Temporal Dynamics in Information Retrieval

Stewart William Whiting

School of Computing Science

College of Science and Engineering

University of Glasgow, Scotland, UK.

A thesis submitted for the degree of

Doctor of Philosophy (Ph.D)

October, 2015

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation 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 dissertation is the result of my own work, under the supervision of Professor Joemon M. Jose and Dr Gethin Norman, and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Stewart William Whiting
October, 2015

Abstract

The passage of time is unrelenting. Time is an omnipresent feature of our existence, serving as a context to frame change driven by events and phenomena in our personal lives and social constructs. Accordingly, various elements of time are woven throughout information itself, and information behaviours such as creation, seeking and utilisation.

Time plays a central role in many aspects of information retrieval (IR). It can not only distinguish the interpretation of information, but also profoundly influence the intentions and expectations of users' information seeking activity. Many time-based patterns and trends – namely *temporal dynamics* – are evident in streams of information behaviour by individuals and crowds. A temporal dynamic refers to a periodic regularity, or, a one-off or irregular past, present or future of a particular element (e.g., word, topic or query popularity) – driven by predictable and unpredictable time-based events and phenomena.

Several challenges and opportunities related to temporal dynamics are apparent throughout IR. This thesis explores temporal dynamics from the perspective of query popularity and meaning, and word use and relationships over time. More specifically, the thesis posits that temporal dynamics provide tacit meaning and structure of information and information seeking. As such, temporal dynamics are a 'two-way street' since they must be supported, but also conversely, can be exploited to improve time-aware IR effectiveness.

Real-time temporal dynamics in information seeking must be supported for consistent user satisfaction over time. Uncertainty about what the user expects is a perennial problem for IR systems, further confounded by changes over time. To alleviate this issue, IR systems can: (i) assist the user to submit an effective query (e.g., error-free and descriptive), and (ii) better anticipate what the user is most likely to want in relevance ranking. I first explore methods to help users formulate queries through time-aware query auto-completion, which can suggest both recent and always popular queries. I propose and evaluate novel approaches for time-sensitive query auto-completion, and demonstrate state-of-the-art performance of up to 9.2% improvement above the hard baseline. Notably, I find results are reflected across diverse search scenarios in different languages, confirming the pervasive and language agnostic nature of temporal dynamics. Furthermore, I explore the impact of temporal dynamics on the motives behind users' information seeking, and thus how relevance itself is subject to temporal dynamics. I find that temporal dynamics have a dramatic impact on what users expect over time for a considerable proportion of queries. In particular, I find the most likely meaning of ambiguous queries is affected over short and long-term periods (e.g., hours to months) by several periodic and one-off event temporal dynamics. Additionally, I find that for event-driven multi-faceted queries, relevance can often be inferred by modelling the temporal dynamics of changes in related information.

In addition to real-time temporal dynamics, previously observed temporal dynamics offer a complementary opportunity as a tacit dimension which can be exploited to inform more effective IR systems. IR approaches are typically based on methods which characterise the nature of information through the statistical distributions of words and phrases. In this thesis I look to model and exploit the temporal dimension of the collection, characterised by temporal dynamics, in these established IR approaches. I explore how the temporal dynamic similarity of word and phrase use in a collection can be exploited to infer temporal semantic relationships between the terms. I propose an approach to uncover a query topic's "chronotype" terms – that is, its most distinctive and temporally interdependent terms, based on a mix of temporal and non-temporal evidence. I find exploiting chronotype terms in temporal query expansion leads to significantly improved retrieval performance in several time-based collections.

Temporal dynamics provide both a challenge and an opportunity for IR systems. Overall, the findings presented in this thesis demonstrate that temporal dynamics can be used to derive tacit structure and meaning of information and information behaviour, which is then valuable for improving IR. Hence, time-aware IR systems which take temporal dynamics into account can better satisfy users consistently by anticipating changing user expectations, and maximising retrieval effectiveness over time.

Acknowledgements

I dedicate this thesis to my parents, Richard and Pamela Whiting. Their love, support and encouragement has been unwavering. Without them I would never have been able to follow this opportunity in my life. I can never express quite how truly appreciative I am.

Writing this thesis has been a long journey. I am deeply thankful to all my friends, family and colleagues who have aided me throughout. I would like to thank my siblings – Marc, Rebecca and Emma Whiting – for being there for me, especially during the difficult past two years. My aunty and uncle, Sandy and Brian Talbot, opened the doors that led me down this path of discovery. Their help, guidance and inspiration changed the way I look at the world, and for that I am ever grateful. I will be forever indebted to my first mentor, Wayne Kerridge, who provided the early inspiration and support that allowed me to develop a career in technology.

My colleagues have been a source of much inspiration and comedy over my years as a PhD student. Guido Zuccon and Teerapong Leelanupab helped me to become established when I first started. Ke Zhou, Jesus Rodriguez Perez, Philip McParlane, James McMinn, Horatiu Bota, Rami Alkhawaldeh, Fajie Yuan and Stefan Raue – their company and collaboration has been a pleasure.

I am extremely appreciative of my supervisor, Joemon Jose, for handing me the opportunity to do a PhD funded by the EPSRC DTA scheme. He granted me the freedom to follow many new research ideas, yet provided the counsel when needed. I would also like to take this opportunity to thank my viva examiners, Arjen de Vries and Milad Shokouhi, for their hard work in providing excellent comments and suggestions to improve the thesis.

From the very start of my time as a PhD student, it has been an absolute privilege to have Yashar Moshfeghi as my mentor. He voluntarily took a central role in helping me shape my research ideas and this thesis, and for that I will be forever appreciative.

Special thanks must go to Omar Alonso, who gave me the incredible opportunity to join Microsoft Research in Silicon Valley as an intern in 2012, and again in 2014. Omar and my other supervisors at Microsoft – Aditi (Shubha) Nabar and Alex Dow – mentored me to develop a research and development approach that has laid the foundations of my career.

Finally, I need to thank my partner, Jodie Clarke, for supporting me during this journey. We have gone through this together, and for that I can never quite thank her enough.

The only constant is change:

> *“All is flux, nothing is stationary; no man ever steps in the same river twice, for it is not the same river and he is not the same man.”*

– Plato, around 369 BC.

Contents

List of Figures

List of Tables

Abbreviations

ARIMA	Auto Regressive Integrated Moving Average
AP	Average Precision
AQE	Automatic Query Expansion
HTTP	Hyper Text Transfer Protocol
IDF	Inverse Document Frequency
IA	Intent-aware
IIR	Interactive Information Retrieval
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LM	Language Model
LSA	Latent Semantic Analysis
MAP	Mean Average Precision
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MRR	Mean Reciprocal Rank
NIST	(U.S.) National Institute of Standards and Technology
P@k	Precision at rank cut-off k
PRF	Pseudo-Relevance Feedback
QAC	Query Auto-Completion
QE	Query Expansion
QRel	Query Relevance (i.e., relevant document judgements)
RF	Relevance Feedback
RM	Relevance Model
RMSE	Root Mean Square Error
TDT	Topic Detection and Tracking
TF	Term Frequency
TREC	Text Retrieval Conference
TS	Time Series
TSN	Temporal Semantic Network
TSQE	Temporal Semantic Query Expansion

Part I

Introduction and Background

Chapter 1

Introduction

1.1 Preface

Locating relevant information with an information retrieval (IR) system has become a ubiquitous activity for many users in an era of information overload. A user with an Anomalous State of Knowledge (ASK) can address their information need by describing it to an IR system (Belkin, 1980) – that is, by posing a *query*. However, often the user is unable to precisely specify their information need, since they lack the knowledge to do so. Hence the user's query is often inadequate, leading to uncertainty about the nature of their information need (Sanderson, 2008). The objective of an IR system is therefore to interpret the information need implied by a query, and identify information most expected to satisfy it.

The query provided by the user is an explicit indication of their information need. Other subtle indications of a user's information need may arise during their interactions. For instance, an implicit situation composed of time, space, social and task contexts (Abowd et al., 1999; Ingwersen and Järvelin, 2005) may help frame the underlying motive of their interaction with the IR system – and thus their intentions and expectations. Context is a complex multi-faceted and interdependent concept investigated from many perspectives, making it difficult to study as a whole. Location in particular is closely related to time, as users in different time zones interact with information that is locally relevant during their normal hours. However, time has many generalisable effects less associated to the location context, including time within the same location (e.g., daytime to evening on the USA east coast), and over longer periods internationally, such as from weeks to months and seasons. Accordingly, this thesis isolates and explores the involvement of the time context in IR.

By its very nature, time is a fundamental sense through which humans experience and perceive the world (Gibson, 1975). Indeed, many human memory and cognitive processes are intrinsically temporal (Pöppel, 1978). Because time is such an omnipresent part of our personal lives and social constructs, many elements of time are interwoven throughout information and interaction. Given the impetus of IR to satisfy users trying to find relevant information, time is therefore of fundamental importance.

The importance of time presents itself in many aspects of IR. Time can distinguish the interpretation of information as new topics emerge, and existing topics evolve in a collection. Furthermore, time can also profoundly influence the intentions and expectations of users' information behaviour, such as defining *why* the user is seeking information and thus *what* they expect to find with an IR system. Information behaviour refers broadly to the many ways in which human beings actively create, seek and utilise (e.g., view, share or edit) information (Bates, 2010). Although these behaviours are closely related, this thesis is most concerned with information seeking behaviour.

Time-based elements are present in information and information behaviour in two distinct forms. *Explicitly*, temporal clues are visibly embedded in informational content as *temporal expressions*. These are the time and date references used to anchor topics related to time-based events and phenomena (Alonso et al., 2011). For example, a document may include meta-data stating it was last amended on the 1st August 2013, at 9:10pm. Alternatively, a document may discuss an event that occurred in the past, e.g., "in September 2010", or that will occur in the future with "next October". *Implicitly*, temporal clues arise from the streams of real-time individual and collective user information behaviour within information spaces. For example, work-oriented information needs are likely to be observed during working hours (e.g., 9am to 5pm, Monday to Friday in the Western world), as opposed to leisure-oriented information needs in the evenings and weekends. Indeed, information behaviour is rarely stationary over time. Users write about, search for and express interest in different topics as interests and expectations change over time (Adar et al., 2007). Patterns and trends in these time-based information behaviour streams are termed *temporal dynamics*. Several temporal dynamics arise as a consequence of temporal information behaviour by individuals, collective groups or the world at large. This thesis seeks to elucidate and expand the frontier of research on the involvement of time in IR (or, *time-aware IR*) by exploring the challenges and opportunities of temporal dynamics.

Temporal dynamics are evident in many of the mainstays of IR, including word, topic, and query occurrence, and indeed, relevance. Temporal dynamics often comprise predictable periodicity and trends, unpredictable surprises and indeed, complex temporal interactions between these elements. These temporal dynamics typically reflect underlying change in interpretation and meaning, such as topics evolving, or external factors (e.g., events) shifting the focus of attention. Despite this, the majority of work in IR has disregarded the involvement of temporal dynamics, instead treating concepts as stationary over time. In this thesis I explore how considering temporal dynamics in IR leads to many opportunities for improving IR system effectiveness, and ultimately, better satisfying users.

1.2 Thesis Statement

This thesis states that temporal dynamics provide tacit meaning and structure of information, and information seeking. Temporal dynamics are a ‘two-way street’ insofar as they must be supported, and conversely, can be exploited. As such, the conjectures set forth by this thesis are two-fold:

- On one hand, current temporal dynamics pose a challenge for IR systems – real-time information seeking must be *supported* for consistent user satisfaction.
- On the other hand, past temporal dynamics offer a complementary opportunity – as a tacit dimension which can be *exploited* to inform more effective IR systems.

The first conjecture relates to supporting users consistently in their activities over time. Since users turn to IR systems to satisfy time-sensitive information needs, an IR system must accommodate the real-time temporal dynamics of changing user needs and expectations. Uncertainty about what the user wants is a perennial problem for IR systems, further confounded by changes over time. To alleviate this issue, an IR system can do two things: (i) assist the user to submit an effective query (e.g., error-free and descriptive), and (ii) better anticipate what the user is most *likely* to want. In this thesis I explore methods to help users formulate queries through time-aware query auto-completion, which can suggest both recent and always popular queries. This approach assists users to pose effective queries consistently over time. Furthermore, I explore the impact of temporal dynamics on the motives behind users’ information seeking, and thus how relevance itself is subject to temporal dynamics. This work motivates the need for time-aware retrieval models which take into consideration past and present temporal dynamics of user motives, to anticipate temporally relevant results.

The second conjecture relates to exploiting previously observed temporal dynamics to inform more effective IR systems. Several temporal dynamics driven by past information behaviour are evident, such as word and phrase use in a time-based collection. Many IR approaches are based on methods which typically characterise the nature of the information through the statistical distributions of words and phrases. In this thesis I look to exploit temporal structure – derived from temporal dynamics – in established IR approaches. I explore how the temporal dynamic similarity of word and phrase use can be used to infer temporal semantic relationships between them. Based on this, I propose a model for enhancing retrieval performance using the temporal semantics uncovered for any query topic, and significantly improve retrieval performance.

This thesis is investigated through three broad research questions, outlined in Section 1.4. In the next section I motivate the foundation of the thesis and research questions.

1.3 Motivation

This thesis is motivated by many factors in society and IR research, which I discuss in the following section. Together, these necessitate and facilitate time-aware IR, and in particular, considering temporal dynamics in IR.

The Rise of Real-time Information Behaviour

The ease of using online IR systems, such as web search engines, has meant that users increasingly turn to them to instantly satisfy all manner of real-time information needs throughout their lives. The prevalence of mobile computing with devices such as smart phones, tablets and watches further facilitates real-time information seeking as users always have IR systems instantly to hand (Tsai et al., 2010). The information needs served by web search engines often relate to common daily tasks such as work, leisure, travel and shopping (Broder, 2002). Many temporal trends and patterns are observed in user's every day tasks and behaviour, and so, are similarly found in accompanying information needs. Indeed, time is highly likely to be even more important in the near future as users become increasingly attached to real-time information services in all aspects of their life through future ubiquitous computing devices. For industrial search engines, satisfying user expectations is paramount – commercial success is pinned on consistently providing quick and satisfactory results, all the time. Consequently, taking the time context into consideration for both efficiency and effectiveness of a search engine is highly valuable. With this in mind, in this thesis I propose several approaches to support increasingly prevalent real-time information behaviour.

As well as supporting daily tasks, the web has become a widespread platform for real-time news and media sources (Teevan et al., 2011). A substantial proportion of day-to-day information needs relate to breaking or ongoing events (Adar et al., 2007; Kairam et al., 2013; Kulkarni et al., 2011). A vast amount of information streams constantly from both established sources, and user-generated content such as blogs and social media. Information creators are as much influenced by time as information seekers, as their attention is on the present. As a result, new information reflects recent and ongoing events and culture as authors write about what is currently interesting. In reciprocation, users are looking to engage with this novel content, and related information. An effective search engine must make sense of this new information and effectively support associated information seeking.

With time becoming a pervasive part of information access, there is a prolific increase in information archives such as news and web snapshots[1] which capture the historic impacts of real-time information and interaction. These archive collections are growing to cover longer periods of time, capturing the flow of past events and phenomena. Time therefore serves as an increasingly important dimension to organise and characterise the meaning of older information during retrieval from these archives, in addition to providing a basis for real-time, recent information retrieval.

Directions in Time-aware IR

Time has long been established as a key element of information and interaction. In the earliest theoretical formulation of the problem of IR, Mooers (1952) viewed the sequential communication between user and system as *temporal signalling*. While this view of time offers a foundation, time has a broad definition and consequently diverse influences on IR, which I explore further in the time-aware IR background presented in Chapter 3. Time-aware IR has sought to address temporal challenges, such as time-based change and meaning in information and information behaviour, apparent for IR in many scenarios (Alonso et al., 2011; Campos et al., 2014; Radinsky et al., 2013b). Although the importance of time in IR has become increasingly apparent, conventional IR models and approaches have concentrated on static collections and information needs. This stationarity assumption is unrealistic in many scenarios, e.g., the web, where temporal change and evolution is ever present; the world does not stand still. While an IR approach may be considered effective based on a static evaluation, it may not necessarily be consistently effective over time as user expectations and the collection change. Consequently, in this thesis I challenge the stationarity assumption by explicitly incorporating temporal information into retrieval approaches.

Temporal dynamics of past information behaviour can offer an additional dimension of insight far beyond statistical distributions averaged over large periods of time (Alfonseca et al., 2009; Efron, 2010; Liebscher and Belew, 2003; Shokouhi and Radinsky, 2012). Temporal dynamics have been used to uncover structure in many other fields involving human behaviour, including financial markets, neurology and psychology (Fu, 2011), but they have not been deeply explored in much of IR – despite their huge potential for characterising the nature of human information behaviour. Indeed, much of IR has disregarded temporal dynamics because of inherent complexity in modelling, processing and evaluation. One of the key aims of this thesis is to explore possible solutions to these challenges.

[1] Such as the Web Archive project, `http://www.archive.org`

Time as an Infinite Dimension

Time is in essence a further dimension of information and interaction. Since time is a continuous stochastic process, it comprises a high dimensional hierarchical space within which any number of time periods and durations can be constructed. Temporal change, or even lack thereof, characterised over infinitely possible time periods offers a multitude of opportunities to derive structure and meaning of the world in terms of modelling changing information behaviour through language, relationships, topics and user interests over time. With this in mind, in this thesis I propose several approaches to exploit the temporal dimension present in many time-based collections to better characterise information, and ultimately improve retrieval effectiveness.

A great deal of existing IR approaches use features based on statistical distributions to infer structure and meaning, for instance, the power-law frequency distribution of words in language observed by Zipf (1949). Intuitively, these statistical distributions differ over time, and so methods reliant on them should not only take the changes into account, but also take the opportunity to exploit the changes.

The overwhelming majority of research in IR has concentrated on the English language. Many of the techniques developed may require further work to apply them effectively to other languages, since they rely on features, syntax or statistics intrinsic to certain languages or cultures. Time abstractly transcends human behaviour and information, regardless of the concrete nuances of language and syntax. Consequently, time offers a independent dimension which can be uniformly applied regardless of language. Indeed, results reported in Chapter 4 demonstrate the opportunity to develop techniques based on time in one language (i.e., English), and apply them to another language (i.e., Chinese).

Advances in Large-scale Data Processing

Time adds an additional dimension to information. In fact, the greater the granularity of time considered, the more dimensionality is increased. Every time period modelled adds a further dimension to what are typically already high-dimensional spaces, e.g., unique words and the co-occurrences between them in a large collection. This explosion in data greatly increases data storage and processing complexity.

Recent advances in distributed systems offer near linear horizontal scaling to accommodate vast data storage and computational challenges with commodity hardware. Such systems alleviate the practical challenges of working with temporal dynamics in IR, and make non-trivial temporal approaches characterising large-scale change over short- and long-term periods computationally feasible in both real-time and batch computing scenarios.

1.4 Research Questions

The thesis outlined in the previous section is addressed by the following three high-level research questions relating to supporting, or conversely, exploiting temporal dynamics. Because this thesis covers broad themes, each research question is prefaced by a motivating setting.

The first two research questions relate to supporting temporal dynamics in real-time information seeking:

- **RQ1:** Query popularity exhibits many patterns and trends over time as common information needs change. *Can these temporal dynamics be supported for consistently effective query auto-completion over time?*
  (This research question is investigated in Chapter 4 on page 58)

- **RQ2:** Query intent exhibits many patterns and trends over time as collective influences and expectations change. *To what extent do query intents change over time, and, can these temporal dynamics be supported?*
  (This research question is investigated in Chapter 5 on page 92)

The final research question relates to exploiting past temporal dynamics captured in time-based information collections:

- **RQ3:** Term popularity exhibits many patterns and trends over time in a time-based document collection. *Can these temporal dynamics be exploited to improve IR system effectiveness over time?*
  (This research question is investigated in Chapter 6 on page 128)

In addition to these broad high-level research questions, in each chapter I further outline four research sub-questions. Each sub-question is concerned with the specific challenges and approaches of the work presented in the chapter, but is discussed and concluded in the context of the respective high-level research question.

1.5 Outline

This thesis explores how temporal dynamics must be supported, and conversely, can be exploited for improving the effectiveness of IR systems. The thesis comprises four parts:

Part I: Introduction and Background

In this part, I introduce and explain the themes addressed throughout this thesis. In Chapter 1, I begin by introducing the broad involvement of time in information and information behaviour. I motivate the need to exploit and support temporal dynamics, and proceed to outline a thesis with two conjectures, and subsequently pose three high-level research questions. In Chapter 2, I provide a general background of IR covering established aspects such as retrieval tasks, models and evaluation techniques. This is followed with a comprehensive background of time-aware IR in Chapter 3. As part of this, I propose a novel map of time-aware IR which serves to conceptualise the involvement of time in *all* aspects of IR. To this end, I examine how elements of time are of interest to information retrieval, and review existing literature to motivate the work presented in later parts of this thesis. Hence, this chapter serves to frame the research presented throughout this thesis.

Part II: Supporting Temporal Dynamics in Information Behaviour

In this part I focus on how time affects user needs and expectations during information behaviour. To this end, I explore methods to better support the real-time temporal dynamics of users' information seeking behaviour.

The fundamental action performed by all users during information seeking is query input, which is a cognitively and physically laborious process. Query auto-completion (or similarly, 'auto-suggest'), aims to alleviate the effort required to formulate each query by offering completed queries the user may wish to submit while they are typing. Since query popularity is affected by several temporal dynamics, query auto-completion must be sensitive to changes in query popularity. In Chapter 4, I explore methods to improve the performance of query auto-completion over time. In addition to query popularity, relevant items for a query may vary as the intention underlying an information need changes over time. In Chapter 5, I study the temporal dynamics of ambiguous and multi-faceted query intents, and consider existing diversity-aware IR frameworks for modelling temporal relevance.

Part III: Exploiting Temporal Dynamics in Collections

In the previous part of this thesis, I explored methods to support real-time temporal dynamics. In this part, I consider the ability of past temporal dynamics to characterise tacit meaning and structure of information contained in a time-based collection.

To this end, I explore methods to exploit the temporal dynamics of word and phrase use in time-based collections to improve IR system effectiveness. In Chapter 6, I begin by proposing an approach for identifying a topic's *chronotype* – that is, the topic-specific temporal semantic relationships betweens words that comprise the topic, as reflected by their temporal dynamic similarity. I further exploit the strongly temporally related terms uncovered by this method for query expansion to improve IR effectiveness.

Part IV: Conclusion

In this final part, I draw the thesis content to a close. In Chapter 7, I summarise the contributions made by this thesis. I discuss the three high-level research questions posed, and conclude the thesis. Furthermore, I make several recommendations for future directions relating to considering temporal dynamics in time-aware IR.

1.6 Publications

Portions of the research presented in this doctoral thesis is included in the following selected peer-reviewed publications (in chronological order):

- **Temporal Dynamics of Ambiguous Queries.** S. Whiting, O. Alonso and J.M. Jose. TAIA 2015 Workshop, SIGIR 2015, Santiago, Chile. (Short Paper)

- **Recent and Robust Query Auto-Completion.** S. Whiting and J.M. Jose. WWW 2014. Seoul, South Korea. (Full Paper)

- **Wikipedia as a Time Machine.** S. Whiting, J.M. Jose and O. Alonso. TempWeb Workshop, WWW 2014. Seoul, South Korea. (Short Paper)

- **Temporal Variance of Intents in Multi-faceted Event-driven Information Needs.** S. Whiting, Z. Ke, J.M. Jose and Lalmas M. SIGIR 2013, Dublin, Ireland. (Short Paper)

- **The Impact of Temporal Intent Variability on Diversity Evaluation.** Z. Ke, S. Whiting, J.M. Jose and Lalmas M. ECIR 2013, Moscow, Russia. (Poster Paper)

- **CrowdTiles: Presenting Crowd-based Information for Event-driven Information Needs.** S. Whiting, K. Zhou, J.M. Jose, O. Alonso and T. Leelanupab. CIKM 2012, Maui, Hawaii. (Demo Paper)

- **Hashtags as Milestones in Time.** S. Whiting and O. Alonso. TAIA 2012 Workshop, SIGIR 2012, Portland, Oregon. (Short Paper)

- **The Essence of Time: Considering Temporal Relevance as an Intent-aware Ranking Problem.** S. Whiting. SIGIR 2012, Portland, Oregon. (Doctoral Consortium Paper)

- **Temporal Pseudo-relevance Feedback in Microblog Retrieval.** S. Whiting, I. Klampanos, and J.M. Jose. ECIR 2012, Barcelona, Spain. (Short Paper)

- **Exploring Term Temporality for Pseudo-relevance Feedback.** S. Whiting, Y. Moshfeghi, and J.M. Jose, SIGIR 2011, Beijing, China. (Poster Paper)

A full up-to-date record of all publications is available online[1]. Several publications including work contained in this thesis are forthcoming at the time of thesis completion.

[1] `http://www.stewh.com/research/publications/`

Chapter 2

General IR Background

2.1 Introduction

In this chapter I provide a general background on the fundamental concepts of information retrieval (IR), along with definitions that are required to understand the topics explored later in this thesis. Subjects covered include basic information retrieval concepts, models, and evaluation methodologies. In the next chapter I provide a background specific to time-aware IR, re-examining many of the concepts presented in this chapter with respect to time.

2.1.1 Chapter Outline

This chapter is organised as follows:

- Section 2.2 introduces types of IR tasks performed by users.
- Section 2.3 discusses the principles of established retrieval models, and proceed to formally define each one.
- Section 2.4 outlines the experimental methodologies typically employed in IR research and development.
- Section 2.5 details common evaluation metrics used in system-oriented retrieval effectiveness evaluation.

2.2 Information Retrieval Tasks

Users interact with IR systems in many ways, with different goals in mind. Various models of user behaviour during retrieval tasks have been explored in IR literature (Borlund, 2003;

Hearst, 2009). These fundamental user interaction models motivate retrieval approaches and evaluation methodologies which lie at the heart of IR research and development.

The majority of IR research has concentrated on developing and evaluating retrieval systems to support *ad hoc* retrieval tasks. In an ad hoc retrieval setting, users state their information need as an isolated query, and the IR system provides a ranked list of results. Consequently, an ad hoc query is based on a one-time interaction with the IR system – the user is not expected to further clarify or explore their information need. Ad hoc retrieval is the focus of retrieval experiments presented in this thesis.

However, this traditional view of user interaction has been challenged by the realisation that users engage with IR systems to answer complex questions, and express a developing information need through multiple queries during a *session*. The user may be uncertain of what they are seeking, and there may not even be a factually correct answer. These search tasks are naturally more *exploratory* in nature, and the search task itself is often a learning process for the user (White and Roth, 2009). Time undoubtedly plays various roles in these scenarios, however evaluating these types of tasks is problematic, let alone with an added temporal dimension, so it is left to future work.

