, *Inf. Process. Manage.* 33(4), 551–572. [28]

Wu, W.-C., Kelly, D., Edwards, A. and Arguello, J. (2012), Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity, *in Proceedings of IIiX '12*, ACM, New York, NY, USA, pp. 254–257. [209]

Wu, W., Li, H. and Xu, J. (2013), Learning Query and Document Similarities from Click-through Bipartite Graph with Metadata, *in Proceedings of WSDM'13*, ACM, New York, NY, USA, pp. 687–696. [177, 180]

Xie, M., Lakshmanan, L. V. and Wood, P. T. (2010), Breaking out of the Box of Recommendations: From Items to Packages, *in Proceedings of RecSys '10*, ACM, New York, NY, USA, pp. 151–158. [54]

Xue, G.-R., Zeng, H.-J., Chen, Z., Yu, Y., Ma, W.-Y., Xi, W. and Fan, W. (2004), Optimizing Web Search Using Web Click-through Data, *in Proceedings of CIKM '04*, ACM, New York, NY, USA, pp. 118–126. [180]

Yilmaz, E., Verma, M., Craswell, N., Radlinski, F. and Bailey, P. (2014), Relevance and Effort: An Analysis of Document Utility, *in Proceedings of CIKM '14*, ACM, New York, NY, USA, pp. 91–100. [26]

Yue, Y., Patel, R. and Roehrig, H. (2010), Beyond Position Bias: Examining Result Attractiveness As a Source of Presentation Bias in Clickthrough Data, *in Proceedings of WWW '10*, ACM, New York, NY, USA, pp. 1011–1018. [64, 207]

Yue, Z., Han, S. and He, D. (2014), Modeling Search Processes Using Hidden States in Collaborative Exploratory Web Search, *in Proceedings of CSCW '14*, ACM, New York, NY, USA, pp. 820–830. [79]

Yujian, L. and Bo, L. (2007), A Normalized Levenshtein Distance Metric, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6), 1091–1095. [183]

Zhai, C., Cohen, W. W. and Lafferty, J. (2015), Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval, *SIGIR Forum* 49(1), 2–9. [151]

Zhou, K., Cummins, R., Lalmas, M. and Jose, J. M. (2012), Evaluating Aggregated Search Pages, *in Proceedings of SIGIR '12*, ACM, New York, NY, USA, pp. 115–124. [77, 79, 80, 163, 206]

Zhou, K., Cummins, R., Lalmas, M. and Jose, J. M. (2013), Which Vertical Search Engines Are Relevant?, *in Proceedings of WWW '13*, ACM, New York, NY, USA, pp. 1557–1568. [80, 164]

Zobel, J. (2018), What We Talk About When We Talk About Information Retrieval, *SIGIR Forum* 51(3). [30]