2.3 Retrieval Models

The primary objective of an IR system is to identify information most likely to satisfy the user's information need. To achieve this goal, an understanding of how humans comprehend and assess information is necessary. This requires an intimate knowledge of the cognitive structures and processes which are responsible for information processing and decision making tasks in the human brain; however, as yet we have not formalised these. Instead, IR researchers propose hypotheses that encapsulate their beliefs about the nature of relevance. These beliefs are operationalised for experimentation by encoding them in a mathematical model, named a *retrieval model*. A retrieval model specifies a framework to: (i) represent the user's information need (i.e., their query), (ii) represent information (e.g., documents), and (iii) match information to the query to identify relevant items. More elaborate models also include a means to estimate the degree of relevance. Relevance estimation allows results to be ranked by their expected ability to satisfy the user – with most relevant results presented first to minimise user effort, thereby increasing overall IR system effectiveness.

Retrieval models are at the centre of a great deal of past and present IR research. In the following sections I present three well-known retrieval model families, each based on different principles and formulations of the problem of IR.

2.3.1 Boolean Model

One of the first retrieval models to be proposed and investigated in IR was the Boolean Model van Rijsbergen (1979). The Boolean Model is based conceptually on set theory, with each document represented as a binary set of its contained terms. No notion of term importance in the query, document or collection is assigned. A query composed of Boolean logic is used to express relevant document expectations. Operators such as `AND`, `NOT` and `OR` provide discriminators which either require or exclude the presence of query terms in relevant documents. Since the retrieval model is based on logic conditions, documents have binary relevance – either: `TRUE` or `FALSE`. There is no estimation of the degree of relevance in this model, so results are provided as an unranked set. Query independent document features, such as age, can be used to induce a result ranking.

Queries based exclusively on boolean logic are a relatively unnatural and difficult means to express complex information needs, especially when the user is uncertain of their precise information need. Moreover, particularly in larger collections, boolean queries can become unwieldy because a large number of conjunctions and disjunctions are needed to sufficiently specify the information need, and reduce the result set size. At the same time, an overly specific query may result in high precision yet very low recall of results. While boolean retrieval models alone are not common place in modern applications, boolean operators are still included as an advanced querying feature for use when good results are difficult to obtain (for example, they can be used when there is ambiguity to exclude unwanted interpretations).

2.3.2 Vector Space Model

Addressing the limitations of the Boolean Model, Salton et al. (1975) propose the popular Vector Space Model (VSM). In contrast to the Boolean Model, VSM supports partial matching and incorporates relevance estimation in result ranking. VSM is based on Euclidean geometry, viewing queries and documents as vectors in a multi-dimensional space where relevance is characterised as the ‘closeness’ between vectors. Aside from performing well empirically, VSM is popular because it provides both an intuitive interpretation of the problem of IR, and readily accommodates more elaborate IR methods such as term weighting and relevance feedback which lead to greater retrieval effectiveness (Croft et al., 2010).

In the VSM approach, a document D_i is represented as a *t*-dimensional vector of terms, such that $D_i = (d_{i1}, d_{i2}, \dots d_{it})$. Similarly, query Q is also represented as a vector in the same *t*-dimensional space, such that $Q = (q_1, q_2, \dots q_t)$. In its simplest form, d_{ij} is the raw frequency (i.e., TF) of the term appearing in the document. Since each term is a dimension in the VSM, a term weighting scheme can be employed to emphasise more significant or discriminative terms. Several variants of term weighting have been explored. Normalization of the term frequency by document length is commonly used to incorporate the importance of the particular term in variable length documents. Proposed by Sparck-Jones (1972), inverse document frequency (IDF) quantifies the information carried by a term. IDF measures term specificity as its ability to discriminate relevant from non-relevant documents, based on the proportion of documents containing that term in the collection. If a term exists in only a few documents then it is highly valuable for retrieval. In contrast, a very common term such as “the” will exist in almost every document in an English language collection, and thus serve little discriminative purpose during retrieval. Document length normalised TF and IDF are normally combined together as measure of both term appearance and importance (i.e., TF-IDF). TF-IDF for term k is calculated as follows:

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

Where $f_{i,k}$ is the frequency of the term, N is the number of documents in the collections and n_k is the number of documents within which term k appears.

With TF-IDF vectors representing terms in a query and documents, vector similarity measures document relevance. Although many approaches for comparing document and query vector similarity have been proposed, empirical evidence has favoured cosine similarity (Croft et al., 2010). Cosine similarity between document D_i and query Q is defined as follows:

$$\mathrm{sim}(d_i, q) = \text{cosine}\,\theta_{\mathbf{d_i},\mathbf{q}} = \frac{\mathbf{d_i} \bullet \mathbf{q}}{\|\mathbf{d_i}\|\,\|\mathbf{q}\|} = \frac{\sum_{j=1}^{t} d_{i,j} \times q_j}{\sqrt{\sum_{j=1}^{t} {d_{i,j}}^2 \times \sum_{j=1}^{t} {q_j}^2}}$$

2.3.3 Probabilistic Models

Probability is concerned with modelling the frequency of uncertain events occurring. In an IR context, probability provides an elegant mathematical framework within which to formally model the relevance of a document to a user’s query, given uncertainty about the user’s precise information need. Probabilistic retrieval models are grounded by the Probability Ranking Principle (PRP) (Robertson, 1997), which states documents that are more likely relevant than non-relevant should be retrieved (i.e., where $P(R|D) > P(\bar{R}|D)$), and further, ranking of retrieved documents should be by their likelihood of relevance. While the PRP provides the foundation of the probabilistic IR approach, it does not give any concrete detail on how to implement such a probabilistic model.

Early probabilistic retrieval models viewed IR as a classification problem, where documents can be classified as either "relevant" or "non-relevant" to a given query. Bayes' theorem[1] is employed to define the probability of document relevance based on the likelihood of drawing query terms from relevant and non-relevant documents. This likelihood can be computed using various models, most notably Okapi BM25 (Robertson et al., 1996) which incorporates document and query term weights. BM25 has had considerable impact in IR due to its empirical effectiveness in many scenarios.

While these traditional probabilistic retrieval models have performed well empirically, their interpretation is loose and not always theoretically principled. As a result, adapting these models to accommodate more elaborate approaches to retrieval is not always intuitive (Hiemstra and Vries, 2000). This has led to the adoption of more formal statistical language modelling approaches in IR (Lavrenko and Croft, 2001). Language modelling has strong a theoretical motivation rooted in the principles of language, as defined in fields such as natural language processing and speech processing. I employ language models for retrieval experiments presented later in this thesis, so I detail their principles in the following section.

2.3.4 Language Models

Unigram language models are the simplest form of language model, modelling individual words (i.e., unigrams) as a probability distribution over the words occurring in language. In essence, a language model represents the probability of observing any given word in a document, query or collection. In IR, a language model can be used to express the topics contained in a document or query. In the case of documents, the more about a topic a document is, the more likely it would be to observe words about that topic in the document. When a word is not observed in a document, smoothing techniques such as the Jelinek-Mercer method estimate a non-zero likelihood the word could have occurred, given the nature of the collection. This avoids zero probability issues when calculating joint probability of terms, allowing partial matching of queries where not all terms appear in the document.

Retrieval based upon language models can be formulated in one of three ways: (i) the probability of generating the document from a query language model (i.e., *Document Likelihood Model*), (ii) the probability of generating the query from a document language model (i.e., *Query Likelihood Model*), or (iii) comparison between the distribution of the document and query, or, *relevance* language models (i.e., *Relevance Model*) (Croft et al., 2010). I detail the Relevance Model interpretation here since it is used experiments presented later in this thesis.

[1] Formally: $P(R|D) = P(D|R)P(R)/P(D)$

The Relevance Model retrieval approach proposed by Lavrenko and Croft (2001) estimates the language model expected to be found in relevant documents (referred to as the *relevance model*), and employs a measure such as the Kullback-Leibler Divergence (referred to as KL divergence) to compute the closeness between document and relevance model distributions. Documents with a language model most similar to the relevance model are deemed more relevant since they contain greater coverage of the relevant topic. KL divergence between two probability distributions, P and Q, is defined as follows:

$$\mathrm{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.1}$$

Of course, the first challenge for this approach is how to derive the relevance model in terms of the probability of words appearing in relevant documents, i.e., $P(w|R)$. In the absence of training data, i.e., known relevant documents, heuristics can be used to approximate the relevance model based on the small amount of evidence provided in the user's query. Alternatively, pseudo-relevant feedback (PRF) can be used to obtain a sample of assumed relevant documents from an initial retrieval using a standard query likelihood model. Typically the top-k (where k is typically in the range 5-30) relevant documents are assumed to be relevant, or at least strongly indicative of relevance, for a query. These documents are used as the training data necessary to directly estimate $P(w|R)$. With this sampled relevance model, KL divergence can be computed to produce a relevance ranking of documents using the equivalent formula:

$$\sum_w P(w|R) \text{log} P(w|D) \tag{2.2}$$

PRF can be problematic as it is very sensitive to a reliable initial retrieval (Carpineto and Romano, 2012). If the initial retrieval is poor, noisy feedback is likely to cause topic drift and further harm retrieval performance. The Relevance Model provides state-of-the-art PRF performance, so I use it as a baseline to further explore temporal PRF approaches in Chapter 6.

2.4 Experimental Methodologies

Evaluation of IR systems has been a controversial topic since the earliest development of retrieval models in the 1960s. Evaluation can be performed using various quantitative or qualitative techniques, which typically fall under two broad scopes. *Efficiency* evaluation is concerned with measuring operational metric such as speed, time and storage requirements.

Meanwhile, *effectiveness* evaluation measures the ability of the system to satisfy user goals during retrieval – for instance, retrieving *only* the most relevant documents, or retrieving *all* relevant documents. Of course, user goals may vary between different tasks (e.g., web search, biomedical search, news search and patent retrieval), so reasonable evaluation must be based on the nature of the desired task. This thesis is primarily concerned with measuring IR system effectiveness. However, many of approaches I propose are done so with a nod towards efficiency, such that they are practical given current computational constraints.

Generally, methods for evaluating IR system effectiveness can be categorised as being either *user-oriented* or *system-oriented* (Voorhees and Harman, 2005). I discuss these evaluation paradigms in the following subsections.

2.4.1 System-oriented Evaluation

System-oriented evaluation is a laboratory-based technique for measuring the effectiveness of IR systems without needing real users at the time of the experiment. Rather than studying individual users and their interactions, system-oriented evaluation is primarily concerned with the ability of an IR system to provide known relevant results for a given test query in an ad-hoc search scenario. As such, system-oriented evaluation requires the creation of a test collection, along with a set of test topics (or, *sample queries*) and associated relevance judgements. After the system provides a result ranking for each query, its performance is characterised (e.g., the presence and rank of relevant and non-relevant documents) using metrics outlined in the following section. These metrics facilitate easy effectiveness comparison between different approaches, parameters and systems.

The Cranfield Studies in the 1960s (Voorhees and Harman, 2005) first highlighted the importance of such test collections as instrumental in scientifically thorough evaluation and development of retrieval models. The Cranfield Evaluation Paradigm, as it is known, is the dominant evaluation technique underpinning much of the evaluation performed in Text REtrieval Conference (TREC) tracks, organised by the National Institute of Technology (NIST). TREC has co-ordinated the construction of open test collections, query samples and relevance judgements for diverse retrieval scenarios, including web, legal, microblog and entity search. The main motivation behind the paradigm is that it facilitates comparative benchmarking of algorithms and techniques, with the scalability and repeatability of the lab environment (Croft et al., 2010) – unlike user studies, which I describe in the following section. This thesis relies extensively on system-oriented evaluation techniques.

While this lab-based approach has undoubtedly been indispensable for the development of IR as a scientifically rigorous discipline, it has received much criticism for its arguably sterile assumptions and overly simplistic view of user information needs. One of the key issues is the highly contended notion of relevance. NIST-employed human assessors evaluate documents based on their perceived relevance given the query topic, yet, relevance is undoubtedly a judgement that is based on complex individual user and context factors such as prior experience, domain expertise, recent information interaction, learning and indeed, time. While collections composed to time-stamped information objects (e.g., tweets and news) exist, there are currently no test collections accompanied by repeatedly collected temporal relevance judgements simulating temporal information needs or queries submitted over time. However, at the time of writing, the NTCIR[1] 12th Evaluation Conference is developing the new "Temporalia" track with some of these goals in mind.

2.4.2 User-oriented Evaluation

For many users, overall user satisfaction is derived not only from the quality of the results, but also increasingly important is the interface which supports the user's interactions. User-oriented evaluation involves direct and measured studies of real users interacting with an IR system, in order to observe the combination of factors that might affect satisfaction.

Studies of this type are concerned with the user satisfaction during typically exploratory interactive IR (IIR) tasks performed using the system. Experiments are often based on providing users complex question answering tasks to complete, or a simulated situation such as a tourist looking for information on where to visit in a city (Borlund, 2003). Questionnaires are used to elicit user opinions and perception, and logged interactions such as result clicks and dwell time can be used to observe and quantify underlying behaviours. User-oriented evaluation offers an avenue for high-quality objective and subjective evaluation, however user studies are expensive and therefore difficult to scale. Zuccon et al. (2013) recommend hybrid evaluation approaches mixing system-oriented evaluation with cheap crowdsourced and lab-based user studies as a possible solution to cost-effective but comprehensive IR system evaluation.

2.5 Evaluating Retrieval Effectiveness

System-oriented evaluation relies on various metrics of retrieval effectiveness to quantify the satisfaction expected to be experienced by a real user. Satisfaction is usually defined in terms of how many relevant results the system provides, and whether those results are highly ranked.

Different effectiveness metrics are helpful to characterise retrieval effectiveness in scenarios where user goals may differ. The most elementary family of retrieval metrics are set-based.

[1] A similar evaluation conference to the TREC series, but based in Asia.

These measures quantify only the existence of relevant results in the top k results (denoted as *measure@*k) retrieved by an IR system. I examine two common set-based retrieval effectiveness metrics in the following subsections.

2.5.1 Recall

Recall measures the ratio of retrieved relevant documents to all known relevant documents for a query. A high recall is desirable in tasks where the user wishes to see all relevant items, for example, in patent retrieval. Recall is formally defined as:

$$\text{Recall} = \frac{\mid \text{relevant documents retrieved} \mid}{\mid \text{relevant documents} \mid} \tag{2.3}$$

2.5.2 Precision

Precision measures the fraction of retrieved documents known to be relevant. High precision is desirable in tasks would rather see only relevant items, for example, web search. Precision is formally defined as:

$$\text{Precision} = \frac{\mid \text{relevant documents retrieved} \mid}{\mid \text{retrieved documents} \mid} \tag{2.4}$$

In reality, a trade-off between recall and precision is often necessary. Increasing recall will inevitably lead to reduced precision as non-relevant results will be introduced to the result ranking, and vice-versa. An understanding of the user's task and goals govern the emphasis on either objective of this trade-off choice.

Of course, the majority of modern retrieval systems produce a ranked list of results to ensure that users see most relevant results first. In ranked retrieval, as well as the presence of relevant results in the first few pages of results, higher ranking of the most relevant results is also desirable. Ranked effectiveness metrics higher weight the presence of highly relevant results in the highest ranking positions. I examine two common ranked effectiveness metrics in the following subsections.

2.5.3 Mean Average Precision (MAP)

MAP offers a single metric to quantify the all-round effectiveness of a retrieval system for a representative set of test topics (e.g., the queries cover diverse popular and unpopular information needs) (Voorhees and Harman, 2005). By evaluating an IR system using a selection of queries with different characteristics, a more realistic view of IR system effectiveness can be obtained. I use MAP to characterise retrieval effectiveness in retrieval experiments reported in Chapter 6.

As its name implies, MAP is computed as the mean over the average precision (AP) obtained for each test topic. AP is a rank-sensitive measure of precision, computed as the average of the precision values obtained at each rank, up to the rank cut-off position k:

$$\mathrm{AP} = \frac{\sum_{k=1}^{n}(P(k) \times \mathrm{rel}(k))}{\mid \text{number of relevant documents} \mid} \tag{2.5}$$

Where $P(k)$ is the precision computed at rank cut-off k, and $rel(k)$ is an indicator function equal to 1 if the item retrieved at rank k is relevant, or 0 if non-relevant.

The AP computed for each sample query q is averaged over the set of all test topics Q to produce the final Mean Average Precision (MAP) for the retrieval system:

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP(q)}}{\mid Q \mid} \tag{2.6}$$

2.5.4 Chapter Summary

In this chapter I introduced the fundamental aspects of IR. I described the common types of IR tasks performed by users, and how these provide the foundation for IR research and development. I briefly summarised the principles of common retrieval model families, namely boolean, vector space, traditional probabilistic and language model approaches. Following this, I outlined the system- and user-oriented experimental methodologies typically employed in IR research and development to quantify user satisfaction, and associated retrieval system effectiveness. Finally, I detailed standard set-based and ranked effectiveness evaluation metrics used to characterise retrieval performance.

In the next chapter I provide a background and motivation on the involvement of time in IR. As part of this, I examine many of the fundamentals of IR presented in this chapter again, with respect to time.

Chapter 3

Time-aware IR Background and Motivation

3.1 Introduction

Broad time-based factors have been explored from many perspectives in information retrieval (IR), establishing the research field of *time-aware IR*. This thesis examines methods for incorporating temporal dynamics into time-aware IR approaches. In the past chapter I introduced the fundamental concepts, models and evaluation strategies underlying established non-temporal IR approaches. In this chapter, I focus on the time-based elements of IR by presenting a novel conceptual map of time-aware IR, followed by an extensive background of related user factors and existing approaches. Consequently, this chapter serves to specifically frame and motivate the philosophical foundations of time in time-aware IR, as well as the concrete research questions and approaches presented in this thesis.

3.1.1 Chapter Outline

This chapter is organised as follows:

- Section 3.2 considers how events and phenomena define time and impact information behaviour.

- Section 3.3 outlines the nature of temporal clues found in information and information behaviour.

- Section 3.4 proposes a conceptual map of time-aware IR. Applying this model as a means to organise the diverse body of relevant theoretical and practical literature, in Sections 3.5 to 3.11, I present an extensive review of time-aware IR research. As part of this, I discuss and illustrate several real collection-based (e.g., word usage popularity) and information seeking (e.g., query and query intent popularity) temporal dynamics that are explored later in this thesis.

- Section 3.12 briefly examines the applications of time in the general field of information management.
- Section 3.13 describes the efficient but flexible representation of temporal dynamics employed for processing by approaches presented later in this thesis.

3.2 Temporal Events and Phenomena

In essence, time is a measure of change. Personal and socially-driven events and phenomena drive change, and therefore delineate time in a dynamic world. Events are loosely defined as a significant happening or occurrence, delimited in scope by time and place with an identifiable set of participants (Allan et al., 1998). At a personal level, an event may be as trivial as travelling to work, or going on holiday. Events include mainstream news, sports and politics, which may be of collective interest to those within particular social, subject or regional crowd. I define phenomena as general changes transcending time that cannot be attributed to any one specific cause, such as drifts in society and culture (e.g., fashion, social norms and religious beliefs). Past, present and future events and phenomena are often discussed as topics in informational content. Meanwhile, current events and phenomena influence real-time information behaviour in terms of what authors write about, and the topics users seek to find out about. Together, these elements of time are woven throughout information and interaction, and therefore comprise various temporal information clues which I detail in the following section.

3.3 Temporal Information Clues

Many time-based clues can provide insight into the temporal meaning and structure of information and information seeking. Temporal information clues may be explicitly embedded in informational content and its meta-data, such as item creation timestamps. Alternatively, several implict temporal information clues may arise from the time-based patterns and trends underlying individual and collective information behaviour. In the following two subsections I elaborate on implicit and explicit temporal information clues.

3.3.1 Explicit Temporal Clues: Temporal Expressions

Explicit temporal clues are embedded by authors within information. Most typically, these are time and date references denoted by *temporal expressions*. Temporal expressions are used by authors to time anchor past, present and future events and topics (Alonso et al., 2011). Temporal expressions can take many forms, categorised as either *absolute* or *relative*.

An *absolute* temporal expression refers to a precise moment or period in time. For example, an exact time or date can be referenced with expressions such as "`1st May 2015`" or "`1/1/2012 1:12pm`", etc. Similarly, a time period can be referenced using expressions such as "`May 2013 to July 2014`" or "`during 2014`", etc. Note that time zones and cultural date formatting must be taken into account when temporal expressions are extracted in different languages, as would be the case for international web documents (Alonso et al., 2011).

A *relative* temporal expression serves as an offset to a known absolute temporal expression – typically the document creation time, or a previously mentioned absolute temporal expression contained in the text. Relative temporal expressions include "`last year`", "`the following month`" and "`yesterday`", etc., as well as named temporal expressions (Brucato et al., 2013) such as "`Christmas Day`" and "`New Year's Eve`". Relative temporal expressions must be normalised into absolute temporal expressions for further computation (Alonso et al., 2011). As part of the Semantic Evaluation (SemEval) evaluation exercises, the TempEval (Uzzaman et al., 2013) tasks facilitated benchmarking of heuristic, machine-learnt and hybrid approaches for extracting and normalising temporal expressions, events and relations.

Since time is a continuous phenomenon, absolute and relative temporal expressions comprise a rich time hierarchy. Where a document contains many temporal expressions, broad temporal expressions such as the month or year are typically elaborated by increasingly granular temporal expressions refining the topics discussed (e.g., referring to "in 2014", then "during May" and "at mid-day on the first Monday", and so forth).

3.3.2 Implicit Temporal Clues: Temporal Dynamics

In contrast to the explicit temporal clues visibly embedded in informational content, *implicit* temporal clues originate from users' time-based information behaviour within information spaces such as the web and social media. Temporal patterns and trends in information (e.g., word use and relationships) and information behaviour streams (e.g., query popularity and meaning) are termed *temporal dynamics*. Many temporal dynamics arise as a consequence of time-based information creation, seeking and utilisation by individuals, collective groups or the world at large. Temporal dynamics are the principal focus for all research questions posed by this thesis. The temporal dynamics I explore in detail later in this thesis include word popularity, word relationships, query popularity and query intent over time.

A temporal dynamic for any given element may refer to a systematic periodicity, such as daily, weekly or seasonal regularity. Alternatively, it may refer to a one-off, irregular or otherwise trending past, present (i.e., recent/ongoing) or future, driven by related events and phenomena. A temporal dynamic might be foreseeable based on previous observations, such as the consistent demand for weekly lottery draw numbers – and thus, *predictable*. Conversely, it may be unpredictable, such as a major natural disaster and the subsequent spontaneously arising information needs. Temporal dynamics often reciprocate, since one temporal dynamic may interact with another. For example, a rise in the popularity of a TV show when it is aired weekly will cause related actors to be equally more popular, and likewise.

Topics discussed in new information, and related information needs drift and evolve over time. Accordingly, temporal dynamics are present in two distinct areas of IR: (i) information collections, and (ii) information seeking. Authors create information discussing current events and phenomena – leading to temporal dynamics in word, phrase and topic occurrence in collections. Methods to exploit past temporal dynamics contained in information collections are explored in Part III of this thesis. Meanwhile, users seek and utilise information relevant to their current situation – often typified by various time-based contexts prompting information needs. Methods to support real-time information seeking temporal dynamics are explored in Part II of this thesis.

3.4 Map of Time-aware IR

IR is a broad discipline transcending the interface of information, users, and the interaction between them during information behaviour. Time is complex and multi-faceted, and therefore naturally underlies many aspects of these IR processes. Understanding the deep involvement of time in IR requires a conceptual overview of both the user- and system-oriented elements of IR, and the various temporal clues present during each. Consequently, in this section I propose a conceptual map of time-aware IR, informed by established conceptual IR models. This map serves as a stand-alone philosophical contribution by this thesis. Furthermore, I use it to structure the extensive time-aware IR literature background provided in this chapter.

Three existing reviews offer extensive literature surveys of time-aware IR. Alonso et al. (2011) broadly outline temporal challenges and opportunities in IR. More recently, Campos et al. (2014) deliver an exhaustive survey of work on time across the spectrum of IR and information management. Meanwhile, Radinsky et al. (2013b) provide a background of work related specifically to temporal dynamics in IR.

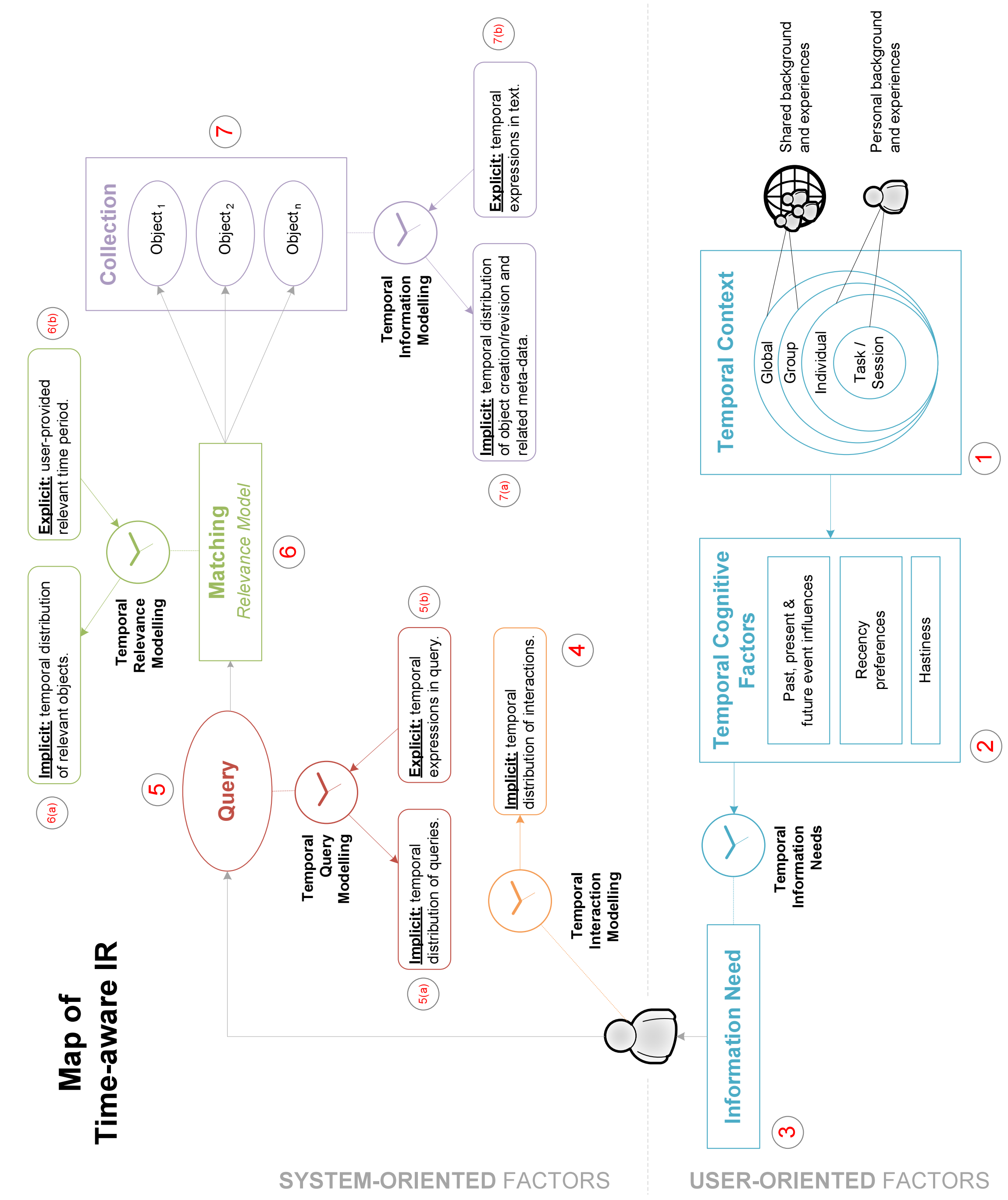

Map of Time-aware IR
SYSTEM-ORIENTED FACTORS
USER-ORIENTED FACTORS
1
Temporal Context
Global
Group
Individual
Task / Session
Shared background and experiences
Personal background and experiences
2
Temporal Cognitive Factors
Past, present & future event influences
Recency preferences
Hastiness
3
Information Need
Temporal Information Needs
4
Temporal Interaction Modelling
Implicit: temporal distribution of interactions.
5
Query
Temporal Query Modelling
5(a)
Implicit: temporal distribution of queries.
5(b)
Explicit: temporal expressions in query.
6
Matching
Relevance Model
Temporal Relevance Modelling
6(a)
Implicit: temporal distribution of relevant objects.
6(b)
Explicit: user-provided relevant time period.
7
Collection
Object 1
Object 2
Object n
Temporal Information Modelling
7(a)
Implicit: temporal distribution of object creation/revision and related meta-data.
7(b)
Explicit: temporal expressions in text.

Figure 3.1: Map of time-aware IR - highlighting the implicit and explicit temporal clues present at each stage of the conceptual IR process, divided into system- and user-oriented factors.

While these previous works offer a comprehensive review of time-aware IR research, neither offer a principled model within which to organise and reason time-aware IR approaches which either exploit or support the many implicit and explicit temporal clues in IR. Since the involvement of time in IR is multi-faceted, I argue such a map is necessary to understand diverse existing work and highlight gaps for future research opportunities.

A typical conceptual model of IR is centred around a user expressing an information need to an IR system – by formulating and submitting a query (Croft et al., 2010). The IR system models a relationship between the query and known information items (i.e., the collection). The relationship is defined by a notion of *relevance*, which identifies the information items deemed most likely to satisfy the user's information need. In interactive IR (IIR) models, the user is acknowledged to work within a number of contexts, such as their current session or work task (Ingwersen and Järvelin, 2005). These contexts form the cognitive foundation upon which the user interacts with the IR system when formulating queries, interacting with search results and assessing relevance. IIR approaches take into account the user's personal background as opposed to 'systems-oriented' approaches which typically disregard individual user-based factors.

These established abstractions form the basis of the time-aware IR model presented in Figure 3.1. Notably, I consider time to be a defining part among context factors which influence a user's cognitive state which in turn may lead to, or otherwise influence their information need. Likewise, I consider temporal clues to be pervasive throughout all aspects of the IR process – from a user expressing an information need through to matching relevant information items. Note that while I isolate the sources and implications of temporal clues in this map, rarely are these clues totally independent of one another. For example, an information need with temporal dimensions may be identified through temporal information modelling, and hence will also benefit from a time-aware relevance model. I focus on the temporal aspects related to six stages of the widely recognised conceptual IR map (numbered 1 to 6 in Figure 3.1). From a user-oriented (i.e., information seeking) perspective I consider: (1) user context, (2) user cognitive factors, and (3) information need formulation. Further, from a systems-oriented perspective I consider: (4) user interaction, (5) query input, (6) matching/relevance, and (7) the information collection.

In particular, I focus on the following established time-aware IR areas: (i) temporal information need identification, (ii) temporal query modelling, (iii) temporal relevance modelling, and (iv) temporal information modelling – each denoted as a clock in Figure 3.1.

In the following sections I elaborate on each of these elements. I discuss the implicit (i.e., temporal dynamics arising from information behaviours) and explicit (i.e., temporal expressions contained in text) temporal clues present, and outline the focus of relevant time-aware IR work which has studied, exploited, or supported each. To illustrate the implicit temporal clues, I present various real temporal dynamics examples.

Note that at the beginning of each research chapter presented in this thesis (chapters 4–7, inclusive), I include an additional brief background presenting chapter-specific literature.

3.5 Time as User Context

Context characterises the *situation* of an entity during interactions (Abowd et al., 1999). In the time-aware IR model presented in Figure 3.1, the user's temporal context is represented by part (**1**).

The notion and essence of context has seen significant investigation throughout multi disciplinary areas of research. Pragmatically, context can be viewed as the interdependence between *who's, where's, when's and what's* which together frame the motivation of *why* an action is performed in a given situation (Abowd et al., 1999). Context and in particular, time, along with the cognitive constructs of a user's perceived situation form the basis of information seeking activity (Ingwersen and Järvelin, 2005).

Informed by the stratified model of user context proposed by Ingwersen and Järvelin (2005), context is abstractly modelled as a nested phenomenon wherein broader contextual factors (e.g. those experienced as part of a group or global context) may influence more personal contexts (e.g. the individual and their past or present task/session). I argue time is intertwined throughout all these contexts since individuals are influenced by broader social contextual cues, and conversely influence one another through further information cascades.

Rarely is a user's query submitted in isolation (White and Roth, 2009), as is the user model assumed by the ad-hoc retrieval setting. In many cases, multiple queries will be be submitted during a single search session or task as the user's exposure to information, and thus understanding of their information need changes. Each query may demonstrate a deeper knowledge of their information need, and thus provide insight into the cognitive processing of the user. Query reformulation and result inspection characterised through time is a fundamental part of this iterative process (Campbell, 2000). Many adaptive IR models which take into account shifting attention over the duration of a session have been proposed. Campbell (2000) propose a time-based exponential decay to model changing information needs. Further, White et al. (2006) propose a more elaborate model of features found in recently viewed information and queries to model changing information needs. Both find that features of recently viewed information are more effective for relevance feedback during ongoing interactions in the search session.

Short- and long-term tasks such as planning a wedding, buying a house or having a baby have several task sequence and time-scale commonalities between users. For example, having baby starts with the early signs, tests and the first scan followed by general questions relating to pregnancy and birth concerns. Towards the birth, information needs are likely to turn to constructing a nursery and associated items. Following birth, nappies, medical help and general advice will likely be needed. This progression will occur in an ordered process and relatively predictable schedule. Richardson (2008) extract these schedules using common patterns of information seeking found in long-term past query logs. There are many possible applications for these types of insights, especially in personalised ad targeting.

Adar et al. (2007) observe how real-time news events and media such as popular television shows (and their associated concepts, e.g. actors and themes) influence a great deal of time-based information seeking activity. Dong et al. (2010b) look to external social sources (i.e., Twitter) which may provide clues of the event context behind information seeking. They show the focus of any relevant recent social discussion is valuable as feedback to improve the ranking of relevant information for current events, since relevance evidence provided by past search behaviour (e.g. result clicks) is sparse or outdated. Meanwhile, Kairam et al. (2013) study how users with different levels of prior knowledge of an event interact differently with related search results – and thus make conclusions on how to predict and offer better search results for various event-based search scenarios.

3.6 Time in Cognitive User Factors

Cognition is the mental process of acquiring knowledge and understanding through thoughts, experience and senses (Eysenck and Keane, 2005). Cognition is naturally at the heart of user interaction in IR. Many time-based notions play a role in various aspects of human cognition, so cognitive factors should be of great interest to time-aware IR. In the time-aware IR model presented in Figure 3.1, the user's temporal cognitive factors are signified in part (**2**).

Cognitive processes informed by a user's context and situations direct information need formulation (Ingwersen and Järvelin, 2005). Since a user's cognitive state is deeply influenced by their contextual situations, there is considerable overlap in work on temporal contexts and temporal cognitive aspects. I distinguish cognition as being an entirely internal state of the user, whereas context often includes external factors.

In the following three subsections I outline three cognitive factors that time-aware IR has broadly addressed, namely: (i) event influences and memory, (ii) recency preferences, and (iii) hastiness.

3.6.1 Event Influences and Memory

Awareness of time-anchored past, present or future events arising from a user's temporal context (i.e., the intertwined global, group, individual or task/session contexts in part (1) of the map, described in Section 3.5), will often be part of the cognitive formation of user expectations.

Explaining the human experience and perception of time has long been the subject of debate in cognitive psychology and neuroscience fields (Pöppel, 1978). Central to the controversy is understanding how time is experienced along with other human senses such as sight, hearing and smell – which together allow us to perceive and interact with our world. From an information behaviour perspective, a fundamental understanding of how humans comprehend time might facilitate a theoretically justified approach to many aspects of time-aware IR. Indeed, the majority of the work in the field of IR has been driven by algorithmic advances, rather than a theoretically principled cognitive basis (Ingwersen and Järvelin, 2005). In IR, human cognition of information is still largely an unopened 'black box', for which we test only inputs and outputs with little knowledge of the internal processes.

One widely held cognitive viewpoint initially proposed by Gibson (1975) is that events are perceivable, but time per se is not. This potentially yields implications for the theoretical underpinning for some areas of time-aware IR. In fact, it suggests time-aware IR research should further explore directions for *event-aware IR*, focusing on modelling and responding to information behaviour surrounding events over time, rather than simply considering time as a continuous phenomenon. Of course, this viewpoint thus necessitates a definitive understanding of what an "event" is in the cognitive sense. Most influentially, Pöppel (1978) proposes a taxonomy of event perception by means of temporal experiences – rather than 'time perception' per se. He suggests a number of elementary time experiences, including (i) duration, (ii) non-simultaneity, (iii) order, (iv) past and present, and (v) change, which includes the passage of time.

Memory recall of events and time-based experiences can undoubtedly inform models of IR, especially since as much as 40% of queries are for re-finding previously seen information (Teevan et al., 2007). Yet, there are seemingly few time-aware IR studies which draw directly from the theoretical foundations of cognitive psychology. Peetz and de Rijke (2013) employ a theoretically justified understanding of memory retention from cognitive psychology in order to model the impact of memory decay in a relevance model. Further considering human memory retention and recall models, Deng et al. (2011) evaluate the value of exploiting past context (e.g., the user's location and time) in addition to information topics when engaging in re-finding tasks. With recent advances in functional magnetic resonance imaging (fMRI) and subsequent studies in IR (Moshfeghi et al., 2013), a deeper understanding of the neural states underlying cognition during IR tasks may yield new directions for future time-aware IR approaches.

3.6.2 Recency Preferences

Recency, or, the desire for the most up-to-date information has become prevalent in many search applications used for real-time information seeking. Most notably, users of event-driven information platforms such as news, user-generated real-time social media (e.g. microblogs) and personal search (e.g. email search) often expect the freshest results. The tendency of human memory to recall and weigh recent experiences more favourably (i.e., the recency effect, and related recency bias (Kahneman, 2011)) is intimately understood in cognitive psychology, yet little studied from a cognitive viewpoint in IR.

Re-ranking the top N relevant results by chronological order (i.e., most recent first) is trivial when they are accompanied by explicit creation or revision time meta data (i.e., *timestamps*) – such as tweets, news documents or blog posts. This approach is employed for the top 30 tweets deemed relevant in the TREC Microblog ad-hoc search evaluation tasks (Ounis et al., 2011).

Of course, information recency and topical relevance are distinct factors. In many cases neither topical relevance nor recency are sufficient alone to rank information effectively for recency queries; relevance may in fact depend on a combination of both factors (Dakka et al., 2012). As such, many time-aware ranking approaches have explored relevance models simultaneously combining recency and topical relevance factors. Extending probabilistic language modelling IR approaches (c.f., Ponte and Croft (1998)), Li and Croft (2003) identify recency queries in a number of TREC collections and incorporate a time-based relevance prior into a language retrieval model (i.e., older documents are less likely to be relevant). With similar intention, Peetz and de Rijke (2013) apply memory retention theories to theoretically justify a time-based relevance prior. In contrast, while Efron and Golovchinsky (2011) employ a similar language modelling approach, they increase language model smoothing for older documents thereby decreasing their distinctiveness during relevance ranking.

Meanwhile, Dong et al. (2010b) propose a fundamentally different approach to recency by examining the current focus of discussion in social media. They assume users are likely to search about what is being talked about, and exploit user-generated real-time content on Twitter to identify topics related to recent events, and rank recently relevant items appropriately. Considering how relevance can change over time is a theme I address in Chapter 5.

Modelling topical relevance and recency simultaneously is not trivial. Consider a user seeking information about a news event. The media which initially broke the news events may be persistently highly relevant throughout the event; but, most recent content may be more sensitive to recency given the evolving focus of discussion. Indeed, this challenge has motivated more robust temporal relevance approaches, which I discuss further in Section 3.10.1.

3.6.3 Hastiness

Hastiness refers to the amount of time a user has available to satisfy their information need. In the sense of time-aware IR, hastiness relates to two aspects: (i) the conciseness of the information retrieved, and hence the time and effort necessary for the user to interpret it, and (ii) the time taken to respond with results following the user's query submission.

While no rational user will willingly want to waste time, a hurried user has an impetus to find succinct information quickly in order to satisfy their information need. The reason for haste is defined by the situational context of the user (Ingwersen and Järvelin, 2005; Savolainen, 2006). For example, a user searching for somewhere to visit with a mobile device is unlikely to want to read exhaustive detail about venues, but rather to see a comparative summary of options to make a quick but informed decision.

Alternatively, hastiness can also relate to the latency between requesting information (i.e., submitting a query to the system) and receiving/viewing results. Commercial search engines have long strived to reduce the latency between user request and response with countless caching and efficiency techniques (Baeza-Yates et al., 2007; Maxwell and Azzopardi, 2014). The prevailing belief is that more responsive (i.e., *faster*) search engines are more satisfying for users, and therefore more commercially effective (Baeza-Yates et al., 2007). Several marked effects on interactive user behaviour, such as increased document dwell time and decreased session queries, have been observed when the latency of query response and document viewing is artificially delayed under experimental conditions (Maxwell and Azzopardi, 2014).

On the contrary, recent research has argued that time is not always a constraint for users. In what has been termed "slow searching", many users are prepared to wait longer for more satisfactory results, especially for complex and exploratory information needs (Teevan et al., 2013). This work challenges the conventional wisdom of instant search results, and paves the way for further development of more computationally complex techniques (for which instantaneous results are currently infeasible), or even, for human augmentation of algorithmic results (e.g., crowdsourced relevance, curation and summarisation techniques).

3.7 Time in Information Needs

From a theoretical perspective, Belkin (1980) proposes that information seeking behaviour arises from an anomalous state of knowledge (ASK), that is, the user's awareness that they lack knowledge – and therefore need to pose questions to address the uncertainty. Stemming from the temporal context and conditioned by temporal cognitive factors, the user develops an information need which they can express to an IR system by means of a query. In the time-aware IR model presented in Figure 3.1, the user's information need is signified by part (**3**).

Literature refers to "temporal information needs" as those information needs which have some clear temporal dimension such as recency, or relating to specific past, present or future time-based events. Temporal query modelling relates to identifying and modelling such information needs (Berberich et al., 2010; Campos et al., 2014); I discuss this work as part of the following section on time expressed in queries.

However, since time is so multi-faceted, in a sense one could argue that *all* information needs are temporal in some respect – albeit perhaps more subtly. Consider a seemingly non-temporal information need such as "`where can i buy` x". Several time-based aspects are likely to play a role in satisfying the user's expectations. For instance, the user will probably want to be recommended a shop that sells the current version of x, is not out of business, and possibly is still open at the time of the information need (that is, if the query has a real-time intention). It is therefore clear that several implicit time-based factors underpin both information seeking and relevance expectations. This thesis aims to serve as an exploration of many of these underlying factors, characterised through temporal dynamics.

3.8 Time in Interaction

In general, human behaviour has many temporal factors. As a result, several implicit temporal clues can frame a user's interactions (and in fact, lack of interactions in the case of absence time) with an IR system over short and long periods of time (Preum et al., 2015; Radinsky et al., 2013b). Elements of time in interaction are characterised in part (**4**) of the time-aware IR model presented in Figure 3.1.

Despite time being a key element of user effort when interacting with information (and consequently, their satisfaction), it has only recently been integrated into IR effectiveness evaluation (Smucker and Clarke, 2012). Traditional IR evaluation measures of system effectiveness, such as MAP, P@k and nDCG, do not consider the time-based effort necessary for a user to inspect and evaluate retrieved information. Following the intuition that users desire both relevant and succinct information in search results, time-biased gain proposed by Smucker and Clarke (2012) models temporal effort (based on elements such as document length and information redundancy) in a generalisable cumulative gain measure. Time-biased gain is useful for evaluating interactive scenarios, such as web search and query suggestions, in a systematic and repeatable systems-oriented evaluation.

Preum et al. (2015) provides a personalised model for predicting when a user will interact with a information system (such as search engines, web and mobile apps, etc.) over time based on common and personal temporal behaviour traits. To quantify the quality of user experience in information systems, various models and metrics of user engagement have been proposed (Lehmann et al., 2012). User engagement takes the experience of using the system as a whole into account, rather than isolated algorithmic evaluation measures, such as recall or precision measures in IR. User engagement is an inherently temporal concept, since users satisfaction may rise or fall over time with temporal factors such as their location or activity. Accordingly, Drutsa et al. (2015) analyse the periodicity of engagement (and conversely, absence) to find several classes of users for which engagement metrics do, and do not exhibit temporal dynamics.

3.9 Time in Queries

Temporal aspects can be implied in many forms when a user poses a query to an IR system, characterised in part (**5**) of the time-aware IR model presented in Figure 3.1. The user may explicitly specify temporal expressions in their query, e.g. "`world cup 2022`". Alternatively, temporal dynamics of collective query popularity may be indicative of underlying query likelihood and meaning. For instance, a recent event might make a particular query topic popular, and equally, influence what the user is expecting to find in results.

In this section I examine work on *temporal query modelling*, which relates to efforts on modelling the temporal aspects of queries submitted to search systems in order to better support them. This includes understanding the explicit time clues expressed in user's queries, as well the implicit patterns and trends in time-based query popularity in user populations. Indeed, more broadly, temporal query modelling also draws on temporal clues from the user's context and the relevance model to achieve its aims.

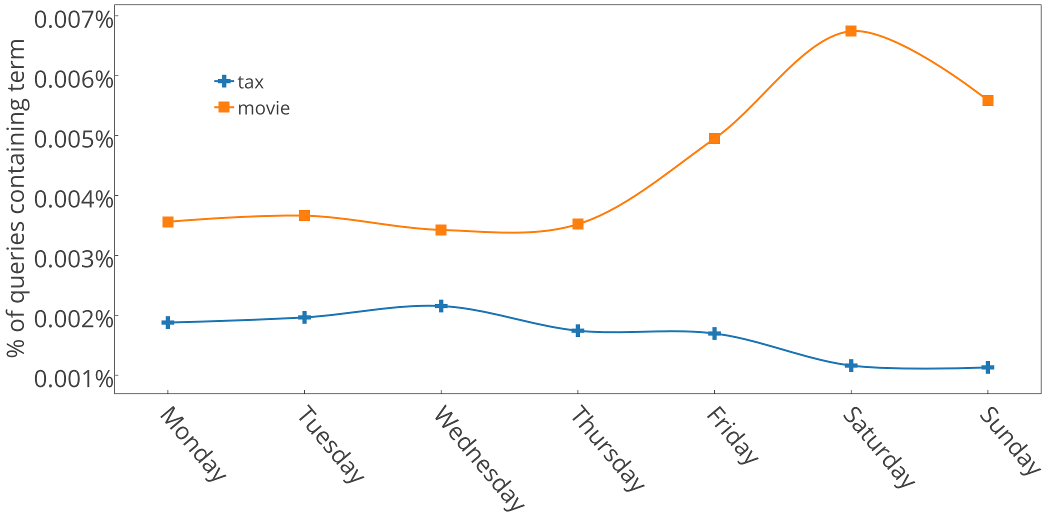


Figure 3.2: Temporal dynamics in the daily popularity of "movie" (i.e., entertainment related) and "tax" (i.e., work related) queries in the MSN query log.

3.9.1 Implicit Temporal Clues: Information Seeking Temporal Dynamics

When a query is posed to a system provides a temporal clue, since it offers a signal of the user's current interest in a particular query topic. When the stream of queries submitted is considered collectively, query popularity varies dramatically over time as real-world events and phenomena influence information seeking behaviour in different population groups. As such, query popularity is subject to a wide range of temporal dynamics including periodic trends and one-off spiking patterns, etc., (Kulkarni et al., 2011).

In Figures 3.2 and 3.3, I present real temporal dynamics in the popularity of queries containing the terms "movie" (i.e., an entertainment related intent) and "tax" (i.e., a work related intent), obtained from the MSN 2006 1 month query log[1].

It is clear from the weekly temporal dynamics in Figure 3.2 that work intents are more common mid-week (i.e., Monday-Friday), and conversely, entertainment intents are most popular at weekends. Similarly apparent in Figure 3.3 is the relative popularity of work intents during Western working hours (i.e., 6am-4pm), in contrast to the popularity of entertainment intent in the evening and at night. Indeed, Kramar and Bielikova (2014) find that context often clearly shifts from work to leisure topics for many users, and highlight the need for personalisation approaches to take this temporal context into account. Such temporal dynamics compose an underlying structure of query popularity over time. I argue that search engines which attempt to support users over time with features such as query auto-completion must

[1] The openly available MSN 2006 query log is used for experiments in Chapter 4, and as such is described in Section 4.4.

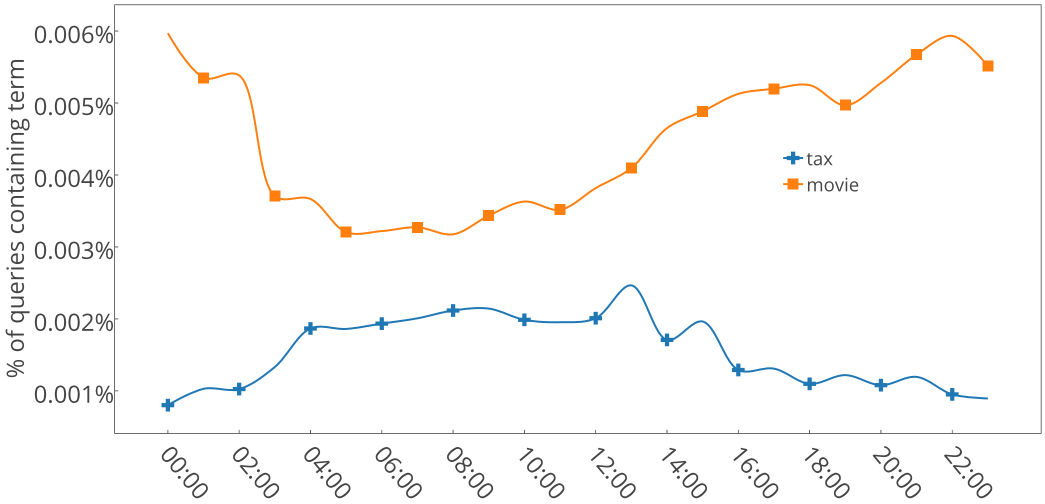


Figure 3.3: Temporal dynamics in the hourly popularity of "`movie`" (i.e., entertainment related) and "`tax`" (i.e., work related) queries in the MSN query log.

take these temporal dynamics into account for consistent effectiveness, and hence explore this problem as RQ1 of this thesis in Chapter 4.

Several studies have analysed the temporal dynamics of information seeking behaviour through temporal query log analysis. Wang et al. (2003) observe the broad temporal dynamics in query popularity and topics in an academic search engine. Beitzel et al. (2004), and later Beitzel et al. (2007) quantitatively study topical trends in common query topics at different times of the day. Zhang et al. (2009) analyse the variance in navigation and transactional queries at different times of day in order to better predict and consequently support user behaviours during search.

An extensive study by Adar et al. (2007) on temporal query activity explores the temporal correlations between topics arising in television shows, news and web search. As part of this work, they characterise the associations and relatively short lag time in information seeking behaviour triggered by real-world phenomena. Based on elements of this behaviour, Zhao et al. (2006) use time-based querying and subsequent result click behaviour to effectively detect and track unfolding events in real-time.

Kulkarni et al. (2011) propose a taxonomy of temporal query patterns based on large-scale query log analysis. They discover that while many queries have relatively stable popularity (e.g. navigational queries such as "`youtube`"), many distinct temporal patterns and trends are evident in query popularity. The taxonomy they propose includes 'wedges' (i.e., equal rise and fall), 'castles' (i.e., persistently increased popularity) and 'sails' (i.e., instant or slow rise, followed by the converse fall). Furthermore, they measure time-based changes in result click entropy, which in turn suggests changing user intentions. I argue this change should be taken into account to provide more effective IR systems, thus motivating RQ2 of this thesis which I investigate in Chapter 5.

Examining the similarity between the time-based popularity of different queries over time, Vlachos et al. (2004) propose various data mining techniques to characterise the periodic temporal patterns between queries. Temporal correlation between temporal dynamics is used to discover semantically similar query topics – e.g. "`christmas tree`" and "`christmas gift ideas`", which are both popular in early to mid-December. In similar fashion, Chien and Immorlica (2005) propose techniques to reliably uncover semantically similar queries from a large-scale query log based on temporal popularity similarity. Building on this work, Alfonseca et al. (2009) use temporally similar queries to quantify temporal semantic associations between the words and phrases contained in web search queries.

Predicting future query popularity has received interest for improving trending query detection and query auto-completion effectiveness. Golbandi et al. (2013) develop a time-aware regression approach to improve query trend detection at different times of day, where different emerging trends are likely to lead to varying outcomes. Shokouhi (2011) proposes several time series modelling approaches based on exponential smoothing to predict future query popularity based on previously observed trends (e.g., the lottery occurring every Wednesday night, and so, users searching for lottery numbers). To this end, Strizhevskaya et al. (2012) explore several more elaborate time series modelling approaches to improve prediction performance. Shokouhi and Radinsky (2012) demonstrates that time series modelling for future popularity prediction considerably improves query auto-completion suggestion ranking. In this thesis, I address several challenges in real-time query popularity prediction to enhance query auto-completion effectiveness in Chapter 4.

3.9.2 Explicit Temporal Clues: Temporal Expressions in Queries

Absolute and relative explicit temporal expressions are included in as many as 1.5-2% of web queries (Brucato et al., 2013; Nunes et al., 2008), such as "`world cup 2012`", "`nyc weekend events`" and "`18th century poets`". These temporal expressions often serve as a means to either temporally disambiguate or specifically qualify the information need expressed. In some cases, relevant documents will have been created during the period expressed by the temporal expression. Alternatively, the temporal expression might refer to relevant events discussed in documents of any age.

Strötgen et al. (2012) integrate temporal expressions identified in a user's query into a relevance model that integrates both textual and temporal elements. Of course, many queries may only be composed of keywords, yet, have an inherently temporal dimension, e.g. "`world cup brazil`", referring to "`world cup 2014`" (Kanhabua and Nørvåg, 2010a). In these cases, the implied time period can be inferred to temporally augment the query and enable temporal relevance modelling to improve retrieval.

3.10 Time in Matching/Relevance

Matching is the process of identifying information items in the collection which are expected to satisfy the user's information need. Relevance quantifies the strength of a match, and facilitates relevance ranking whereby items with the greatest expected utility can be presented to the user first. In the time-aware IR model presented in Figure 3.1, matching and relevance is characterised in part (**6**).

Research on the fundamental nature of relevance as a dynamic and multi-dimensional concept has long since been at the heart of the interface between information science and IR. Several extensive works have wrestled with philosophical, cognitive and practical definitions of relevance (Saracevic, 2007). Context is a fundamental part of relevance. Since context is dynamic, so too is relevance across users, and indeed, for the same user over time (Saracevic, 2007). Elaborating on the role of time, Hjørland (2010) suggests user-based relevance theories tend to ignore the social nature of the world, of which time is a key element – instead treating the user as an isolated individual. User-oriented relevance has concentrated on discovering a general psychological mechanism residing in the mind of each individual user. Instead, Hjørland (2010) recognises that knowledge is expanding and changing all the time. Therefore, as relevance has been shown to be closely related to the user's previous information interaction, he posits that information needs or relevance as developing inside the mind of a user cannot be understood disregarding the development in our collective knowledge, and thus, the time reflected throughout information collections.

In this section I examine work on *temporal relevance modelling*, which relates to incorporating temporal aspects into IR relevance models which have conventionally focused solely on topical relevance.

3.10.1 Implicit Temporal Clues: Relevant Item Distribution

Several temporal dynamics arise when relevance is considered over time. Firstly, in a collection where items are timestamped, the post-retrieval timestamp distribution of top retrieved items (i.e., those deemed most relevant) often contains implicit time clues of events and phenomena related to the query posed. Secondly, items considered relevant to a given query may change over time as users' underlying intentions behind the information need changes.

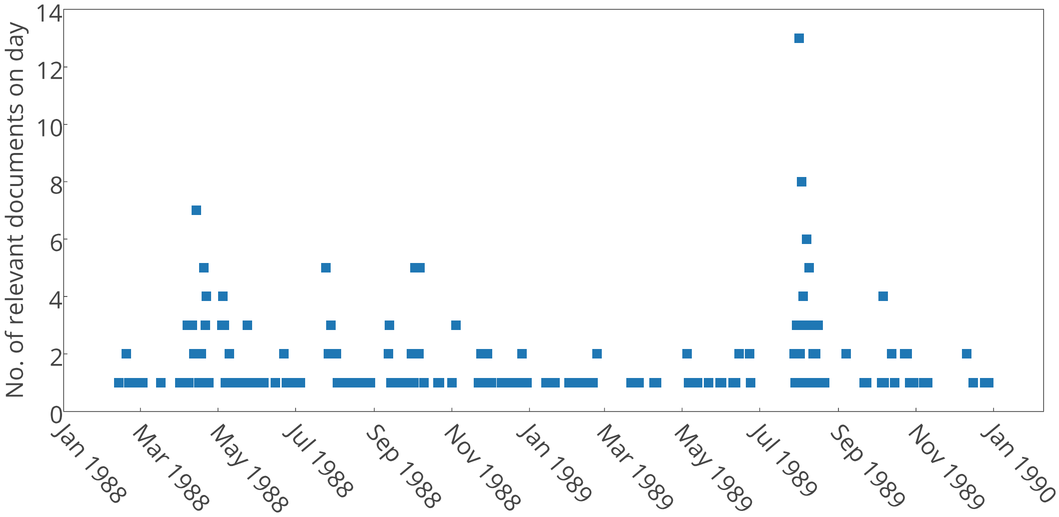


Figure 3.4: Temporal dynamics in the timestamp distribution of relevant documents for TREC-2 topic 64 (“`hostage taking`”) in the AP 88-89 news wire collection.

Indeed, these items may or may not be timestamped. Hence, there are more subtle temporal behavioural factors involved. I discuss these temporal dynamics in the following subsections.

Retrieved Item Timestamp Distribution

Collections containing timestamped information items often cover events (e.g., news with a publication time, or email with a sent/received time). Many information needs for these collections are time-sensitive in the sense that the topics they cover may relate to specific periods in time (Dakka et al., 2012; Jones and Diaz, 2007).

Illustrated in Figure 3.4 are temporal dynamics in the relevant document timestamp distribution for the query “`hostage taking`”, in the TREC AP88-89 news wire collection[1]. While hostage taking is discussed for the duration of the collection, there are clear events represented by spikes in relevant documents, most notably in May 1988 and August 1989. These periods correspond with reporting on major hostage taking events.

Following an initial retrieval, with analogous intuition to pseudo-relevance feedback, the patterns contained in the timestamp distribution of the top-k topically-relevant results may be considered an indicative sample of the time distribution of *all* relevant results (e.g. periodic or recency skewed, etc.). A substantial body of works exists for exploiting this temporal dynamic in time-aware IR tasks, including identifying highly relevant time periods for exploration, diversity, performing query expansion and query performance prediction.

[1] The Associated Press (AP) 1988-89 test collection is described and used for experiments in Chapter 6

Jones and Diaz (2007) propose several techniques to identify and classify temporal information needs based on the relevant document timestamp distribution. Following classification of a temporal query, query performance prediction is enhanced by modelling the time-based distribution of all results.

Dakka et al. (2012) use the relevant item timestamp distribution to estimate temporal priors which are used as feedback for established retrieval models to improve retrieval effectiveness. Efron et al. (2014) expand this idea to Microblog search, and propose a framework for what they term the ‘temporal cluster hypothesis’ – positing that relevant documents often cluster together in time (i.e., for a specific event). Rather than modelling priors, IR effectiveness in this scenario is improved by biasing temporally adjacent results using a temporal density function based on kernel density estimation. Meanwhile, Berberich and Bedathur (2013) diversify retrieval results by selecting most the relevant items from all temporal clusters.

Further exploiting the fact that relevant documents for many time-based information needs tend to cluster in time, Peetz et al. (2014) extract distinctive terms found in bursts of documents detected in the relevant item timestamp distribution in order to expand queries and improve retrieval performance. Meanwhile, Massoudi et al. (2011) improve retrieval by expanding queries with distinctive terms found in the most recent relevant items, thus capturing only the most recent discussion topics.

Since relevant item timestamp distribution is an implicit artefact of the retrieval model, it is relatively sensitive to the underlying retrieval model effectiveness. Consequently, when the initial retrieval is poor, then it is unlikely to sufficiently reveal an accurate timestamp distribution of relevant results. Accordingly, further work is needed to understand how reliable this feature is for temporal modelling in different scenarios.

Relevant Item Distribution

In contrast to the relevant item timestamp distribution, temporal dynamics in relevant items may be observed as users select different items for the same query over time (and thus, indicate temporal relevance changes). This change reflects a shift in *intent* underlying a user’s query. This notion of time-sensitive relevance was first noted by Kulkarni et al. (2011) after observing increased entropy in search result clicks for some queries over time.

For example, consider the non-specific web search query “`party planning`”. Depending on the time of year, the type of party the user is likely to be planning will vary, and therefore, so too the relevant search results they are likely to desire. To illustrate the change in party planning intent, in Figure 3.5, using Google Trends data over two years I present the popularity temporal dynamics of the each party type intent, e.g., “`halloween`”, “`christmas`”,

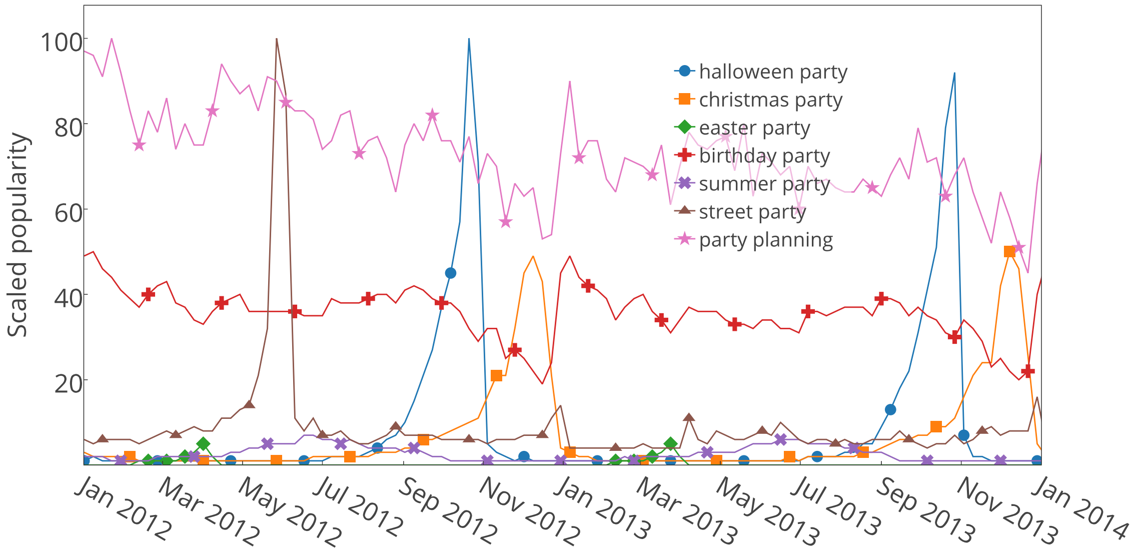


Figure 3.5: Temporal dynamics of six possible intents for the web search query "`party planning`", based on data from Google Trends.

"`easter`", "`birthday`" and "`summer`". Also included is the scaled popularity of the "`party planning`" query itself, showing occasional temporal correlation with the intent popularities – for relatively large query bursts around major events – over time. Note that the scaling is relative only, since Google does not release absolute popularity statistics for queries.

In Figure 3.5, it is clear that birthday parties are relatively popular uniformly throughout the year, albeit with a decline over summer in favour of more general summer parties. Clear temporal dynamics are present indicating the seasonal change in party planning intent, such as Christmas party planning in the run up to Christmas. While the majority of temporal dynamics in this example are periodic, there could equally be one-off short-term bursts of changing query intent, such as for "`street party planning`" during the celebration of the British Queen's Diamond Jubilee celebrations[1] in the United Kingdom at the start of June, 2012.

Keikha et al. (2011) take a different view on the temporal dynamics of relevant items to improve blog discovery. To this end, they examine the relevance of the a blog to a given query over different periods of time. A blog which is consistently relevant for a topic over all periods time is considered to be most relevant for the query.

Central to many of the temporal dynamics themes discussed throughout this thesis, Radinsky

[1] The Diamond Jubilee event was the period of national celebration for the 60 year reign of Queen Elizabeth II. Local community street parties were widely organised across the country in celebration.

et al. (2013a) and later Radinsky et al. (2013b) propose an extensive framework for modelling temporal relevance of individual search results based on past and present temporal dynamics. In particular, they propose predictive models incorporating smoothing, trends, periodicities, and surprises in time-based behaviour, such as queries and result clicks. Results demonstrate significant improvement in IR effectiveness by incorporating temporal dynamics. To open new directions in modelling and operationalising temporal relevance, I explore temporal dynamics of query intent in relation to modelling temporal relevance for RQ2 of this thesis in Chapter 5.

3.10.2 Explicit Temporal Clues: User-stated Relevant Time Periods

In many search systems, such as news and archive search, the user is able to specify a time period to rigidly filter the timestamps of relevant results.

Despite this being a common feature, there has been surprising little research on how users interact with this feature, and its impact on satisfaction. Based on analysis of the TREC Microblog collection, Lin and Efron (2013) suggest allowing the user to specify ‘hard’ and ‘soft’ relevant time periods. This finding is driven by the expectation that imposing a exact ‘hard’ relevant time period is problematic when the user may not actually know the exact time period they should be looking for, and so simply biasing certain periods (i.e., ‘soft’ relevance) is therefore more likely to lead to satisfactory results.

When information objects are timestamped, these approaches are easily applicable. However, a great deal of information is not accompanied by explicit creation or revision timestamps, such as web pages. To overcome this problem, more elaborate techniques must be employed to estimate the information age. If dated revision notes are contained in the text, information extraction (e.g. temporal expression tagging) might be appropriate. In some cases, HTTP headers accompanying a web page may contain a last revision date. Alternatively, de Jong et al. (2005) and Kanhabua and Nørvåg (2008) demonstrate language modelling can be used to estimate a document’s age based on the usage of temporally specific terms contained only in documents of a known age. Alternatively, for web documents, Nunes et al. (2007) find links to and from timestamped web documents, such as news articles, can be modelled to infer age.

3.11 Time in Information Collections

Time is reflected in many ways in information collections. Collections may be static, that is, no new items will ever be added. Alternatively, they may evolve as new items are added incrementally, often as a real-time stream of incoming items. In either case, items in the collection may or may not be individually timestamped. Depending on the composition of the collection, it may have breadth, or conversely, depth of items over relatively short- or long-term periods of time. Documents may be revised, as for Wikipedia articles or web documents, leading to implicit patterns as topics persist and evolve across changes. Documents often also contain explicit temporal expressions relating to new time-anchored topically related events. In the time-aware IR model presented in Figure 3.1, the collection is characterised in part (**7**).

Domain specific meta-data often accompanies information, forming temporal structures such as linking and social networks surrounding the information. On the web, hyperlinks between pages form a rich structure which informs information meaning and authority (Nunes et al., 2007; Page et al., 1999). In user-generated content, meta data such as authors, tags (e.g., hashtags), endorsements (e.g., retweets or likes) and sharing serve as temporal clues for meaning and 'interestingness'.

Efforts to understand the temporal nature of information in collections – *temporal information modelling* – is by far the most diverse and developed area of time-aware IR. This branch of time-aware IR tends to be concerned with understanding topics and the words and relationships from which they are composed, and how they change over time. Furthermore, temporal change in domain-specific meta data offers a rich insight into the meaning and importance of information over time, which in turn can be used as a temporal signal to inform better IR models.

3.11.1 Implicit Temporal Clues: Collection-based Temporal Dynamics

Collection-based temporal dynamics are derived solely from changes in the content, or accompanying meta data of information items created over time. In the following section, I explore a number of areas where temporal dynamics have been investigated for various applications.

I use the TREC Associated Press (AP) 1988-1989 news test collection and TREC-1 test topics to provide many of the real temporal dynamic examples in this section. The collection covers almost two years, and is composed of around 150K time stamped news wire articles. Part III of this thesis explores temporal dynamics contained in information collections. In these chapters I introduce several other time-based test collections, along with techniques to exploit the temporal dynamics contained within them to improve IR system effectiveness.

Information Diffusion

Information spread (or, *cascades*) throughout social networks is an intrinsically time-based stochastic process, for which temporal dynamics are a key factor. A significant body of work characterising trends prevalent in various domains exists. Kumar et al. (2004) and Gruhl et al. (2004) model topics spreading through web blog discussion. Leskovec et al. (2009) study and characterise the temporal dynamics in dissemination of breaking news through media outlets. Impressively demonstrating the spread of information over time, Sakaki et al. (2010) show how earthquakes can be detected in real-time from social media posts.

Temporal Meta-data and Structure

A network formed of links between information items is a valuable source of insight for many non-content based factors such as authority and interest. Dai and Davison (2010) study evolving web page linking activity to incorporate freshness into PageRank-based web page authority estimation. Choi and Croft (2012) exploit the time of re-tweets on Twitter to extract relevant time periods, with the hypothesis that when people retweet information it is because they find it interesting, consequently providing a temporal relevance signal.

Temporal Topic Modelling

Topic models are probabilistically composed models comprising common word occurrence and co-occurrence forming the topics present in a collection of documents. Among the most commonly used topic modelling approaches is Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003). While LDA was not originally proposed to be time-aware, it has since been adapted to run sequentially across time-ordered documents contained in sequential time periods (namely, ‘buckets’), thus facilitating modelling of time-evolving topics (Blei and Lafferty, 2006).

Kleinberg (2002, 2006) identifies hierarchies of bursting terms in time-based collections in order to discover changes in the focus of discussion, such as different themes of emphasis in the annual US State of the Union address.

Wang et al. (2007) simultaneously model topics appearing in two streaming collections in order to identify the same events being discussed in English and Chinese language. Patterns discovered can provide insight into the latent topic evolution. Similar to the findings of their work, in Chapter 4, I demonstrate that temporal dynamics are language agnostic, and therefore techniques exploiting or supporting them are readily generalisable, e.g., to IR systems in different scenarios and languages.

Retrospective (i.e., past collection) and online (i.e., real-time streaming collection) topic detection and tracking (Allan, 2002) and event detection (Allan et al., 1998) has developed several approaches for identifying new events and drifting topics based on many collection-based temporal dynamics, such as the trajectory of word and phrase usage and specificity (He et al., 2007).

With application to IR, Metzler et al. (2012) instead looks to characterise past events in a Twitter collection to retrieve them as cohesive units for a given query, as opposed to retrieving their individual often out-of-context tweets.

Content-based Temporal Dynamics

In many collections, the same document can be revised and updated over time. Persistent and new or evolving content appearing in a single information item over time leads to content-based temporal dynamics, or synonymously, *content dynamics*.

Changing content is especially common in web pages. An incremental record of the changes can be recorded (as is the case for the collaboratively edited Wikipedia encyclopedia), however in many cases it must be inferred from comparing snapshots of the page over time. Links added and removed from the content provide temporal meta-data and structure, discussed previously in Section 3.11.1. Content structure implied by sections and headers can also offer rich insight into the nature of temporal change.

Fetterly et al. (2003) crawled 150M pages every week for 11 weeks to make several observations of content dynamics. Counter to expectations, they found longer documents tended to have more frequent and extensive changes made to them. Notably, they also found for web pages that change, they do so consistently over time. As a result, web documents must be regularly crawled, and retrieval models must be sensitive to temporal content changes which impact both whole page relevance, and the composition of a collection in terms of statistical term distributions.

Adar et al. (2009) study a directed sample of 55K highly popular web pages (e.g. portal pages) to identify stable and dynamic content. They found that these highly used pages tended to change significantly much more frequently than suggested by previous studies – many every hour. They provide various analyses of the content change at the whole page, page structure and term level, and propose several related models for effectively predicting future change.

Acknowledging that web page content is rarely static, Elsas and Dumais (2010) incorporate content dynamics into a relevance ranking model to improve navigational query effective-

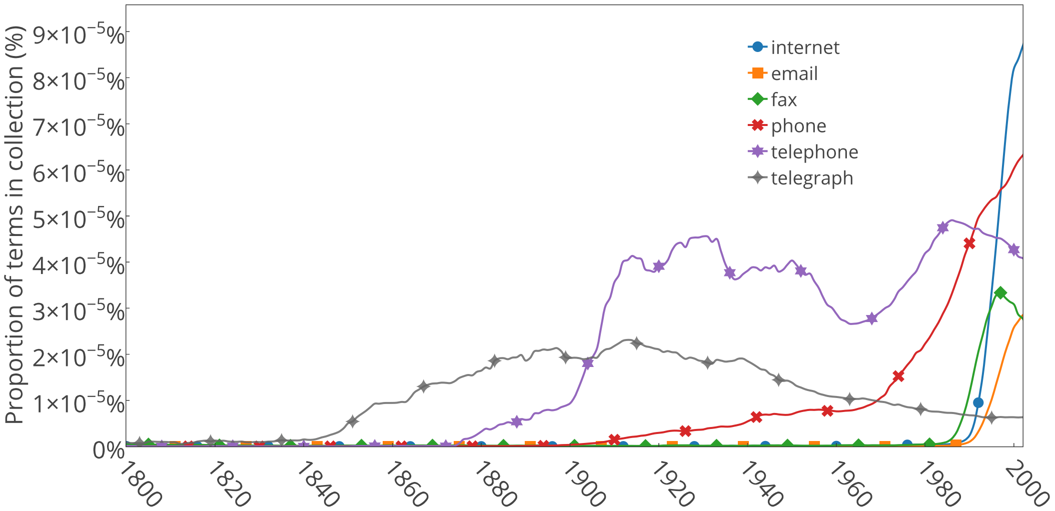


Figure 3.6: Temporal dynamics in the annual frequency of communication methods discussed in the Google Books corpus (Michel et al., 2010) of printed works published between 1800 and 2004.

ness. To improve retrieval, the proposed model exploits the temporal dynamics of document content to favour web documents which persistently contain relevant terms.

I further explore the hypothesis that authors write about currently relevant topics that are of interest to information seekers in Chapter 5, establishing a relationship between content-based temporal dynamics and temporal query intents for event-driven queries.

Lexical Change

Lexical change is defined as the changes in word and phrase use over time. In a time-based collection, word frequency is subject to temporal dynamics as cultural change, events and phenomena all influence what information authors discuss over time. This is most apparent when one term fades into obscurity from common use while another becomes popular.

From a social observation point of view, temporal dynamics of word frequency are reflective of societal change over time (Michel et al., 2010). In Figure 3.6, I illustrate the temporal dynamics of the communication technologies “internet”, “email”, “fax”, “phone”, “telephone” and “telegraph” appearing in printed books between 1800 and 2004 in the Google Books collection[1]. The temporal dynamics for each technology are very prominent, relating to its adoption, use and subsequent obsolescence. For example, “telegraph” became popular at the start of the 1900’s, but begins to decline once the “telephone” era begins, and then later, the boom of the “internet” age from the mid-1990’s.

[1] Google Books is a large-scale digitised library of dated publications. Counts of words and phrases appearing in publication content each year are released for research at: https://books.google.com/ngrams

Odijk et al. (2012) present a time-aware exploratory search system that visualises temporal dynamics of terms in a collection. Often the same named entity might have alternative but equivalent references over time. For example, "Hillary R. Clinton" was also referred to as the "New York Senator" between 2001 and 2008. Kanhabua and Nørvåg (2010c) and Kanhabua and Nørvåg (2010b) improve retrieval by expanding queries with such time-based synonyms. Efron (2013) suggests the problem of finding time-based synonyms is similar to cross-language IR. With this in mind, he proposes cross-temporal retrieval as a framework to combine temporal evidence, and support effective retrieval of historic documents containing words and phrases that are similar in meaning to a query expressed in modern English.

In this thesis, I explore the value of temporal dynamics to infer semantic similarity between index terms, and subsequently improve IR effectiveness in Part III.

Semantic Similarity Change

In linguistics, semantics is the study of word *meaning*. Integral to a word's meaning is its relationship (i.e., *semantic similarity* or synonymously, *semantic relatedness*) with other words and phrases, and what they represent. For example, words such as "`oil`" and "`energy`", and "`doctor`" and "`medicine`" have some notion of semantic association for many people – although the precise nature of the relationship may depend on individual experiences and perception, of which time is likely to play role. Humans innately build models of semantic similarity based on their knowledge, capturing various underlying cognitive reasoning and beliefs (Dumais, 2004; Radinsky et al., 2011). Several semantic representation models have been proposed in an effort model and quantify latent "common sense" semantic knowledge (Dumais, 2004; Gabrilovich and Markovitch, 2007; Mikolov et al., 2013; Radinsky et al., 2011).

Characterising relationships between words and phrases plays a fundamental role in many IR techniques, including query expansion, clustering, topic modelling and disambiguation. Temporal change in the semantic relationships between words and phrases over time is therefore of great importance for developing robust time-aware IR models. In Figure 3.7, I illustrate anecdotally the temporal dynamics of term co-occurrence in the long-term British Parliamentary Hansard collection[1]. Co-occurrence in debates – measured by the Jaccard coefficient – is plotted between the term "`energy`" and related terms: "`coal`", "`nuclear`", "`renewable`" and "`oil`" in the Hansard collection between 1945 and 2004. Co-occurrence is computed in discrete two year time frames to capture reliable patterns and trends. Notably, British government energy policy and major world events are reflected by the promi-

[1] The Hansard is the official transcript of debates in the public chamber of the House of Commons by elected members of parliament (MPs) in the British Government, from 1803 to present.

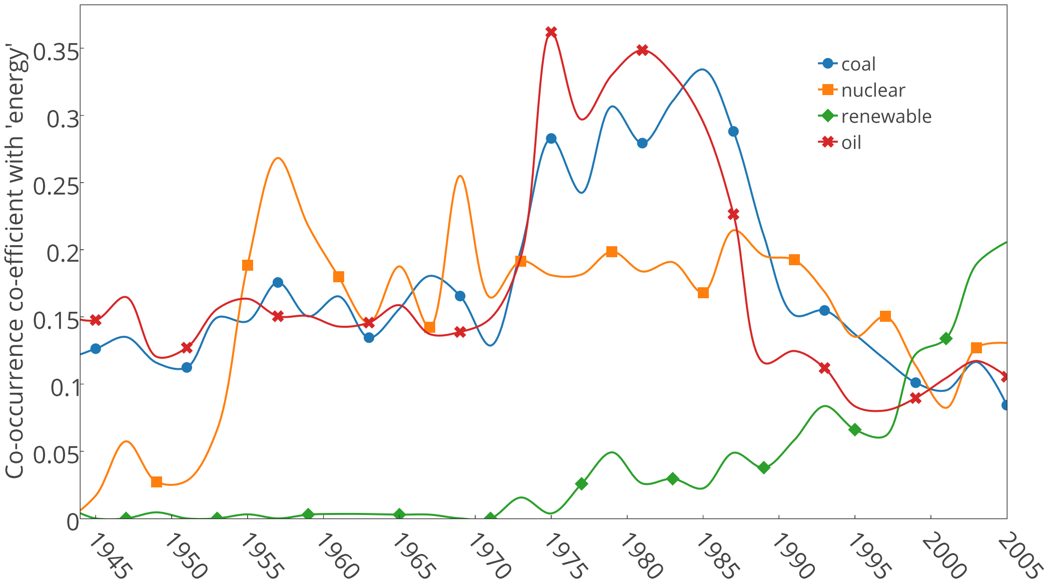


Figure 3.7: Two-yearly Jaccard co-efficient (i.e., document-level co-occurrence) of the term "`energy`" and related terms:"`coal`", "`nuclear`", "`renewable`" and "`oil`" between 1945 and 2004 in the *Hansard*, the UK parliamentary debates record.

nent co-occurrence temporal dynamics. Rising co-occurrence with "`energy`" in the 1970's and 1980's for "`coal`" and "`oil`" arises from increased concern and discussion about the reliance on these energy sources following the 1973 oil crisis, and increasing coal prices caused by supply shortages after widespread miners' strikes. The rapid rise of nuclear energy is clearly visible in the early 1950's, as is the gradual increase of renewable energy from the late 1970s. Similarly evident is the relative fall in oil, coal and nuclear energy in favour of renewable energy through the 1990's.

Past work has explored semantic similarity and co-occurrence temporal dynamics from different perspectives. Odijk et al. (2012) visualise the change in highly co-occurring communities of entities over time to aid exploratory search. Wijaya and Yeniterzi (2011) use the co-occurrence evidence provided by 5-grams in the Google Books collection (Michel et al., 2010) to illustrate the changes in close proximity term co-occurrence (i.e., within 5 words), modelled as a graph over hundreds of years. Radinsky et al. (2011) demonstrate similar word frequency temporal dynamics to be indicative of semantic relatedness. Consequently, they propose a measure of temporal semantic relatedness based on the correlation between word frequency temporal dynamics, and find it highly correlates with human judgements of semantic relatedness.

However, none of the aforementioned studies on semantic relatedness temporal dynamics apply their findings to retrieval models which can in turn improve time-aware IR effectiveness.

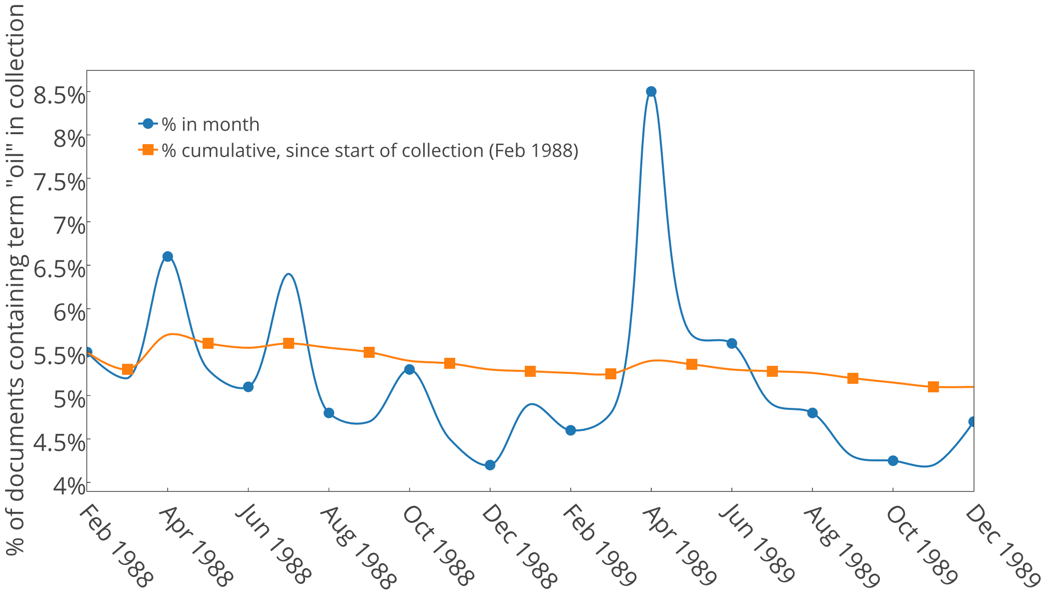


Figure 3.8: Temporal dynamics in the monthly specificity of the word "`oil`" in the Associated Press 1988-89 news wire collection.

In this thesis I posit that these temporal dynamics are valuable for improving retrieval effectiveness for time-based collections. As such, for RQ3 of this thesis, in Chapter 6, I study how the topic-specific semantic similarity of terms changes over the duration of a collection, and how these relationships can be exploited to improve IR effectiveness.

Word Specificity

Word specificity quantifies how many items a given word appears in within the information collection. Words which appear in relatively few items in a collection are more specific, and therefore more discriminative for distinguishing relevant from non-relevant items (Luhn, 1958). Recall, as I considered in Chapter 2, word specificity is a vital part of the majority of modern retrieval models – typically quantified using inverse document frequency (Sparck-Jones, 1972), or derivatives.

Much like word frequency over time, word specificity is also subject to temporal dynamics as documents discussing a topic and overall composition of the collection evolves. In Figure 3.8 I illustrate temporal dynamics of term specificity, measured as the percentage of collection documents containing the word "`oil`" each month, and cumulatively in the AP 1988-89 collection. Note, the greater the percentage of documents containing a word, the less discriminative it becomes for IR purposes of distinguishing relevant documents.

The specificity of "`oil`" deviates dramatically from month to month; events such as the Valdez oil tanker disaster causing supply concerns are prominent (the spike around April

1989). Consequently, retrieving information about "oil" may be more unreliable during low specificity periods. Cumulatively, the specificity of "oil" is much less volatile, although it does trend downwards over the duration of the collection. If the collection was expanded to cover decades, then greater fluctuation may be observed.

Two notable studies have explored time-aware word specificity models for IR. Liebscher and Belew (2003) argue that words whose frequency is high early in the period covered by a collection should be more favoured than more recently popular terms, however perform no IR experiments to conclude on their assumption. In contrast, Efron (2010) posits that temporal dynamics of a word's frequency can be used to better measure word specificity than non-temporal approaches (e.g. TF-IDF). More specifically, he finds that words with a future stable temporal dynamic, as predicted from past temporal dynamics, tend to be less discriminative for retrieval, and vice-versa. This corresponds with work presented in Chapter 6, which found stop words have few distinctive temporal dynamics because they are used consistently, regardless of time.

3.11.2 Explicit Temporal Clues: Temporal Expressions

Many documents contain explit temporal expressions relating to events being discussed. Conventional retrieval models will literally match these terms, disregarding any temporal hierarchy or specificity they imply.

For queries including a time period (e.g. "1990-1992" or "during 2000"), Khodaei et al. (2012) consider documents to be potentially relevant if there is overlap in the time periods expressed in the query and time periods expressed in the document content. Berberich et al. (2010) integrate temporal expressions into a time-aware probabilistic language model to improve retrieval effectiveness. Accordingly, queries containing temporal expressions such as "in the 1990s" will better match information containing one or more dates expressed in the relevant 1990s period. Strötgen et al. (2012) explore several additional characteristics of temporal expressions in documents, such as position, repetition, and hierarchy in order to determine the top relevant temporal expressions in a document during retrieval.

3.12 Time throughout Information Systems

There are of course several broader research fields concerned with general information management. Many of these diverse fields employ techniques rooted in the IR domain, most notably unstructured text similarity and ranking techniques. Given the ubiquity of time, many of these fields have also explored various time-aware approaches which build upon intuitions and techniques presented in the previous sections.

A great deal of engineering effort is concerned with building efficient, reliable and scalable search systems. Since much large-scale information behaviour is time-based, elements of time hold several implications for the caching strategies necessary to build efficient search engines (Baeza-Yates et al., 2007). Effective web crawlers must understand the constantly evolving and dynamic nature of the web, such as sources of important new information from communities in flux (Cho and Garcia-Molina, 2000). Temporal change necessitates intelligent crawlers which incrementally capture fresh information (e.g., news articles) which are likely to be required to satisfy forthcoming information needs (Chakrabarti et al., 1999).

Time is at the heart of information filtering (IF), where a time-ordered stream of incoming information objects is screened to pro-actively match relevant items to interested users. In IF, information topics and user interests change over time (Arampatzis, 2001). Similarly, information interpretation and novelty is inherently temporal in IF. The TREC Knowledge Base Acceleration evaluation task (Frank et al., 2012) resulted in many novel time-aware approaches to IF. Time-based adaptive term weighting models have been integrated successfully in IF models (Arampatzis, 2001).

Summarisation is the process of distilling information into a shorter yet still representative form. Many information streams such as news wires contains a great deal of redundant information which time-based summarisation can discard while maintaining the key points of the stories. The TREC Temporal Summarisation (i.e., TREC-TS) task (Aslam et al., 2013) developed an evaluation methodology for such a scenario. Various work has developed techniques to summarise the changes occurring over time to a single evolving web page, such as Wikipedia (Georgescu et al., 2013b; Jatowt and Ishizuka, 2004; Whiting et al., 2012b).

Somewhat related to summarisation, approaches for visualising time-based information have seen a great deal of interest. Timelines composed of time-anchored events and relationships are a popular tool for visually representing the flow of time and events (Alonso et al., 2009). Furthermore, timeline-inspired complex user interfaces have been developed to visualise both information change and related meta-data to uncover more subtle time-based phenomena such controversial change and debate (Wattenberg et al., 2007).

In multimedia retrieval, temporal distributions of user-generated multimedia tags (such as day time with sunshine, night time with stars, etc.) have been used to automatically expand tag coverage and improve image retrieval effectiveness (McParlane and Jose, 2013; McParlane et al., 2013).

Time has many implications in the accumulation, management and understanding of vast historic information archives, such as digital libraries and web archives (Kanhabua, 2009;

Masanès, 2007).

Text classification has long been an active area of research with wide applications. Temporal topic change has been studied in patent classification, and is viewed as a key requirement for reliable classification approaches given the progression in inventions over time (Mourão et al., 2008).

Recommender systems which pro-actively push expected relevant items to users have gained notable popularity in recent years. The role of time in user expectations and needs is studied from many aspects in recommender system approaches and evaluation (Koren, 2009; Lathia et al., 2010).

A great deal of research has sought to observe temporal correlations between real-world phenomena and information behaviour temporal dynamics, in order to facilitate future predictions. Ruiz et al. (2012) correlate Twitter discussion with stock trading volume and prices, and use the insight to enhance automated trading strategies. Others, such as Achrekar et al. (2011) have used large-scale information seeking trends and Twitter discussion to predict related flu trends.

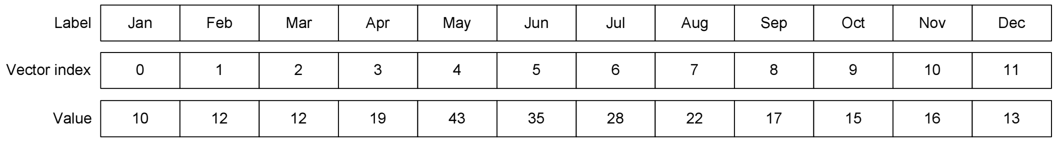
3.13 Working with Temporal Dynamics

In the previous sections I highlighted several temporal dynamics present throughout information and information behaviour. Temporal dynamics are realised from *time series* derived from aggregating raw temporal data streams, such as terms appearing in a time-based document collection (i.e., where the documents have a timestamp), or distinct queries and interactions recorded in search engine query logs. A temporal dynamic refers to the time-based patterns and trends evident in a time series. Each time series can be extracted from a retrospective collection, or, obtained in real-time. In the following section I formally describe how a time series is represented for processing.

3.13.1 Time Series Representation

Abstractly, a time series is a sequence of successive data points over time, where each data point represents an observation made during a uniform *time frame*. As such, data points are ordered chronologically.

For processing, a time series is represented as a vector data structure, as shown for a 12 month time series in Figure 3.9. The time series comprises of successive time frames, each referenced by a unique vector index offset. The *temporal resolution* of the time series is the time frame size used (e.g., n minutes, hours or days, etc.). Depending on the application,



Label	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Vector index	0	1	2	3	4	5	6	7	8	9	10	11
Value	10	12	12	19	43	35	28	22	17	15	16	13

Figure 3.9: Conceptual time series vector for 12 months of data at a monthly resolution.

a suitable temporal resolution can be selected based on the temporal coverage and change frequency necessary. I discuss the temporal resolutions selected for each experiment in this thesis in the relevant chapters.

Formally, time series X starting at time t_s, with temporal resolution t_r has ordered time frames $frame_i$ defined as follows:

$$\begin{aligned} X = (\ t_s \leqslant [frame_1] < t_s + (1 \times t_r),\ t_s + (1 \times t_r) \leqslant [frame_2] \\ < t_s + (2 \times t_r) \ldots t_s + (n - 1 \times t_r) \leqslant [frame_n] < t_s + (n \times t_r)\ ) \end{aligned} \tag{3.1}$$

Normalisation

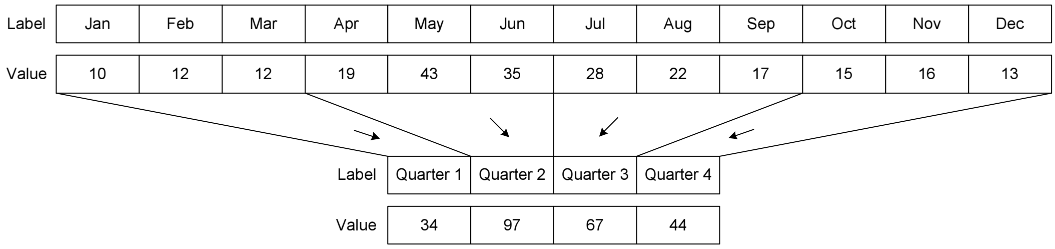
Background temporal effects such as international time zones, day/evening/night time and weekdays/weekends cause natural fluctuations for many phenomena – e.g., there is less social media discussion at night because users are sleeping. As such, for many applications a time series containing raw observations will need to be normalised to remove natural 'background' temporal dynamics, and expose unexpected temporal patterns and trends.

Normalisation expresses the observation in each time frame as a relative proportion of the overall sum of the phenomena observed in that time frame. In the case of the time series of a term's frequency in the collection, this means expressing the term's popularity in each time frame as a proportion of the sum of the frequency of all terms observed in the collection during the time frame. This is formalised generally as follows:

$$frame_{i',X} = \frac{frame_{i,X}}{\sum_{Y \in y} frame_{i,Y}} \tag{3.2}$$

Where $frame_{i,X}$ is the raw i'th time frame and $frame_{i',X}$ is the normalised i'th time frame of time series X. y is the set of all time series characterising the given phenomena (e.g., the time series for all terms in the collection).

Throughout this thesis I use the nomenclature X/Y, and X'/Y' to refer to two non-normalised



Label	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Value	10	12	12	19	43	35	28	22	17	15	16	13

Label	Quarter 1	Quarter 2	Quarter 3	Quarter 4
Value	34	97	67	44

Figure 3.10: Conceptual illustration of time series down-sampling from monthly to quarterly temporal resolution.

and two normalised time series of any given phenomena, respectively.

Down-sampling

While a time series can be initially represented in a high resolution, it is often desirable to reduce dimensionality of the observations to smooth erratic movements, and decrease the volume of data for storage and processing.

Figure 3.10 illustrates conceptual time series down-sampling from months to quarters. Down-sampling is achieved simply by uniformly aggregating (i.e., summing) contiguous time frames. Note that time series down-sampling must be performed prior to any normalisation.

3.13.2 Characterising Temporal Dynamics

A diverse tool kit of techniques rooted in statistics (e.g., time series analysis in finance and economics), physical sciences (e.g., signal processing) and data mining has emerged to explain and model stochastic processes characterised by time series data (Fu, 2011; Radinsky et al., 2013b). Common tasks include time series modelling for future forecasting, periodicity detection and correlation. This thesis employs several of these techniques; I describe the related approaches in the context of each relevant chapter.

3.14 Chapter Summary

The definition of time is simultaneously diverse and multi-faceted, as shown by its involvement in all user- and system-oriented aspects of the conceptual IR map. I outlined the nature of explicit temporal clues (i.e., temporal expressions) found in information content, as well as implicit temporal clues (i.e., temporal dynamics) arising from streams of information behaviour occurring over time. I explored the nature of time in information behaviour and time-aware IR. To that end, I proposed a conceptual map of time-aware IR, and used it to organise the diverse body of relevant theoretical and practical literature. As part of this model, I discussed and illustrated several real collection-based (e.g. word usage) and information seeking (e.g. query and intent popularity) temporal dynamics. Finally, I detailed the practical representation of temporal dynamics used by experiments later in this thesis. Overall, time poses many challenges, but also yields many opportunities for IR. Work presented in the remainder of this thesis serves as an exploration into methods to support and exploit one particular element of time – temporal dynamics – in IR systems.

Part II

Supporting Temporal Dynamics in Information Seeking

Chapter 4

Recent and Robust Query Auto-completion

In Chapter 3, I explored several temporal dynamics evident in information seeking activity, in particular, patterns and trends in query popularity over time. Consistently supporting users who are engaging in real-time information seeking driven by events and phenomena requires approaches which are sensitive to temporal dynamics. One of the foremost challenges for users during retrieval is formulating an adequate query to express their information need sufficiently to a retrieval system – and thus increase the chance of receiving satisfactory search results. In this chapter, in order to *explicitly* assist users to formulate queries which can better satisfy their information needs over time, I propose and experiment with novel time-aware approaches for query auto-completion (QAC) in web search – a ubiquitous activity performed hundreds of millions of times every day.[*]

4.1 Introduction

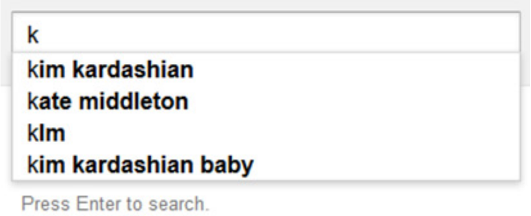
Cognitively formulating and physically typing search queries is an especially time-consuming and error-prone process. Spelling mistakes, forgetfulness and information need uncertainty often make textual query input laborious (Moshfeghi and Jose, 2013). In response, search engines have widely adopted QAC as a means of reducing the effort required to submit a query (Bar-Yossef and Kraus, 2011; Shokouhi and Radinsky, 2012). Indeed, beyond query input in search systems, text input auto-completion has also become popular in many other applications in which there is likely to be common input between users, such as inaccurate touch-screen text input, content tag selection, text 'hashtagging' (e.g., "`#topic`") and domain-specific search (e.g. maps, jobs and people). As the user types their query into the search box, QAC suggests possible queries the user may have in mind (which I refer to as

[*] Research presented in this chapter is published in Whiting et al. (2013a) and Whiting and Jose (2014).

completion suggestions), beginning with the currently input character sequence (i.e., *prefix*). Recent work has examined approaches for making QAC robust to spelling mistakes (Duan and Hsu, 2011) and term re-ordering.

The primary objective for effective QAC is to: (i) present the user's intended query after the fewest possible keystrokes, and (ii) at the highest rank in the list of completion suggestions. The most common approach to QAC is to extract past queries with each prefix from a query log, and rank them by their past popularity (Bar-Yossef and Kraus, 2011); this assumes current query popularity is the same as past query popularity. Although this approach provides satisfactory QAC on average, it is far from optimal since it fails to take into account clues such as time or user context which often influence the queries most likely to be typed. As a result, this thesis chapter explores QAC approaches which are sensitive to changing query popularity – where the popularity is not predictable from long-term past query popularity observations.

As the web increasingly becomes a platform for real-time activity, news and media, time plays a central role in information behaviour. A substantial proportion of the daily query volume is the result of users turning to search engines for information about recent and ongoing events and phenomena (Adar et al., 2007; Kairam et al., 2013; Kulkarni et al., 2011). Indeed, 20% of daily Google queries have not been seen in the past 90 days[1], with 15% have never seen before[2]. While the long-tail will inevitably account for a large proportion of these queries, many will be the result of short-term temporal events. We illustrate this with the following example.



k

kim kardashian

kate middleton

klm

kim kardashian baby

Press Enter to search.

Figure 4.1: Google auto-completion suggestions for the query prefix '***k***'. Screenshot taken September 23rd 2013, during the ongoing Westgate shopping mall terrorist attack in Kenya. Persistent browser cookies were cleared to avoid any individual personalisation effects.

Figure 4.1 shows the four completion suggestions offered by Google for the single character query prefix '***k***' on September 23rd, 2013. The list of completion suggestions indicates the historically most likely queries to be submitted with the prefix, possibly in the context of some

[1] http://googleblog.blogspot.co.uk/2009/12/this-week-in-search-121809.html

[2] http://investor.google.com/corporate/2013/founders-letter.html

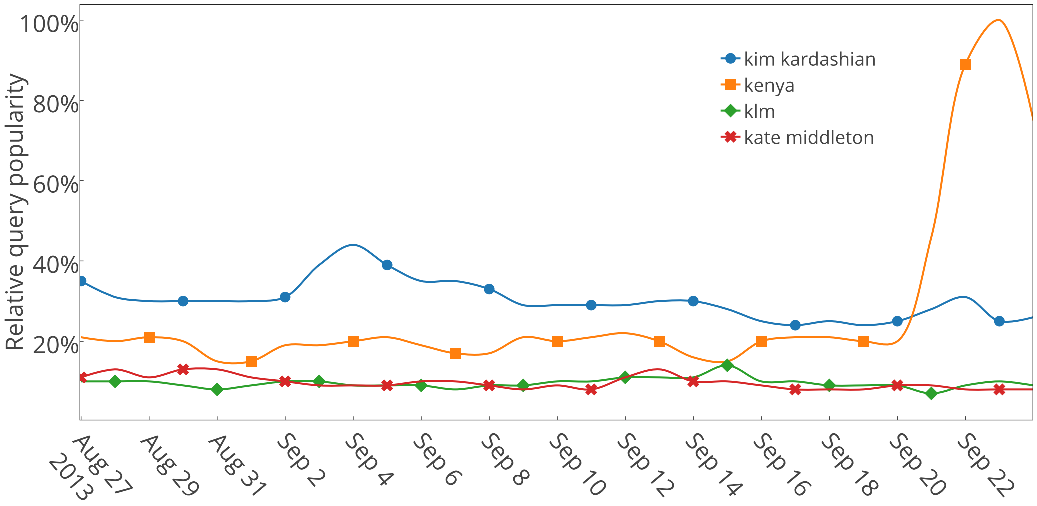


Figure 4.2: Google Search Trends indicating temporal popularity of the completion suggestions in Figure 4.1 during August and September 2013.

undisclosed ranking features such as user location. Despite the recency and prominence of the Kenya Westgate mall terrorist attacks, the query ‘kenya’ ranks very low in the completion suggestions. In Figure 4.2, I show the dramatic change in query popularity caused by the events – ‘kenya’ becomes by far the most popular query. Yet, despite the fact that ‘kenya’ is trending because of the ongoing events, Google’s QAC fails to support users searching for information about the event as it ranks completion suggestions based on the past query distribution - which is no longer appropriate. Further compounding this issue, QAC for short prefixes (i.e. 1-2 characters) is often unsuccessful as there are such a large number of possible completion suggestions (Bar-Yossef and Kraus, 2011). It is typically consistently popular ‘head’ queries that are provided as completion suggestions for such short prefixes (evidenced by the celebrity queries shown in this example). Therefore, there is a need to take the temporal aspect into account for effective QAC. This work is an attempt towards this objective.

Furthermore, queries that need to be included in completion suggestions fall into two main categories. The first category corresponds to *predictably* popular queries which are: (i) consistently popular, (ii) temporally recurring (e.g. at Christmas, in January, etc.) or (iii) known/foreseeable events and phenomena (e.g. TV episodes, sporting events, expected weather etc.). The second category corresponds to *unpredictably* popular queries related to entirely unforeseeable current events and phenomena (e.g. breaking news). Of course, although these may be unpredictable prior to the event occurring, once the query popularity is trending, then further popularity may become predictable based on short-range trends. Indeed, queries are likely to switch between these categories over time, making longer-term predictions problematic. Therefore, achieving optimal QAC effectiveness for all users, on average, is a trade-off between opposing objectives: (i) time-sensitivity, or *recency*, and (ii) *robustness*. Recency requires that completion suggestions include emerging and increasingly popular queries. Conversely, robustness requires that completion suggestions also reliably include long-term and consistently popular queries. These two goals are at odds – completion suggestions comprised only of short-term popular queries (e.g. in the last hour) might lead to lower ranking of many consistently popular queries. Alternatively, completion suggestions comprised of long-term popular queries (e.g. in the last year) will likely exclude the most recently popular queries. We propose an approach to address this trade-off by developing models which take recent evidence into account when possible.

4.1.1 Motivation

To be effective over time, QAC approaches must be time-sensitive since large-scale news, events and world phenomena play a central role in collective searching behaviour, and in turn dramatically affect query popularity (and hence the likelihood of a user typing a query) over time (Adar et al., 2007; Kairam et al., 2013; Kulkarni et al., 2011). While relying on a long period of past query log evidence to rank completion suggestions will ensure QAC is robust for consistently popular queries, it will also have the effect of smoothing over sensitivity to recently popular queries. For example, imagine a scenario in which a query, say q_1, is consistently popular, occurring 1,000 times every day in the query log. Aggregating query popularity over the past 30 day period would mean that a new query, say q_2, which became popular today, would need to occur 30,000 times before it outweighed the long-term popular query – despite the fact it is far more popular today than q_1 (and so should be a temporally higher-ranked completion suggestion). At the same time, reducing the aggregation period risks not fully representing the long-term q_1 query popularity, allowing arbitrary temporal fluctuations to reduce its ranking. Consequently, solutions addressing this problem are studied in this chapter.

Rather than relying passively on previously observed query popularity distributions for QAC, there is an opportunity to improve QAC over time by anticipating the current query popularity distribution based on previously observed trends – and better rank completion suggestions at each moment in time. As a result of this, in the past, long-term time-series modelling for query trends has been used to improve QAC for predictably recurring temporal query trends, such as seasonal (e.g., Christmas) or weekday/weekend related queries (Shokouhi, 2011). However, long-term time-series modelling can break down for many temporal events that are not always constant. For example, holidays (e.g., Easter) and natural phenomena such as the weather/seasons (e.g. planting spring bulbs) occur on an indeterminate schedule, where rigid temporal modelling of such queries might result in increased popularity prediction at incorrect times. Furthermore, for previously unseen and therefore unpredictable recent queries, time-series modelling often proved problematic due to lag and over-fitting (Shokouhi, 2011). Chang et al. (2014) note that time series which exhibit sudden spikes and heavy tails are often failed by most existing time-series models. This is due to the fact that time- series models often struggle to incorporate increasing trends quickly enough, and even then, continue to predict increased popularity for some time after brief periods of popularity. Aside from the prediction issues of long-term time-series modelling, I am motivated to explore alternative approaches since many QAC systems will not have long-term and in-depth past query logs available for robust modelling. Indeed, many organisations may only have more recent smaller query logs available, yet still need to 'bootstrap' effective QAC systems with far less data and resources. As such, in this work I take a conceptually different approach to anticipating current query popularity for effective completion suggestion ranking: (i) relying only on recent query popularity evidence, and (ii) short-range query popularity prediction based only on relatively recently observed trends.

4.1.2 Research Questions

In this chapter I investigate the following specific research sub-questions, with regard to **RQ1** of this thesis: "Query popularity exhibits many patterns and trends over time as common information needs change. *Can these temporal dynamics be supported for consistently effective query auto-completion over time?*".

- **RQ1.1:** How effective is QAC in diverse search scenarios?
- **RQ1.2:** Can relying only on recent query popularity evidence improve QAC effectiveness?
- **RQ1.3:** Can an optimal trade-off between recency and robustness be achieved for QAC?
- **RQ1.4:** Can short-range query popularity prediction (i.e., based only on recent trends) improve QAC effectiveness?

4.1.3 Chapter Outline

This chapter is organised as follows:

- Section 4.2 presents past work related to QAC effectiveness and efficiency, and in particular, improving QAC completion suggestion ranking.
- Section 4.3 outlines existing approaches to QAC (used as baselines), and proposes novel approaches for QAC.

- Section 4.4 describes the experimental setup and procedure used in this chapter, including query log datasets and real-time simulation methodology.
- Section 4.5 presents baseline and experimental QAC approach results, and discusses these results with regard to the four research sub-questions outlined in the previous section.
- Section 4.6 makes conclusions on the work and findings presented in this chapter.

4.2 Related Work

The majority of QAC research has centred on the inherent engineering complexity of providing efficient, responsive and scalable approaches that are resilient to typing errors (Bast and Celikik, 2011; Chaudhuri and Kaushik, 2009; Duan and Hsu, 2011; Fafalios et al., 2012; Kastrinakis and Tzitzikas, 2010).

Despite the prominence of QAC, there have been relatively few studies on improving QAC effectiveness; most likely down to the fact that there are few suitable query logs available outside industrial search engine companies for experimentation.

Exploiting the user's personal context and past query sessions has led to considerable increases in QAC effectiveness. For example, Shokouhi (2013) exploits user profiles in terms of a user's demographically biased topics to model the likelihood a user is to issue certain queries, and therefore personalise completion suggestion ranking (e.g., younger men searching for cars, older women searching for knitting, and so on). Similarly, Bar-Yossef and Kraus (2011) exploit common query activity between users to improve QAC for other similar users. In this work I focus solely on the temporal factor of QAC. Time and personalisation are not mutually exclusive, so I leave integrating temporal and personalised approaches to future work.

Shokouhi (2011) examines techniques for predicting query popularity based on previously observed temporal patterns. Extending this work, Shokouhi and Radinsky (2012) apply long-term time-series modelling of past temporal query patterns to improve QAC effectiveness for the proprietary Microsoft Bing query log. Popular queries recurring during specific temporal intervals, such as day/night, day of week, month, etc. were modelled to anticipate query popularity for completion suggestion ranking at different times. Shokouhi and Radinsky (2012) propose the short time window technique I experiment with in this paper as a hard baseline (MLE-W, which they refer to as p_1, p_3, etc.). They note its relative effectiveness, particularly for correctly predicting short-term highly temporal and unpredictable queries for which time-series modelling is problematic. However, no detailed analysis of the performance impact of the time window period for each prefix length is provided. With related temporal intuition, Sengstock and Gertz (2011) demonstrate, but not evaluate, a system which can rank completion suggestions depending on the time of day. In contrast to QAC approaches dependent on a past query log, Bhatia et al. (2011) propose techniques to extract common terms and phrases found in indexed documents and rank them for as completion suggestions.

Short-range query popularity prediction has seen little attention. Golbandi et al. (2013) develop a regression model to detect bursting queries for enhancing trending query detection. Meanwhile, Strizhevskaya et al. (2012) measured prediction accuracy of various daily query popularity prediction models for the proprietary Yandex query log.

4.3 Query Auto-Completion Approaches

In this section I formalise the problem of QAC, and propose four distinct approaches (and one 'meta' approach) to rank query completion suggestions based on query popularity observed over past periods of time. I focus independently on the recent temporal aspect of QAC in this work, and leave integrating user context, such as the user's past query topic preferences (Bar-Yossef and Kraus, 2011; Shokouhi and Radinsky, 2012) and location, to future work.

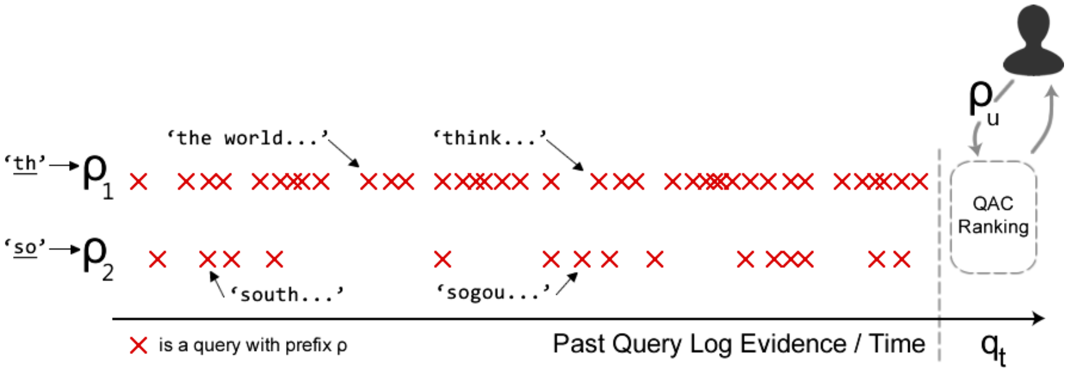


Figure 4.3: Conceptual query auto-completion approach based on previously observed queries, represented by each '×'.

QAC is formally defined as follows: a set of possible query completion suggestions S at time q_t are ranked for the user-provided query prefix q_ρ. Only signals prior to q_t (e.g., past query popularity, or the user's query history) are available for ranking completion suggestions. The top k completion suggestions are provided to the user with minimal latency for their optional selection. This scheme is conceptually illustrated in Figure 4.3 for two different two character prefixes: ρ_1 ('th') and ρ_2 ('so'). I use this conceptual diagram to visualise each approach presented in this section.

Two distinct approaches can be employed for completion ranking in QAC, in essence: (i) those which assume current query popularity is the same as recently observed query popularity, and (ii) those which predict current query popularity based on recent query popularity trends (i.e. the expected optimal completion suggestion ranking at q_t). In Sections 4.3.1, 4.3.2 and 4.3.3, I present approaches based solely on past query popularity distributions. Following these, in Section 4.3.4 I propose an approach which ranks completion suggestions based on short-range predicted query popularity. Finally, in Section 4.3.5, I propose a meta approach which uses online learning to adaptively optimise the parameters of the aforementioned approaches.

4.3.1 Maximum Likelihood Estimation (MLE-ALL)

The common and straight-forward approach to completion suggestion ranking is Maximum Likelihood Estimation (MLE), based on past query popularity obtained from prior query log evidence (Bar-Yossef and Kraus, 2011). In essence, MLE assumes that the current query popularity distribution will be the same as that previously observed, and so completion suggestions are ranked by their past popularity in order to maximise QAC effectiveness for all users, on average.

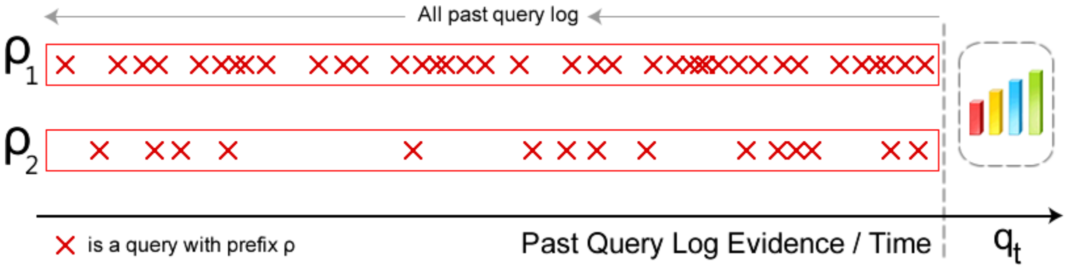


Figure 4.4: Conceptual query log and MLE-ALL query auto-completion approach.

In past work, MLE is used as a baseline and referred to as '*MostPopularCompletion*' (Bar-Yossef and Kraus, 2011; Shokouhi and Radinsky, 2012). I also use MLE, aggregating *all* query log evidence prior to the time of the user's query prefix input time (i.e., q_t) as our soft experimental baseline **MLE-ALL**, visualised in Figure 4.4.

MLE for prefix ρ and each query q with probability $P(q)$ in all past queries Q_ρ prefixed with the characters ρ, is formalised as:

$$MLE(\rho) = \underset{q \in Q_\rho}{\arg\max} P(q) \tag{4.1}$$

4.3.2 Recent Maximum Likelihood Estimation (MLE-W)

The MLE completion suggestion ranking approach introduced in the previous section relies on the query popularity aggregated since the start of the past query log. This QAC approach was first outlined by Shokouhi and Radinsky (2012) as a hard baseline.

Over time, the accumulation of query occurrences for popular queries smooths temporal variation, and thus promotes only the most consistently popular queries as completion suggestions. Recently emerging and popular queries will be outweighed by consistently popular queries, even though the emerging query might be considerably more popular at q_t.

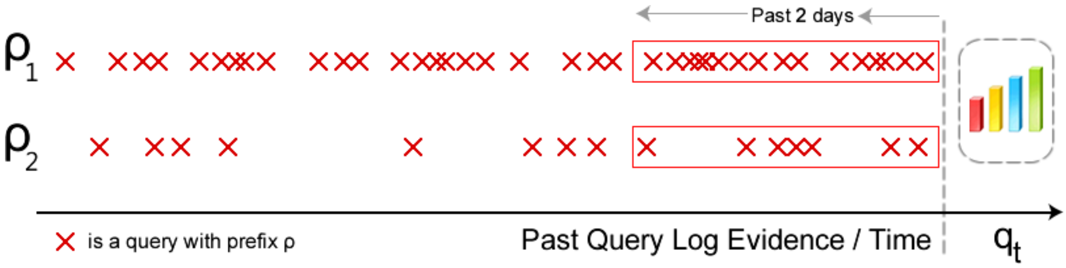


Figure 4.5: Conceptual query log and MLE-W query auto-completion approach.

To increase time-sensitivity I therefore propose using only the last N days of query log evidence, visualised in Figure 4.5. N = 2, 4, 7, 14 or 28 days is used for computing $P(q)$ at q_t (i.e., a single *sliding window period* of past query log evidence). I refer to this established approach as **MLE-W**, and consider it our hard baseline.

The intuition underlying this approach is that a limited recent period of queries may more accurately reflect the current query popularity distribution (Shokouhi and Radinsky, 2012). Similarly, although consistently popular queries will still be adequately reflected in the distribution, their total frequency will no longer be great enough to outweigh the frequency of popular queries that only burst for short periods.

4.3.3 Last N Query Distribution (LNQ)

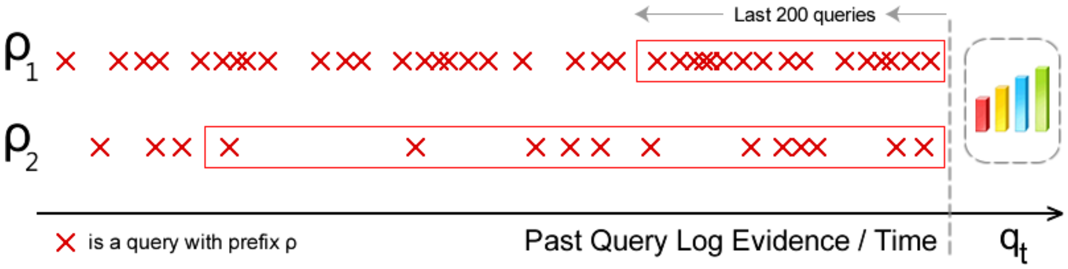


Figure 4.6: Conceptual query log and LNQ query auto-completion approach.

Algorithm 1 Last N Query Distribution (LNQ) for query prefix ρ

```
Require: $N$ = number of last queries to track, $n$ = flood limit for duplicate queries, $QueryStream_\rho$ = chronologically ordered stream of queries with prefix $\rho$.
$Deque_\rho \leftarrow$ FIFO deque for prefix $\rho$        ▷ Instantiate deque for queries with prefix $\rho$
$DequeSize_\rho \leftarrow 0$        ▷ Track number of queries in the deque
for all query $q$ at time $q_t$ in $QueryStream_\rho$ do
        Output completions at $q_t \leftarrow$ MLE($Deque_\rho.Contents()$)        ▷ Output completions, but continue algorithm execution

    if $Deque_\rho.Count(q) + 1 <= n$ then        ▷ Ensure $n$ for $q$ not exceeded
        $Deque_\rho.Push(q)$        ▷ Add query to deque
        $DequeSize_\rho \leftarrow DequeSize_\rho + 1$        ▷ Increment queries in deque
    end if

    if $DequeSize_\rho > N$ then        ▷ Ensure $N$ not exceeded
        $Deque_\rho.Eject()$        ▷ Remove oldest $q$
    end if
end for
```

Prefix popularity, like many linguistic phenomena (such as word and phrase use in corpora) exhibits a power-law (or, *long-tail*) frequency distribution. As such, there are a small number of prefixes that are very common (e.g. '`the`' for a 3 character prefix), and a large number of prefixes that are very uncommon (e.g. '`zzt`'). In early preliminary work, I found that although imposing a short sliding window for MLE-W could marginally improve QAC effectiveness overall, it often harmed completion ranking for less common long-tail prefixes – which together represent a large proportion of overall queries. As common prefixes account for a large volume of queries, query popularity distributions will be adequately reflected in a short sliding window. However, for less common prefixes, a short sliding window may reduce the available query popularity evidence over time - leading to poor completion suggestion ranking based on only the most recent and potentially randomly occurring query distributions, rather than robust longer-term distributions.

To address this issue, I investigated assigning sliding windows to different prefixes based on their popularity. Common prefixes were assigned a shorter sliding window (e.g., 2-7 days), while less common prefixes were assigned longer sliding windows (e.g., 7+ days). Empirical experiments found this approach problematic as the optimal sliding window identified for each prefix during the training period often changed since the prefix popularity can also change dramatically over time (for instance, if a query with a less common prefix becomes popular). I propose a continuous parameter learning meta approach to tackle this issue in Section 4.3.5.

Alleviating the need to impose any rigid sliding window time period, I propose the Last N Query Distribution approach, visualised in Figure 4.6. LNQ tracks the last N queries observed for each prefix, and exploits their popularity distribution to rank completion suggestions at q_t in the same fashion as MLE (i.e., by most popular completion). This technique is inherently adaptive as it does not impose any strict time cut-off for past evidence. The trade-off between robustness and recency is controlled by capacity parameter N – a greater

N promotes robustness by ranking completion suggestions using more past evidence, and conversely, a smaller N encourages recency by only using the most recent queries for the prefix in completion ranking.

An extremely popular query, especially for very short prefixes (e.g. 2 characters) may drive out other less common queries when N is relatively small. To avoid this, I introduce a second 'flood limit' parameter to the model, n, to constrain the upper limit of duplicate query instances to track (i.e. the maximum number of times the query '`google`' for prefix '`go`' is considered in the past N queries). When $n \geq N$, there is no upper limit on any single query.

Despite the practicality of this approach, to the best of our knowledge it has been neither proposed nor studied in past literature. I refer to the Last N Query Distribution approach as **LNQ**.

Practical Implementation

LNQ is implemented by maintaining an adapted first-in-first-out (FIFO) double-ended queue[1] with capacity N for each prefix.

As each query is observed for a prefix (i.e., after the user has submitted the query), it is pushed to the deque, so long as the constraint defined by n is maintained. If the deque exceeds N, the oldest query observed is removed. Abstract implementation of LNQ is presented in Algorithm 1.

4.3.4 Predicted Next N Query Distribution (PNQ)

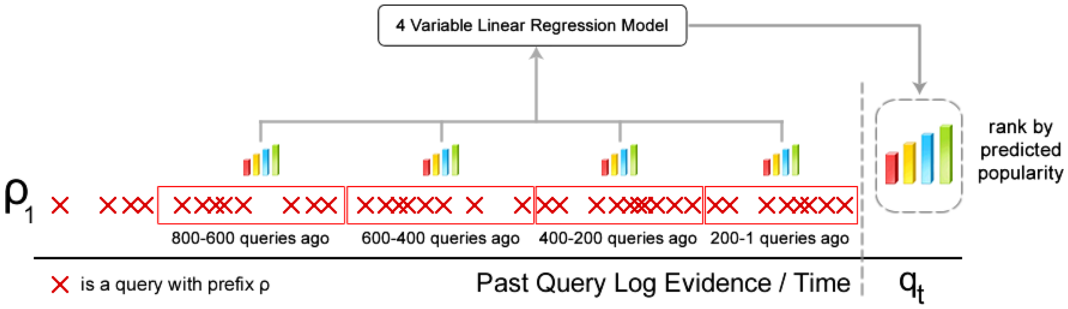

4 Variable Linear Regression Model
rank by predicted popularity
ρ1
800-600 queries ago
600-400 queries ago
400-200 queries ago
200-1 queries ago
× is a query with prefix ρ
Past Query Log Evidence / Time
qt

Figure 4.7: Conceptual query log and PNQ query auto-completion approach.

As query popularity distributions are not stationary over time (Kulkarni et al., 2011; Shokouhi, 2011; Shokouhi and Radinsky, 2012), the query popularity distribution observed in the past, no matter how recent, will in many cases be different to the actual present distribution at

[1] Referred to as a *deque* data structure by Knuth (1997)

q_t. Consequently, for queries which fluctuate in popularity over time, QAC based passively on past query popularity distributions may always have a lag in effectively suggesting and ranking completions.

Rather than assuming present query popularity distributions are the same as those previously observed, I propose to predict short-range (i.e., current) query counts using recently observed query popularity trends (e.g. stationary/rising/falling query counts), visualised in Figure 4.7. In contrast to past work (Shokouhi and Radinsky, 2012), rather than exploiting long-term repetitive trends (e.g., Christmas, New Year, Easter, etc.), this work relies on predicting current query popularity based only on the most recently available data.

Time-series modelling techniques have been used to predict query popularity. Auto-regressive time-series modelling has proven effective for long-term predictably recurring queries (Shokouhi and Radinsky, 2012; Strizhevskaya et al., 2012), e.g., those likely to be seen daily, monthly, quarterly, etc. Such techniques require a vast amount of past data to avoid overfitting, fall down with indeterminate time schedules and are often ineffective for unforeseeable newly popular queries. Trending query detection can be improved through query count prediction using minute-based linear auto-regression models for the 30 minutes prior to q_t (Golbandi et al., 2013). Of course, trending queries are concentrated in the very recent query log, making such specific short-range prediction feasible.

Time-series query popularity prediction approaches require adequate query popularity data to construct a robust time-series model which can make reliable predictions. However, since query popularity exhibits a power-law distribution, the majority of queries are relatively low frequency, and subsequently have relatively sparse time-series, making accurate predictions problematic. Yet, these queries still need to be effectively ranked in QAC.

Henceforth, to side-step the issue of sparse and potentially erratic query popularity time-series, I frame the problem of query popularity prediction for completion suggestion ranking as: *given trends in the last N queries observed with a prefix, what is the chance of observing the query again in the next N queries with the prefix?* The last N queries may have been observed in the last 5 minutes, or 5 hours, and so forth, depending on the prefix popularity. Although this approach is still time-aware, it is not dependent on query popularity in a rigid temporal scale, but instead the rate of all queries with the prefix observed over time. I note my approach is not exclusive to long-term temporal modelling (Shokouhi and Radinsky, 2012), and I leave combining the two distinct approaches to future work[1].

[1] Currently, the publicly available short-term query log data severely limits any longer-term temporal modelling opportunities.

Using the predictive model outlined in the following section, this approach ranks completion suggestions based on their predicted popularity distribution at q_t. I refer to the Predicted Next N Query Distribution approach as **PNQ** in Section 4.5, where I note predictive model accuracy and present QAC experimental results.

Incrementally-trained Predictive Model

To capture recent periods of query popularity, a sequence of successive LNQs is employed. Rather than using a single LNQ containing the last N queries observed with a given prefix (as in the previous section), I chain multiple LNQs together to track windows of the last N queries with the prefix. For example, visualised in Figure 4.7, are four $N = 200$ LNQs used to track the last 200, 200-400, 400-600 and 600-800 queries recently observed for a given prefix (i.e., tracking the trends in overall, the last 800 queries observed with a given prefix). The recent query count distribution contained in these four LNQs is used by the regression model to predict upcoming query popularity counts for each distinct query in the next 100 queries with the given prefix. This predicted query count is thus used for ranking the completion suggestions at time q_t. In Section 4.5 I denote the successive LNQs used for the regression model in PNQ experiments with the notation $M \times N$, where M is how many LNQs are chained together, and N is the capacity of each separate LNQ.

I propose a multiple linear regression model based on the query counts observed in the successive LNQs, to predict the count of recently observed queries in the next 100 queries with the same prefix. In this work I train a single regression model for all prefixes – future work will explore a regression model for each prefix, thereby reflecting its innate temporal characteristics. Several regression models were tested in early empirical studies, however multiple linear regression was selected because it was the best performing approach. Future work will explore more elaborate and time-aware non-linear regression models.

Conventional machine learning techniques rely on batch learning – in which a finite and stationary set of training examples is used to produce a static model which best satisfies the optimisation objective, e.g. minimising a cost function. For a regression model, this often means using the standard least squares method. The linear least squares approach minimises the squared difference between the target and predicted variable over all available training instances to fit appropriate model parameters. This approach is perfectly adequate when the number of training instances fits in memory, and the nature of the model is unlikely to change. However, for query popularity prediction, there is a large-scale online stream of training instances with unforeseen latent and interacting temporal factors which may affect a query popularity model, such as the time of day (Golbandi et al., 2013), day of week, public holidays and weather, etc. Hence, a model trained in one time period may fail to generalise to a later time period.

Recent research attention has focused on developing machine learning techniques (e.g. for classification, regression and ranking) that can incrementally refine their model to constantly maintain and improve accuracy based on new training instances seen in an evolving data stream (Zliobaite et al., 2014), such as that available from a query log. These techniques are deemed *active* learning approaches.

Conceptually, in an online scenario such incremental models still provide predictions based on past training (although, more recent learning may take precedence), however, they continuously incorporate training instances as they become available to further reduce error through feedback. From a practical point of view, this strategy for model learning offers two major benefits in a query popularity prediction scenario. Firstly, the model can be incrementally fitted as new training data becomes available, relieving the need to store all past training examples to re-learn the model from scratch each time it needs updating – which would be prohibitively slow. Secondly, since there is a large volume of training data, in-memory optimisation is infeasible, so incremental approaches offers a means of scaling up the learning to all available training data.

In order to learn a model for query popularity prediction, I employ Stochastic Gradient Descent (SGD) (Zhang, 2004). SGD incrementally fits a regression model by reducing prediction error following each observed training instance. Full details on efficient SGD model training implementations can be found in Zhang (2004). I summarise SGD for the purpose of this application below.

The SGD cost function, $cost(\theta, (x^{(i)}, y^{(i)}))$, is modified from batch gradient descent, such that the cost of model parameter θ is computed for each specific training example (rather than over the entire training set simultaneously):

$$cost(\theta, (x^{(i)}, y^{(i)})) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{4.2}$$

Following the ith instance, the SGD algorithm updates the model parameter θ_j based on the difference between the predicted and actual instance value (i.e., *gradient*), and the learning rate. This process is formalised by Algorithm 2.

I set a constant SGD learning rate of 0.01. Since SGD is sensitive to scaling (Zhang, 2004), query counts are normalised between $[0, 1]$. Traditional k-fold cross-validation is not applicable to streaming settings since it would disorder the temporal data sequence (Gama et al.,

Algorithm 2 Stochastic Gradient Descent (SGD) Algorithm

```
Require: $m$ = number of training instances, $n$ = number of instance features, $\alpha$ = learning rate).
for $i := 1, 2 \ldots m$ do
    for $j := 0, 1 \ldots n$ do
        $\theta_j := \theta_j - \alpha(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$
    end for
end for
```

2014). Re-ordering the input training set is commonly used in SGD training to ensure that overfitting to a specific training instance sequence does not occur. This approach is not logical in this scenario, since there is natural ordering in the time-based stream of past queries and furthermore, I continue learning for the full duration of the dataset.

PNQ Model Training Routine

For PNQ, the following routine is used to record training data from the stream of queries, and train the query popularity prediction model using SGD:

1. After every 100 queries observed with a given prefix, the successive LNQ past query counts for each respective q with the prefix are stored as a future training instance.

2. Following a further 100 queries, the prediction target variable – each q's count in the following 100 queries with the prefix, is now known.

3. SGD adapts the linear regression model co-efficients based on each q training instance.

4.3.5 Online Parameter Learning (O-...)

All the experimental approaches I have outlined in the past three sections (i.e. MLE-W, LNQ and PNQ) have one or more parameters that require tuning for optimal QAC performance (e.g. window size for MLE-W, N/n for LNQ, and $M \times N$ for PNQ).

In Section 4.5, I explore the parameters necessary to achieve best performance for each prefix length and query log, on average. Of course, since different prefixes have varying characteristics (recall the power-law distribution of prefix popularity I discussed in Section 4.3.3), prefixes of the same length (e.g., '`th`' and '`zh`') may have different optimal QAC approach parameters, and these could change over time as new behaviour emerges (e.g. a major new event causing a great deal of change for a previously stationary prefix).

The final approach (in essence, a *meta*-approach) I propose in this chapter is based on online parameter learning for the aforementioned approaches. This requires that the best performing parameters for each approach, and prefix, are continuously re-evaluated and selected to improve QAC at q_t. To this end, I exploit the nature of QAC whereby the correct completion suggestion is known almost immediately after suggestions were initially provided – whether because the user has selected it, or manually typed it. This characteristic enables online measurement of the effectiveness of each parameter, in real-time, and thus facilitates online QAC optimisation. Since all the QAC approaches I propose are light-weight, this approach is still very practical in real-world applications.

Online parameter learning tracks the QAC performance (i.e., the mean reciprocal rank, discussed in Section 4.4.2) of each individual approach parameter for each prefix, in a training horizon of the last k queries observed with that prefix (from hereafter referred to as Δ). At q_t, the best performing parameter is used to rank the completion suggestions for the prefix, hence, the parameters are free to change if another becomes optimal over time.

Since I use the previously proposed approaches in combination with this meta-approach, I refer to these as **O-MLE-W**, **O-LNQ** and **O-PNQ** in Section 4.5.

4.4 Experimental Setup

I conduct experimentation using AOL 2006, MSN 2006 and Sogou 2008 query log datasets. By experimenting with the millions of queries contained in each query log, I obtain a representative view of how each approach would perform in a real-world setting. Comparing results from multiple query logs validates QAC approaches in different scenarios in which there may be diverse user population biases, temporal behaviours and sparsity levels.

4.4.1 Query Log Datasets

AOL (Pass et al., 2006), MSN (Craswell et al., 2009) and Sogou 2008[1] are publicly available[2] query log datasets. Each dataset is provided in a raw format containing all user interactions in chronological order. Each interaction can represent a query submission (with no result clicks), a result click or result page change, etc. Each interaction in the dataset therefore duplicates the query originally typed for identification. For realistic QAC experimentation I extract only the queries a user physically input (which I refer to as *typed queries*).

I consider the first appearance of a query in a user's session (and the associated timestamp) as typed queries. Where not provided in the dataset, sessions were identified using a 30 minute interaction time-out (Catledge and Pitkow, 1995). Details of each query log and its associated typed queries dataset are provided in Table 4.1.

[1] `http://www.sogou.com/labs`

[2] MSN available on request from (Craswell et al., 2009).

Table 4.1: AOL, MSN and Sogou query log dataset statistics.

	AOL	**MSN**	**Sogou**
Language	US English	US English	Simp. Chinese
Start date	2006-03-01	2006-05-01	2008-06-01
End date	2006-05-31	2006-05-31	2008-06-30
Training days	28	14	14
Testing days	64	17	16
Days covered	92	31	30
Typed queries			
Total	18.1M	11.9M	25.1M
Avg. per hour	11,764	29,069	65,364
Avg. per day	196,411	382,589	836,660
Query length (chars)			
Avg.	17.03	17.38	6.53
Stdev. (±)	11.02	12.15	4.01

AOL Issues

Early analysis discovered a relatively large number of short bursts of suspected bot spamming activity in the AOL query log. Many generic queries (e.g., ‘`personalfinance`’, ‘`aolcelebrity`’ and ‘`computercheckup`’ – both with and without spaces) occur with remarkably uniform interaction spacing (e.g. every 30-60 seconds).

Although AOL contains more queries than MSN, it also has greater breadth by covering a three month period. However, the dataset is sampled by user, rather than randomly over all queries (Adar et al., 2007). Therefore, the breadth of coverage may have biased querying distributions since many web search users engage in information re-finding behaviour (as many as 40% of queries are for re-finding) (Teevan et al., 2007). Moreover, there is a missing day of data – May 17th (Adar et al., 2007).

Comparing Datasets

Although the exact sampling and construction of each query log is unknown, together these three datasets have differing characteristics that allow us to make in-depth comparisons between QAC approaches. AOL seems much more raw compared to Sogou and MSN, which have both been filtered to remove identifying information and adult queries. Spam issues I highlighted for AOL may prove problematic for any results based on it alone. As MSN and Sogou provide one month datasets, they offer greater daily depth in comparison to AOL’s breadth over three months.

Because AOL and MSN are English language query logs, they contain queries in the 26-character Latin (i.e., *a*-*z*) alphabet. These queries are on average around 17 characters in length. In contrast, Sogou queries are almost all in the simplified Chinese alphabet. This alphabet has around 3,500 commonly used characters accounting for 99.5% of all characters used (Da, 2004), although search query language distributions often differ (Chau et al., 2009). Accordingly, Sogou queries are on average only 6.5 characters in length, as each character carries much greater information than the Latin alphabet. As the proposed QAC approaches are blindly based on single characters and the temporal dimension, the alphabet/language of the queries is irrelevant.

Temporal Trends

Many studies have offered extensive insight into the temporal trends of longer-term but proprietary query log data (Beitzel et al., 2007; Kulkarni et al., 2011; Shokouhi, 2011). In the experimental datasets I identified many popular queries with the majority of their occurrences concentrated within short periods - that is, queries related to unpredictable ongoing events and phenomena which need to be included in completion suggestions. In AOL, popular short-term temporal queries include: '`amelia earhart pictures`', '`karl der grosse`', '`the simpsons live action`' and '`leisure suit larry`'. Likewise, in MSN among the most popular are: '`stephen colbert`', '`poison milk`', '`ohio bear attack`' and '`kimberley dozier`'. I expect that the proportion of unpredictable recent queries will have substantially increased in more recent query logs, given the rise in real-time internet media.

4.4.2 Experimental Procedure

In this section I outline the experimental procedure I use to investigate the four research questions proposed in Section 4.1.

Real-time QAC Simulation

Previous work has used past query logs as the ground truth during simulated evaluation (Bar-Yossef and Kraus, 2011; Shokouhi and Radinsky, 2012). My experimental methodology also employs this strategy. To that end, I simulate a real-time user search scenario, such that when the user types a prefix, they receive completion suggestions based *only* on evidence observed prior to the time of their input. QAC effectiveness is measured by the presence and rank of a ground-truth match in each set of completion suggestions.

The chronologically ordered query log datasets (discussed in the previous section) provide a stream of ground-truth user-typed queries. I assume that the user would have selected their completed query if it had been presented to them. QAC approach performance is evaluated over *all* typed queries in these datasets - *not* just a subset. This ensures a representative evaluation of QAC effectiveness in a real scenario. I discuss matching ground-truth queries and completion suggestions, and measuring the effectiveness completion suggestions, in the following section.

I begin by evaluating baseline approaches for all 2- to 5-character query prefixes. I find sparsity caused by the long tail distribution of prefix popularity is compounded by the sampled experimental query logs for more specific prefix lengths of 4 or more characters. In these cases, time has less significant effect on the rankings since it sufficiently narrows down the completions suggestion desired – at least over the shorter periods of time I am limited to experimenting with. Consequently, my experimental QAC approach evaluation concentrates on 2-3 prefix characters, where temporal change causing optimal completion suggestion ranking changes is more pronounced with the greater space of possible completion suggestions. In any case, longer and therefore rarer prefixes pose several challenges that require a different set of techniques for effective completion suggestion ranking (Mitra and Craswell, 2015). Completion of ordered and un-ordered separate word or phrase prefixes is also left to future work.

Experiments for each prefix length were run independently, hence we assume a successful completion suggestion for a 2-character prefix has no bearing on the later evaluation of a 3-character prefix. Further work is needed to examine the conditions of this assumption. In particular, it is reasonable to expect that a user may choose to type another character to bring a lower ranked completion suggestions to the top ranking, if the combined effort of doing so is still less than selecting the lower ranking. This user model may vary between users, and indeed devices based on the physical effort of reading the display and providing further input. A deeper understanding of this user model is likely to open up several opportunities in aspects such as novelty (or conversely, redundancy), diversity and readability in QAC approaches.

To emulate a real user interface scenario (such as that shown in Figure 4.1), experiments are based on the user seeing at most the four highest-ranking completion suggestions for each prefix they input.

Evaluation Metric

As in past QAC work (Bar-Yossef and Kraus, 2011; Shokouhi, 2013; Shokouhi and Radinsky, 2012), I rely on Mean Reciprocal Rank (MRR) to measure the effectiveness of each QAC approach. Reciprocal Rank (RR) has typically been used for evaluation in retrieval situations in which there is a single relevant ranked item (Croft et al., 2010). MRR reflects the user interaction model of QAC: a higher-ranked completion suggestion is more beneficial, but the difference in ordering of lower-ranked completion candidates is less significant. That is to say, there is less noticeable difference between a correct completion suggestion ranked at either the 3rd or 4th position, compared to the 1st or 2nd position. For a set of completion suggestions S and the user's intended (i.e., completed) query q', RR is computed as:

$$\mathrm{RR}(\mathrm{q'}, \mathrm{S}) = \frac{1}{Rank(q' : S)} \tag{4.3}$$

If no match for q' is present in the ranking, RR is by default $= 0$, ensuring no divide-by-zero computational errors occur. MRR is computed as the arithmetic average of RR for all queries (thus it is in the range $[0, 1]$). Because of the large number of queries over which I compute the MRR (i.e., 12-25M queries), it is relatively insensitive, so I round MRR values to 4 decimal places.

Because of the non-linear discounting of RR, a small percentage increase in MRR can be the result of a much larger percentage increase in completions with an RR > 0 (depending on the distribution of RRs obtained for each completion rank). In terms of reputation for a real-world QAC system, I argue that even showing an appropriate completion suggestion at a lower rank will be valuable, so even small improvements in RR are noteworthy.

I consider a literal lower-case exact string match between completion suggestion and ground-truth as a successful match. Although there often may not be an exact literal match between the user's intended (ground-truth) query and an appropriate completion suggestion (e.g., '`A.A.`', '`AA`' and '`American Airlines`' are semantically equivalent). Similarly, the ground-truth query might be a spelling mistake which would not have occurred if the user had been given the correct completion. I leave this advanced evaluation to future work.

4.4.3 Experiment Settings

In the following sections I outline experimental settings used in the real-time QAC simulation experiments.

Baselines

As in recent past work (Bar-Yossef and Kraus, 2011; Shokouhi and Radinsky, 2012), I consider MLE-ALL (i.e., '*MostPopularCompletion*') as our soft baseline. Early results (reported in Appendix A) showed MLE-W almost always outperforms MLE-ALL, so I considered it as an appropriate hard baseline – with which I compare all the experimental approaches in Section 4.5.

I am unable to reliably re-implement past long-term temporal modelling as a QAC baseline (Shokouhi and Radinsky, 2012), as the query logs I have available are all very short-term compared to the proprietary multi-year datasets the paper authors utilised. As I would need to separate the datasets into train and test periods (and only test on periods after training for temporal validity), over-fitting to a small training set, or alternatively, relying on a very small and potentially unrepresentative test set would be inevitable. In any case, I consider work relying on short-term query trends to be complementary to long-term query trend modelling, and expect future work to investigate combining the techniques.

Query Log Training Periods

Since queries vary over time, QAC performance can also fluctuate. It is therefore unfair to compare two approaches using the same query log over different time periods. Consequently, *all* QAC approach comparisons I make (e.g., relative MRR change) are computed over the same query log time periods (and thus, the same set of test queries) to ensure robust findings.

In most real-world scenarios, a pre-existing query log would be used to train (or, *bootstrap*) a QAC system prior to use. In line with this, I dedicate the first 14 days of Sogou and MSN, and 28 days of AOL to this purpose. The remaining duration of the query log is used for testing. Accordingly, all MRR evaluation results presented are computed for completion of all queries from 00:00 on 2006-03-29, 2006-05-15 and 2008-06-15 onwards for the remaining duration of AOL, MSN and Sogou, respectively. This approach is visualised conceptually in Figure 4.8 for the AOL query log.

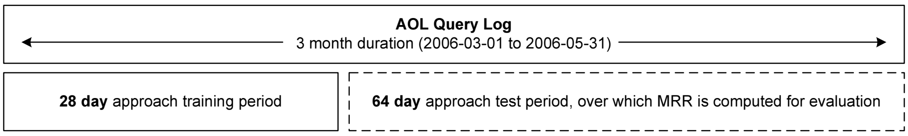

AOL Query Log
3 month duration (2006-03-01 to 2006-05-31)
28 day approach training period
64 day approach test period, over which MRR is computed for evaluation

Figure 4.8: QAC approach training and testing time periods for the AOL query log. MRR is computed for queries in the test period only, ensuring results between QAC approaches are comparable.

4.5 Results and Discussion

I conducted a large number of experiments to test QAC approaches with several parameters for three query logs and prefixes of 2-5 characters. In this section I present the most interesting positive and negative results that experimentally validate the approaches we propose,

Table 4.2: **MLE-ALL** Approach (Soft Baseline) - QAC MRR observed after initial training periods of 14 days (MSN/Sogou) and 28 days (AOL), with query prefix ρ lengths of 2-5 characters.

	Query Log / MRR		
ρ	**AOL**	**MSN**	**Sogou**
2	0.0962	0.1124	0.4117
3	0.1527	0.1702	0.5487
4	0.1969	0.2106	0.5970
5	0.2304	0.2340	0.6129

and offer insight for real-world QAC system design. Section 4.5.7, I draw on these results to discuss the four research sub-questions set out for this chapter in Section 4.1.2.

Statistical Significance

The aggregate statistical power of MRR measured over tens of millions of queries in the datasets with which I experiment means that all the results I report are statistically significant according to common *t*-tests (Ellis, 2010). My analysis therefore concentrates on the effect size over the baseline of each approach – that is, relative MRR change.

In Table 4.2, I report the absolute MRR achieved by MLE-ALL QAC for prefixes of 2-5 characters for each query log.

4.5.1 Soft Baseline: MLE-ALL

Most apparent, for all query logs, is that QAC is considerably more effective with a longer (i.e., more specific) prefix. This is unsurprising, given that each extra character in the prefix vastly reduces the space of possible completion suggestions, therefore increasing the chance of a completion suggestion match (Bar-Yossef and Kraus, 2011).

QAC is always more effective for MSN than for AOL, although at 4-5 characters their performance begins to converge. Expectedly, QAC is much more effective for Sogou – attributable to the greater specificity of each character in the larger Chinese alphabet (and hence, shorter Chinese queries).

4.5.2 Hard Baseline: MLE-W

In Table 4.3, I report the MRR change achieved by MLE-W QAC relative to the MLE-ALL soft baseline, with a 2- to 28-day sliding window period for prefixes of 2-5 characters. A

Table 4.3: **MLE-W** Approach (Hard Baseline) - QAC MRR change (relative to the MLE-ALL soft baseline) observed for past 2- to 28-day sliding window periods of query log evidence, with query prefix ρ lengths of 2-5 characters for AOL, MSN and Sogou query logs. Best performing sliding window periods are highlighted.

	Query Log / Sliding Window Period (days)				
ρ	2 days	4 days	7 days	14 days	28 days
			AOL		
2	-1.12%	+0.47%	**+1.12%**	+1.03%	+0.69%
3	-11.18%	-4.91%	-2.01%	-0.45%	**+0.14%**
4	-19.58%	-11.46%	-6.71%	-3.14%	**-1.04%**
5	-26.00%	-17.01%	-11.16%	-5.97%	**-2.39%**
			MSN		
2	+2.60%	**+2.95%**	+2.28%	+0.68%	-
3	-5.13%	-1.50%	-0.11%	**+0.27%**	-
4	-11.05%	-5.77%	-2.56%	**-0.58%**	-
5	-15.84%	-9.45%	-4.97%	**-1.47%**	-
			Sogou		
2	**+3.59%**	+3.36%	+3.03%	+2.47%	-
3	-5.46%	-2.96%	-1.52%	**-0.19%**	-
4	-7.81%	-4.86%	-2.90%	**-0.87%**	-
5	-10.22%	-6.60%	-4.11%	**-1.37%**	-

28-day sliding window period MRR is not reported for MSN and Sogou because they do not cover enough time to reliably compute it.

In all cases, using a sliding window period of past evidence has a considerable effect on overall QAC performance compared to the soft baseline. For AOL there is a sliding window period which can improve QAC performance over MLE-ALL for prefix lengths of 2-3 characters, albeit only marginally for 3 character prefixes. A 7 day sliding window period improves QAC performance by up to 1.1% for shorter prefixes of 2 characters.

Meanwhile for MSN, QAC for 2- or again, marginally, 3-character prefixes outperforms the soft baseline when using a small sliding window period. Specifically, I see the best performance improvement, of almost 3%, when using a 4-day sliding window.

Finally, for Sogou, although QAC performance increased by up to 3.6% for 2-character prefixes with a 2-day sliding window period, it is always harmed for longer prefixes.

For all query logs, however, using a sliding window period between 2- and 28-days always impairs QAC performance compared to the soft baseline for 4- or 5-character prefixes. This is because the sliding window filters out infrequent potential completion suggestions for less common prefixes, thereby reducing the evidence available for completion suggestion ranking. However, the detrimental effect of a sliding window on QAC performance reduces as the sliding window time period is increased. I expect experiments on longer-term query logs are likely to find optimal sliding window periods greater than 28 days which are capable of improving QAC performance for these longer prefixes. As I am limited to the shorter-term experimental query logs, I concentrate on 2-3 character prefixes where time has a more immediate impact since the space of possible completions is much wider for such non-specific prefixes (and hence QAC ranking changes are more necessary). As such, all further results of the proposed approaches reported in this section concentrate on 2- and 3-character prefixes, and are reported relative to the highlighted hard baseline performance[1].

4.5.3 Approach: LNQ

In Table 4.4, for clarity I report the MRR change achieved by LNQ QAC relative to the MLE-W hard baseline (highlighted in Table 4.3) for prefixes of 2-3 characters. As a result, in this and further results tables, a positive percentage is reflective of an improvement in QAC performance over the hard baseline. I experiment with an LNQ capacity parameter of $N = 100, 200, 400, 800, 1200$. Preliminary experiments showed that n (i.e., the flood limit parameter, proposed for avoiding extreme spiking queries or spam saturating the LNQ and making it ineffective) had a negligible effect on my results (at least when n was $\geqslant 0.5N$), so I experiment with N and n as a single parameter (thus disregarding n in this instance). It is worth noting that undisclosed filtering processes for spam and bot querying may have already been applied to the testing query logs used in this work, thus explaining this effect. Future work, particularly if applying these approaches to real-world industrial QAC tasks will need to study the effect of n as a query flood limit since extreme spiking queries and spam are increasingly prevalent, and thus must to be taken into account for robustness.

LNQ with one or more parameter settings is able to significantly outperform the hard baseline for all query logs for prefix lengths of 2-3 characters. The greatest LNQ QAC performance improvements are observed for Sogou, where any N/n parameter setting is able to exceed the hard baseline performance for all prefix lengths. In particular, for 2-character prefixes LNQ exceeds hard baseline performance by 4.9% and 3.9% for Sogou and MSN, respectively.

[1] With exception of the Sogou 3 character prefix, where the soft baseline marginally outperforms the best available hard baseline. In this case, I treat the higher-performing soft baseline as the hard baseline to compare proposed approaches against.

Optimal N and n parameter settings vary between query logs. Sogou is less sensitive to N, whereas MSN and AOL performance fluctuates more with different settings. This may in part be because Sogou has stronger temporal query distributions since it has the most queries concentrated in the shortest time, along with very specific prefixes. Interestingly, a clear pattern emerges whereby the optimal N and n parameters for LNQ are near constant for each query log, regardless of the query prefix length, indicating that parameters are dependent on the query log and its intrinsic temporal characteristics rather than prefix length.

Table 4.4: **LNQ** Approach - QAC MRR change (relative to the MLE-W Hard Baseline) observed for each N/n parameter, with query prefix ρ lengths of 2-3 characters for AOL, MSN and Sogou query logs. Best performing N/n are highlighted.

	Query Log / LNQ N (and n) Parameter				
ρ	100	200	400	800	1200
			AOL		
2	-8.26%	-3.28%	-0.53%	+0.46%	**+0.83%**
3	-4.04%	-0.95%	+0.44%	+0.78%	**+0.81%**
			MSN		
2	-2.98%	+1.24%	+3.27%	+3.82%	**+3.92%**
3	-1.78%	+0.81%	+1.92%	**+2.16%**	+2.05%
			Sogou		
2	+4.58%	**+4.88%**	+4.65%	+4.09%	+3.66%
3	+3.17%	**+3.21%**	+3.03%	+2.75%	+2.55%

4.5.4 Approach: PNQ

In Table 4.5, I present the MRR change achieved by PNQ QAC, relative to the MLE-W hard baseline, for prefixes of 2-3 characters. We experiment with PNQ linear regression models of $10{\times}50$, $5{\times}100$, $20{\times}50$, $10{\times}100$ and $5{\times}200$ (this $M{\times}N$ nomenclature is explained in Section 4.3.4).

With the appropriate regression model, PNQ is able to marginally outperform LNQ for all query logs and prefix lengths, except for MSN with 3-character prefixes. However, despite the additional overhead of making predictions, there is relatively little improvement over the optimal LNQ parameter.

In contrast with MSN and Sogou, PNQ for AOL is very sensitive to the parameters, although AOL on the whole does have much lower QAC effectiveness than both MSN and Sogou.

There is no optimal parameter for all query logs, although the same parameter is optimal for different prefix lengths in the same query log.

Table 4.5: **PNQ** Approach - QAC MRR change (relative to the MLE-W Hard Baseline) observed for each $M \times N$ regression model, with query prefix ρ lengths of 2-3 characters for AOL, MSN and Sogou query logs. Best performing regression models are highlighted.

Query Log / $M \times N$ Regression Model					
ρ	10×50	5×100	20×50	10×100	5×200
AOL					
2	-0.25%	-0.35%	+0.82%	**+0.85%**	+0.80%
3	+0.68%	+0.61%	**+1.12%**	**+1.12%**	+1.07%
MSN					
2	+3.29%	+3.30%	+3.86%	+3.95%	**+4.06%**
3	+1.60%	+1.70%	+1.96%	+2.02%	**+2.11%**
Sogou					
2	**+5.14%**	+5.13%	+5.11%	+5.11%	+5.04%
3	**+3.38%**	+3.36%	+3.35%	+3.34%	+3.29%

PNQ Predictive Model Accuracy

Although I focus on the final outcome (i.e., the QAC performance, measured using MRR) in this scenario, to illustrate the accuracy of the PNQ multiple linear regression model in predicting upcoming query counts. In Table 4.6, I report training set size and predictive model accuracy measures (i.e., RMS Error and R^2 co-efficient of determination) for 2 character prefixes in AOL, MSN and Sogou for five $M \times N$ regression models. The accuracy measures are computed over 2 days of the query log from day 2 onwards (i.e., after an initial training period). I filter long-tail queries with a single occurrence from training – so only queries with a frequency of two or more are used in the training set.

I focus further analysis for a 5×200 linear regression model. For all collections, there was almost no significant difference in predictive accuracy between 20×50, 10×100 and 5×200 regression models. In the first two days of AOL, MSN and Sogou, there were 160K, 287K and 170K training instances generated, respectively.

For each query log, the regression model co-efficients were as follows (the left-most co-efficient is for the query count in the most recent 200 queries in this model):

- AOL: $\{0.19, 0.1, 0.07, 0.07, 0.07\}$

Table 4.6: **PNQ** query popularity (in next 100 queries observed with prefix) $M \times N$ prediction model accuracy for 2 character prefixes for AOL, MSN and Sogou query logs. Best performing regression models are highlighted.

	Query Log / $M \times N$ Regression Model				
Measure	10×50	5×100	20×50	10×100	5×200
			AOL		
R^2	0.894	0.894	**0.898**	0.898	0.898
RMS Error	0.941	0.941	**0.671**	0.672	0.673
| Train Set |	93K	93K	160K	160K	160K
			MSN		
R^2	0.919	0.918	0.917	**0.916**	0.916
RMS Error	1.142	1.144	0.816	**0.817**	0.820
| Train Set |	154K	154K	287K	287K	287K
			Sogou		
R^2	0.979	0.980	0.984	**0.985**	0.985
RMS Error	2.060	2.007	1.482	**1.448**	1.451
| Train Set |	133K	133K	170K	170K	170K

- MSN: $\{0.26, 0.11, 0.05, 0.04, 0.03\}$
- Sogou: $\{0.4, 0.05, 0.03, 0.017, 0.01\}$

A varying temporal dynamic between query logs is apparent given the varying co-efficients between the models. Predicting query counts in the next 100 queries with a given prefix, these models achieved R^2 of 0.9, 0.92 and 0.98, and a root mean squared error (RMSE) of 0.67, 0.82 and 1.5 for AOL, MSN and Sogou, respectively.

4.5.5 Approach: O-MLE-W, O-LNQ & O-PNQ

In Table 4.7, I report MRR change, relative to the hard baseline, achieved by the online parameter learning meta-approaches: O-MLE-W, O-LNQ and O-PNQ, with a learning horizon $\Delta = 100, 300, 600$ queries. For each QAC approach (i.e., MLE-W, LNQ and PNQ), to rank completion suggestions for each prefix at q_t, I identify the highest performing approach parameter for each prefix based on past queries with the same prefix observed in the learning horizon Δ. Each QAC approach may use the parameters from the previously reported experiments (e.g., $N/n = 100, 200, 400, 800, 1200$ for LNQ).

The results obtained by the online parameter learning meta-approach represent the greatest QAC effectiveness achieved by any of the approaches I propose, even if only marginally for

Table 4.7: **O-** Online Parameter Learning Approaches - QAC MRR change (relative to the MLE-W Hard Baseline) observed for each approach and training horizon (Δ queries), with a query prefix ρ length of 2 characters for AOL, MSN and Sogou query logs. Overall best performing approach and Δ are highlighted.

	Query Log / Δ		
Approach	100	300	600
	AOL		
O-MLE-W	+1.45%	**+1.57%**	+1.54%
O-LNQ	-0.36%	+0.41%	+0.72%
O-PNQ	+0.54%	+0.63%	+0.70%
	MSN		
O-MLE-W	+3.18%	+3.45%	+3.40%
O-LNQ	+3.32%	+3.90%	**+4.11%**
O-PNQ	+3.80%	+3.90%	+3.94%
	Sogou		
O-MLE-W	+3.83%	+3.75%	+3.67%
O-LNQ	+5.42%	**+5.43%**	+5.42%
O-PNQ	+5.27%	+5.28%	+5.28%

MSN and Sogou. In particular, I find O-LNQ performs best for MSN and Sogou, while O-MLE-W is the best approach for AOL. I outline underlying sampling and spam issues of the AOL query log, which may be responsible for this inconsistent finding, within the context of RQ1.1 discussion. For MSN and Sogou there is typically little difference between O-LNQ and O-PNQ performance. Relying on a greater number of past queries in the training horizon (i.e., $\Delta = 300, 600$) yields greatest QAC performance; however, performance is often insensitive to Δ, especially for Sogou.

4.5.6 Time-based Approach Performance

To characterise QAC approach performance over time, in Figure 4.9, I present the 6-hourly MRR of four QAC approaches. Each approach uses the best performing parameters for Sogou with a 2-character prefix: MLE-ALL (soft baseline), MLE-W (2-day sliding window - hard baseline), LNQ ($N/n = 200$) and PNQ (10×50).

Visibly apparent is the QAC performance fluctuation for all approaches over time. The 'cold-start' period of relatively low MRR (i.e., when there is little past query popularity evidence to rank completion suggestions) is prominent in the first four 6-hour periods which exhibit comparatively low MRR.

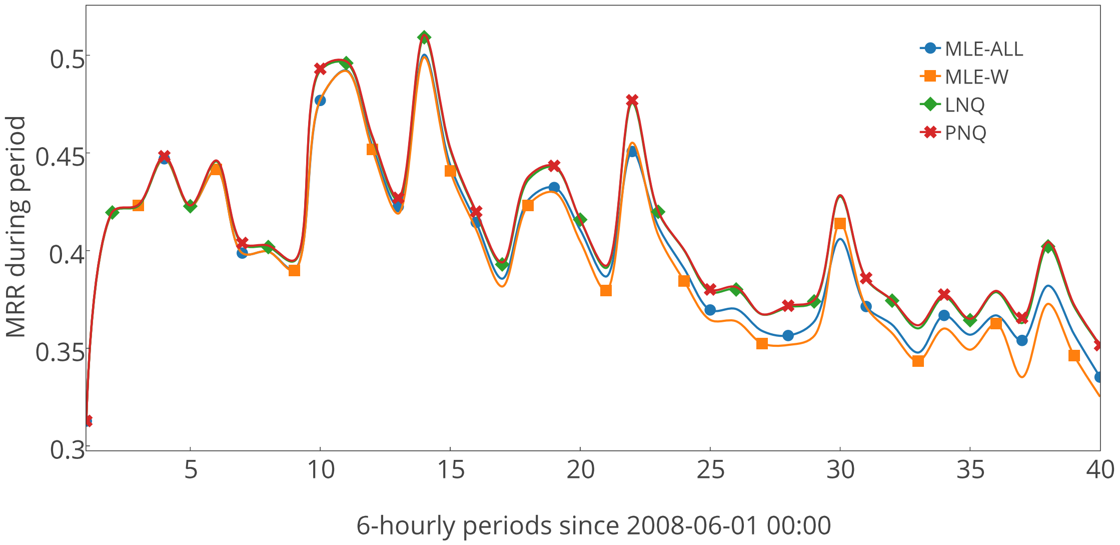


Figure 4.9: QAC performance every 6-hourly period in June 2008, for 4 QAC approaches with Sogou and a 2-character prefix.

There is some degree of daily day to night time cyclical pattern in MRR (i.e., the repeated spikes in MRR). With most people asleep at night, new and developing events and their associated previously unseen or unpopular queries will not be seen by a search engine in such great quantity, thus querying behaviour is likely less spontaneous, and thus MRR increased. Notably, there is no clear long-term pattern, suggesting a random degree of query popularity distribution over time with which QAC struggles to assist reliably.

Although the performance of all approaches is largely similar to begin with, beyond this (i.e., from twenty 6-hourly periods onwards), substantial variance between approaches is observed. MLE-W starts by under-performing compared to MLE-ALL. However, in the latter half of the month (i.e., sixty 6-hourly periods/15 days onwards), MLE-ALL gets progressively and consistently worse compared to MLE-W, most likely because it becomes increasingly insensitive to changing query distributions because of the accumulation of overbearing historic query popularity evidence. LNQ and PNQ remain optimal as they are adaptively slide to utilise the most recent query popularity evidence, therefore maintaining their effectiveness.

LNQ and PNQ mostly reflect one another in performance, with occasional minor deviations. Both approaches consistently outperform MLE-ALL and MLE-W approaches at almost all times in the latter half of the month, and increasingly so towards the end of the month.

4.5.7 Discussion

In the following section I discuss the results I presented in the previous section in relation to the four research sub-questions of this chapter (i.e., RQ1.1-4).

RQ1.1: How effective is QAC in diverse search scenarios

Results presented in Section 4.5.1 demonstrate that QAC effectiveness varies considerably in different scenarios, that is, for different IR systems (i.e., different query logs in this evaluation) and prefix lengths in my experiments. QAC performance for shorter, non-specific prefix lengths (i.e., 2-3 characters) is always relatively low, as there is a large space of possible completions and much greater uncertainty regarding the user's intended query (Bar-Yossef and Kraus, 2011).

Chinese users have far more effective QAC since the Simplified Chinese alphabet is much larger than the Latin alphabet, reflected by a much shorter average Chinese query length (7 characters, compared to 17 for the Latin alphabet).

Even for the Latin alphabet, QAC effectiveness is not equal between scenarios – at least for shorter prefixes, e.g., 2-4 characters, in which MSN achieves up to 16.8% greater baseline QAC performance than AOL. Indeed, AOL performs relatively poorly in all our experiments and sometimes shows considerably differing approach performance and optimality (e.g., for O- experiments where O-MLE-W is best for AOL, in contrast to O-LNQ for MSN and Sogou).

In part, this may be caused by a number of factors which may limit AOL-only findings. Firstly, although AOL contains more queries, these are spread sparsely over a three month period, whereas MSN queries are concentrated in a one month period. This may suggest that a search engine serving more queries is able to make better completion suggestions since it has a larger sample of changing behaviour. Additionally, AOL's day of missing data (Adar et al., 2007) may to some degree harm QAC effectiveness following the affected period. Furthermore, there may be underlying demographic differences between users of the two search engines that lead to changes in query distributions, for instance, MSN and Sogou users may be more responsive to ongoing events. The sampling used to construct AOL may also distort results. While MSN and Sogou were sampled from randomly selected queries over their duration, AOL was constructed by randomly selecting a subset of users, and including all their queries in the query log. This tracking of individual's activity, and perhaps their re-finding bias, is likely to skew query distributions. Finally, bot spamming activity in AOL may degrade QAC performance for real users by distorting query popularity distributions in the derived query log. Regardless of AOL's possible issues, I still consider AOL results indicative if the behaviour is echoed by MSN and Sogou results.

Effectiveness converges for 5-character prefixes, probably because QAC is simply constrained by how much past evidence is available, and a 5-character prefix hugely narrows possible completion suggestions in any case.

RQ1.2: Can relying only on recent query popularity evidence improve QAC effectiveness?

In Section 4.5.2, I found that QAC effectiveness can be improved by using a recent sliding window period of query log evidence to rank completion suggestions. Although MRR gains are relatively small (1.1-3.6% above the soft baseline), they are consistent for 2-character prefixes over all query logs.

Although performance improvements for each prefix length and sliding window period are different for each query log, there is a clear overall relationship between prefix length and optimal sliding window period emerging from the results. QAC for shorter prefixes performs optimally with a shorter sliding window of evidence, and conversely, QAC for longer prefixes performs best with a longer sliding window of evidence. However, for prefixes of 3 or more characters, the 14-28 day sliding window period upper limit I experiment with is not enough to begin observing notable improvements consistent with this trend (I cannot reliably increase the sliding window period for our short-term experimental query logs). We would expect experiments on longer-term query logs to begin seeing improvements based on this trend.

Even though relying on recent evidence (i.e., using a short sliding window) increases the time-sensitivity of QAC, it is apparent that it can also harm the robustness of QAC, reducing performance overall. Less common prefixes (e.g. for rarer queries) and longer prefixes (e.g. 4-5 characters) rely on a longer sliding window period to include the evidence necessary to rank them effectively, since they are less likely to have been used recently. In these cases, imposing a sliding window period is an inherently flawed approach as it excludes potentially valuable past query log evidence.

Regardless of language, time is a common characteristic that affects all QAC scenarios. As evidenced by Figure 4.9, QAC effectiveness is not stationary over time. That said, the impact of time is very different for each query log, suggesting that each query log scenario has differing temporal characteristics.

RQ1.3: Can an optimal trade-off between recency and robustness be achieved for QAC?

In Section 4.5.3, I presented the results of our proposed LNQ approach, which addresses the flaw of imposing a sliding window period of evidence for all prefixes. Rather than imposing a strict past evidence time period, LNQ tracks the past N queries observed with each prefix and ranks completion suggestions by their distribution. The N parameter allows the model to be adjusted for robustness (greater N) or recency (lower N), while accommodating the varying popularity of different prefixes. n ensures no single query can flood the model if it becomes extremely popular, or abused (e.g., spammed).

LNQ results observed for MSN and Sogou are especially encouraging, with up to +3.9% and +4.9% QAC performance increases over the hard baseline for many N parameters, demonstrating that an optimal trade-off between recency and robustness can be reached. Improvement is marginal for AOL, but exploring further N parameters may yield better results. What is in essence a very straight-forward and practical technique is capable of making considerable gains over the hard baseline.

Different prefixes benefit from different LNQ N parameters at different times. This is demonstrated by O-LNQ in Table 4.7, where online learning of the optimal LNQ N parameter per prefix for LNQ leads to the best performance I observe overall for 2-character prefixes in MSN and Sogou: +4.1% and +5.4% over the hard baseline, and correspondingly a +7.2% and +9.2% QAC effectiveness improvement over the soft baseline.

RQ1.4: Can short-range query popularity prediction (based on recent trends) improve QAC effectiveness?

Given the effectiveness of the stand-alone LNQ approach, I adapted it for short-range query popularity prediction based on recent trends. In Section 4.5.4, I presented results of the PNQ approach which uses an incrementally-learnt regression model to predict current query popularity for completion suggestion ranking.

In almost all cases, PNQ was only able to marginally improve the best performing LNQ approach for each query log (i.e., by a fraction of a percent). Despite the relatively high prediction accuracy of the query count regression model (especially for R^2, which is most pertinent given the rank-ordered nature of QAC), PNQ only offers small gains over LNQ. Future work will investigate more elaborate incrementally-learnt regression models to improve PNQ performance through increased temporal sensitivity.

4.6 Chapter Conclusions

In this section I conclude on the sub-questions investigated in this chapter. I considered the trade-off between recent (i.e., time-sensitive) and robust query auto-completion (QAC). The objective was to develop effective QAC approaches that can provide both consistently popular and recently popular query completion suggestions alongside one another, when necessary.

First, I outlined two existing common approaches to QAC, which are based on the most popular queries observed previously, and in recent periods of time (i.e., MLE-ALL and MLE-W). I proposed a new QAC approach based on tracking the last N queries with a given prefix (i.e., LNQ). I extended this approach to anticipate upcoming query popularity changes using query count prediction based on an incrementally-learnt short-range linear regression model (i.e., PNQ). Finally, I proposed an online parameter learning meta-approach to determine the optimal parameters of the aforementioned QAC approaches in a typical online QAC setting.

Using three real query logs, namely AOL, MSN (both English language) and Sogou (Chinese language), I simulate a realistic real-time QAC system and experiment with each QAC approach for query prefix lengths of 2 to 5 characters over tens of millions of queries. The diverse temporal, language and demographic characteristics of each query log allows QAC approach performance comparison in differing real-world scenarios. As all three query log datasets are publicly available, all results presented in this paper are reproducible. Through extensive empirical investigation I study the effectiveness of each approach with various parameter settings.

I find that using all past query log evidence to rank completion suggestions (i.e., MLE-ALL, my soft baseline) leads to under-performing QAC. However, using a sliding window period of past evidence (MLE-W, my hard baseline) can improve QAC performance up to 3.6% above the soft baseline.

Tracking the last N queries observed for a prefix to rank completion suggestions (i.e., LNQ) produces strong performance improvements in most cases, especially when the trade-off between robustness and recency, controlled by N, is optimal. Notably, QAC for MSN and Sogou is improved by +7% and +8.7%, respectively, for prefixes of 2 characters over the MLE-ALL soft baseline. LNQ is a highly practical and efficient approach which can be easily trained on- or off-line for effective recent and robust QAC performance. Notably, online learning to select the optimal LNQ N parameter for each prefix continually over time leads to the best QAC performance increase I observe over the soft baseline for 2 character prefixes in MSN and Sogou. In particular, O-LNQ for MSN achieves +7.2% (an absolute MRR of

0.125), and +9.2% (an absolute MRR of 0.45) for Sogou. Overall, this finding represents reproducible state-of-the-art QAC performance for these query logs. Employing a relatively high-accuracy recent trend linear regression model to rank completion suggestions based on short-range predicted query counts (i.e., PNQ) offers little improvement over an appropriately selected LNQ model.

Overall, building an effective QAC system requires careful selection of approach, parameters and query popularity predictive model to suit the intrinsic temporal, language and demographic characteristics of the scenario; I found no combination that is assured to be effective in all situations. I expect the insights and patterns observed for each approach in this work will inform later QAC system research and development.

Chapter 5

Temporal Relevance as Time-sensitive Search Result Diversification

In the previous chapter I proposed QAC approaches which *explicitly* help users to formulate descriptive and error-free queries consistently over time. However, as well as query popularity changing over time, so to does the meaning (or, *intent*) of many queries as recent and ongoing temporal events and phenomena influence information needs and expectations. In this chapter, I explore methods to *implicitly* support users over time with time-sensitive search result ranking. More specifically, I consider approaches to take into account the focus of temporal interests during relevance ranking, that is, *temporal relevance*. The lack of suitably long-term open query logs for characterising temporal dynamics of user behaviour inhibits this area of research; I first address this issue by examining the suitability of open Wikipedia datasets as a surrogate to proprietary commercial query logs. With a view to operationalising temporal relevance as time-sensitive intent-aware search result diversification, I begin by studying the impact of temporal dynamics on ambiguous query intents. Following this, I proceed to explore novel methods to discover and model multi-faceted query intents for event-driven information needs.*

5.1 Introduction

The primary objective of any information retrieval (IR) system is to satisfy the user's information need. However, there is often uncertainty about the information need caused by vague (i.e., under-specified) or otherwise ambiguous queries (Agrawal et al., 2009). In the case of web search, this problem is highlighted when queries are typically only two to three words in length (Jansen and Spink, 2006; Sanderson, 2008; Silverstein et al., 1999), and the collection

*Research presented in this chapter is published in Whiting et al. (2013b), Whiting et al. (2014) and Whiting et al. (2015).

covers a very broad space of topics which the user does not realise will impair the specificity of their provided query.

Relevance uncertainty is especially problematic for *ambiguous* and *multi-faceted* queries. In some cases the user may provide a query which is ambiguous because it has many possible interpretations, of which only one would be considered relevant. For example, the user's motivation behind the query "`jaguar`" may be to find information related to (i) the Jaguar vehicle brand, (ii) animal, (iii) Apple OS X operating system version, or (iv) Fender Jaguar guitar. Song et al. (2007) estimate that 16% of web search queries are ambiguous in nature, so supporting temporal dynamics of ambiguous queries will be potentially high impact. On the other hand, a query is considered multi-faceted if it describes a broad topic consisting of multiple facets (or synonymously, *sub-topics*), with each relating to a specific aspect of the broader query topic. The user may be most interested in information related to one or more of these possible facets. For example, the user's motivation behind the query "`steve jobs`" may be to find information about his (i) death, (ii) biography, (iii) movie, or (iv) keynote presentations, and so on. Jansen et al. (2007) suggest that 25% of web search queries are generally vague and therefore multi-faceted in nature, so much like ambiguous queries, supporting temporal dynamics in multi-faceted queries will also be potentially high impact. For the remainder of this chapter, I refer to the intended interpretation of an ambiguous query, or similarly, the desired facet(s) of a multi-faceted query as the user's *query intent*.

In a sense, the vast majority of search queries could be considered either ambiguous or multi-faceted given the inherent uncertainty of almost all but the most trivial of queries (e.g., exact navigational queries such as "`google.com`"). Without further clarification it is impossible to know the user's precise intent for either type of query. Of course, the user could be requested to elaborate their query to better state their intention. However, to minimise effort it is more desirable to first present results which cover a range of possible query intents, and therefore might satisfy their information need without further interaction – thus maximising the effectiveness of the IR system.

Considering that users often have differing query intents, *intent-aware search result diversification* is a commonly employed strategy to maximise the chance of satisfying users posing ambiguous or multi-faceted queries (Santos et al., 2011). The premise of intent-aware diversification is to interleave results covering as many query intents as possible to produce an *optimal* intent-aware ranking, which maximises the chance of satisfying the most users. Query intent ranking is based on several signals, in particular: (i) the relevance of each search result to the user's query, (ii) the relevance of each search result to each known intent, and (iii) the previous popularity of each intent. In essence, the search results that cover the most popular query intents, and are themselves most relevant, should be higher ranked in results.

Accordingly, an integral part of effective intent-aware diversification is: (i) characterising each possible intent t_i of query topic T, and (ii) quantifying the popularity of each intent. This popularity is characterised with respect to other query intents as the query intent likelihood $P(t_i|T)$ – that is, the probability that the query intent is that which the user will be interested in. Possible query intents can be modelled by observing query reformulations following inadequate queries (i.e., query log mining) (Santos et al., 2011), derived from structured knowledge bases (Hu et al., 2009), document clustering and topic modelling, or hybrid approaches combining multiple techniques (Nguyen and Kanhabua, 2014). The popularity of a query intent $P(t_i|T)$ is typically modelled using past preferences shown towards each query intent – measured through past query reformulation or search result click-through rates for items deemed relevant to the query intent (Santos et al., 2011).

The existence of query intents, and their accompanying popularity has conventionally been considered as *stationary* – that is, neither change over time. However, as I explored in Chapter 3, temporal dynamics have many effects on information seeking behaviour which in turn affects query and intent popularity (Radinsky et al., 2013b). Temporal events (e.g., TV shows, politics, sport and public holidays, etc.) and phenomena (e.g., seasonal weather and cultural change, etc.) are behind many information needs (Adar et al., 2007; Kulkarni et al., 2011), and therefore influence what users are most likely expecting to find in results. As such, to be consistently effective, a search engine must anticipate and accommodate such time-based changes when necessary.

Query intent has been observed changing due to both periodic (and therefore, potentially predictable) and irregular (i.e., event-driven) influences over time (Adar et al., 2007; Kulkarni et al., 2011; Radinsky et al., 2013b; Zhou et al., 2013). For intent-aware ranking to be consistently optimal over time, temporal dynamics in query intent popularity must be reflected by changes in intent-aware ranking.

By considering that new query intents emerge over time, and that the popularity of all query intents is subject to temporal dynamics, I challenge the prevailing belief of stationary query intents in pursuit of a framework for integrating a notion of *temporal relevance* in time-aware IR. Temporal relevance in this sense is defined as the variability of information item usefulness for satisfying a particular information need, given external temporal factors such as the current time or recent or upcoming events, which may dictate the motivation and therefore expectations of the user.

I first focus on ambiguous queries, and study how temporal dynamics affect the optimal intent-aware ranking over time. Next, I turn to multi-faceted queries, and in particular, those driven by temporal events such as unfolding news stories. As these queries occur in real-time, there is often little past evidence to reliably structure query intents and characterise their popularity – making effective intent-aware ranking challenging. I therefore consider whether temporal dynamics contained in the collection can be used to reliably model query intents and derive their popularity over time. With respect to this, I evaluate the effectiveness of characterising emerging query intents and their popularity through changes in the content (i.e., *content dynamics*) related to the information need.

5.1.1 Motivation

The search results preferred by users have been observed changing over time for many queries (Kulkarni et al., 2011; Radinsky et al., 2013b). Temporal relevance is founded on the idea that relevance of information to any given information need is itself is subject to temporal dynamics. More specifically, the same literal query topic (e.g., "`tax`") might have differing relevant items depending on *when* the query is posed. For example the user may be referring to tax return filing help just before the filing deadline. Likewise, they may be seeking the dates when they will receive any overpaid tax back following the deadline. While these common information needs may be predictable every year, a new announcement to the tax code may cause a brief one-off change to this routine.

Although topical relevance (Mizzaro, 1997) is undoubtedly foremost, temporal relevance can be viewed as the complementary pertinence of an item given recent or ongoing events and phenomena that are likely to have motivated the user's information seeking behaviour. The distinguishing factor of temporal relevance is that it affects *all* information items – not just timestamped items in a time-based collection, as is the case for typical approaches to temporal relevance, e.g., recency queries in news search (Dakka et al., 2012). Indeed, recent and ongoing events and phenomena, whether predictable or unpredictable, might result in old or non-timestamped items becoming highly relevant, albeit only temporarily.

Frameworks for intent-aware diversification are well established (Santos et al., 2011), and thus suitable for integrating this notion of temporal relevance alongside established notions of topical relevance. Rather than modelling temporal relevance of individual items based on past click through rates, which may be unreliable for measuring past relevance of items because of previous ranking biases (Radlinski et al., 2008), I instead propose to model the temporal dynamics of query intent popularity within an intent-aware diversity framework. Query intents offer a logical unit to reason past temporal popularity (if evidence is available), and to consider in relation to recent and ongoing events and phenomena which might influence temporal relevance. Accordingly, considering emerging query intents and temporal query intent popularity is an appropriate first step towards operationalising this notion of temporal relevance.

Modelling temporal relevance in an intent-aware framework is a complex problem consisting of many research challenges. A broad research agenda towards solving this problem involves the following challenges:

1. Quantifying the impact of query intent popularity temporal dynamics on existing intent-aware ranking approaches – thus, motivating the need to address the problem of temporal relevance. I study this as part of RQ2.1 and RQ2.2 for ambiguous queries, and RQ2.3 for multi-faceted queries in this chapter.

2. Discovering emerging or evolving query intents, and deriving their temporal popularity. There may be insufficient existing query log evidence available, or, it may not reflect current real-time query intents enough to reliably perform this task using conventional intent mining approaches. I investigate this problem as part of RQ2.4 for multi-faceted queries in this chapter.

3. Modelling past query intent popularity temporal dynamics to predict future intent popularity temporal dynamics.

4. Combining past and present query intent popularity to produce a consistently optimal intent-aware temporal relevance ranking.

5. Finally, time-sensitive intent-aware rankings will inevitably affect user behaviour, so evaluating user satisfaction over time becomes necessary. Established IR intent-aware system effectiveness evaluation metrics do not take temporal dynamics into account. Consequently, these metrics are limited when considering changes over time (Zhou et al., 2013), hence new approaches for time-aware evaluation need to be investigated.

In this thesis, I focus on point 1 and 2 of this research agenda. To that end, I quantify the impact of temporal dynamics on effectively satisfying ambiguous queries. Further, I study methods to model emerging and evolving query intents, and the temporal dynamics of their popularity for event-driven multi-faceted queries. Although the work presented in this chapter does not seek to completely solve the complex problem of addressing temporal relevance, it does contribute a thorough definition of the problem, motivate the impact of future work and highlight novel approaches (e.g., intent-aware temporal ranking) ripe for further exploration. Hence, this chapter establishes an significant foundation for future work.

5.1.2 Research Questions

In this chapter I investigate four specific research sub-questions, with regard to **RQ2** of this thesis: "Query intent exhibits many patterns and trends over time as collective influences and expectations change. *To what extent do query intents change over time, and, can these temporal dynamics be supported?*".

The first two research sub-questions relate to the temporal dynamics of *ambiguous* queries:

- **RQ2.1:** To what extent are ambiguous queries affected by occasional (i.e., event-driven) ranking changes?
- **RQ2.2:** To what extent are ambiguous queries affected by consistently periodic (e.g., hour-to-hour, day-to-day, etc.) ranking changes?

The following two research sub-questions relate to the temporal dynamics of event-driven *multi-faceted* queries:

- **RQ2.3:** Does content structure of related information reflect query intents for event-driven multi-faceted queries?
- **RQ2.4:** Do content dynamics of this structured information correlate with query intent popularity for event-driven multi-faceted queries?

5.1.3 Chapter Outline

This chapter is structured around answering the research sub-questions related to ambiguous and multi-faceted queries. Accordingly, it is organised as follows:

- Section 5.2 presents background work on intent-aware IR, and existing temporal relevance approaches.
- Section 5.3 explores Wikipedia as a surrogate source for the long-term temporal dynamics found in a large-scale query log, since such a query log is not openly available for research.
- Section 5.4 examines the impact of temporal dynamics on optimal intent-aware ranking for ambiguous queries, and presents findings.
- Section 5.5 examines modelling event-driven multi-faceted query intents and their popularity using content dynamics, and presents findings.

- Section 5.6 provides overall conclusions on the work and findings presented in this chapter.

5.2 Related Work

In this section I present work related to this chapter, in particular, intent-aware search result diversification and modelling temporal relevance.

A substantial body of recent IR research has investigated intent-aware (also referred to as "diversity-based", or "sub-topic") retrieval approaches for modelling and satisfying all likely query intents during search tasks, where ambiguity or multi-faceted information needs cause relevance uncertainty (Carbonell and Goldstein, 1998; Zhai et al., 2003). Previous work has identified temporal dynamics in query intents by measuring change, in terms of entropy, in clicked results for the same query over time (Kairam et al., 2013; Kulkarni et al., 2011). Several temporal clues, such as relevant document timestamps and recent query popularity with result clicks, have been used in task of news vertical selection to decide whether recent news results should be promoted for a general web search query (Arguello et al., 2009). However, the impact of temporal dynamics on intent-aware diversity approaches has been largely ignored, instead assuming intents and their popularity remain static over time.

Past work has explored temporal dynamics in query intent from different perspectives. Based on temporal query log analysis, Kulkarni et al. (2011) suggest changes in query intent for common queries, leaving analysis to later work. Radinsky et al. (2013a,b) model temporal periodicities, surprises, trends and random noise in the preferences shown to individual search results over time to predict future preferences. The work presented in this chapter differs from this, as rather than modelling and predicting popularity of individual search results in isolation, I instead look to analyse the impact of time on intent-aware ranking. Moreover, I concentrate on using query intents to robustly model temporal interests, rather than the temporal relevance of individual items. Kairam et al. (2013) highlights the need for intent-aware diversity for event-driven information needs.

Sun et al. (2010) offer a system to visualise how the result rankings for a query change over time to test their proposed measure of result ranking similarity, however do not quantify the overall extent of temporal ranking changes. I discuss the measure they propose and its limitations in this scenario in the context of measuring temporal ranking changes in Section 5.4.1. Often the same queries re-occur annually, e.g., "`sigir`". Metzler et al. (2009) uses the time of the query to disambiguate the most likely year reference expected to be found in the relevant results, such as "sigir 2015" if the query is submitted after the burst of interest in SIGIR 2014. This technique is limited to annually re-occurring queries, rather than more general temporally changing queries (e.g., hourly and daily, etc.). Nguyen and Kanhabua (2014) mine dynamic query intents using recent query logs and content changes, and find their time-aware approach beats several state-of-the-art non-temporal intent mining approaches.

In our early work (Zhou et al., 2013), we conducted an analysis of how ambiguous query intent popularity changes monthly over a year. Based on the changes in individual query intent likelihood, we found around 35% of ambiguous queries have modest or more temporal intent change. Using these statistics, we simulated the impact of temporal dynamics on existing system-oriented diversity evaluation approaches and found they were ineffective. In this thesis I continue the analysis by further examining the temporal impact on ambiguous queries. Accordingly, I address the fact that changes in query intent popularity do not always affect search rankings. Furthermore, changes for less popular query intent are less important. Additionally, I analyse temporal dynamics over varying periods of time, e.g., hourly, daily and monthly to provide a much more in-depth understanding.

5.3 Wikipedia as a Source of Temporal Dynamics

To address the research questions set forth in this chapter, it is first necessary to observe how information, and related information behaviour change over time. The first challenge for this work, and indeed any future open research in this area, is in fact a *data problem*.

A real search engine query log contains a wealth of history about what users search for, and the results they prefer over time by way of result clicks. Analysing this previously collected evidence allows us to structure query intents and characterise temporal dynamics. Unfortunately a suitably large-scale and long-term[1] query log is not available to researchers outside of commercial search engines. In any case, for the most recent and ongoing events, a query log alone may not have had the chance to capture much evidence of the latest ongoing changes in information seeking behaviour. An alternative means of observing and studying information behaviour temporal dynamics is therefore needed. Tackling this issue, in this section I explore using openly available data from the Wikipedia collaboratively edited encyclopaedia as a surrogate query log.

As all the experiments conducted in this chapter rely extensively on Wikipedia to structure query intents and characterise their temporal dynamics, this section serves to motivate and justify its application in time-aware IR.

[1] In the context of this work, I consider a year or more to be appropriately long-term since it would contain both predictable periodic change (e.g., hourly, daily, monthly and seasonal) as well as many significant events and phenomena.

5.3.1 Wikipedia Characteristics

Wikipedia is the 7th most visited website on the internet[1]. Since its creation in early 2001, the founding English language encyclopaedia project has grown to include well over 4.7M structured articles covering a wide range of topics, with an average of 20.35 revisions per article[2]. For most countries, the local language Wikipedia domain is among the most visited websites.

Wikipedia facilitates a large-scale crowdsourced effort to continuously create and edit articles containing structured knowledge. Each article contains information and multimedia related to a particular topic, such as people, places or events. In many cases, authoring activity is triggered by users reporting ongoing real-world events, often very soon after they occurred, leading to an ever-evolving large-scale source of temporal information (Georgescu et al., 2013a,b; Steiner et al., 2013). Structured knowledge from Wikipedia has been used extensively in general IR research. In the following sections I discuss characteristics that make Wikipedia suitable for time-aware IR research.

Freshness and Timeliness

The latency of events being reflected in Wikipedia temporal dynamics is important for time-aware IR research. An ever increasing amount of editing activity is triggered by users reporting ongoing real-world events, often very soon after they occurred (Keegan et al., 2013; Vis, 2009). For mainstream events, Osborne et al. (2012) state that Wikipedia lags behind Twitter by about two hours on average, based on hourly page view statistics. However, Steiner et al. (2013) estimate time lag using the real-time article edit stream to be within 30 minutes, with major global news usually reflected in minutes (although initial edits are typically small and incremental). Worst-case scenarios are less studied. While anecdotal examples do not reflect overall timeliness, they do give insight into the temporal dynamics across all news sources.

For example, Whitney Houston's death was first reported on Twitter by the niece of the hotel worker who found her at 00:15 UTC on the 12th February 2012. After spreading through Twitter, at 00:57 that the Associated Press verified and broke the news on their Twitter feed. The first edit to Whitney Houston's Wikipedia page to reference her death ("*has died*") was at 01:01 (UTC). High-frequency editing of unfolding details followed, citing available sources. Steiner et al. (2013) uses the example of the resignation of Pope Benedict XVI, noting that the English and French Wikipedia articles were first edited at 10:58 and 11:00, respectively,

[1] According to Alexa Statistics in February, 2015 (`http://www.alexa.com/topsites`)

[2] According to official Wikipedia Statistics in June, 2013 (`https://en.wikipedia.org/wiki/Special:Statistics`)

which is impressive given that Reuters broke the news on their Twitter feed at 10:59, following the Vatican's public announcement at 10:57:47.

Topic Coverage

The Wikipedia project is available in 285[1] languages, each containing language/location-specific and translated articles. Although the large quantity of articles suggests extensive topic coverage, a study by Halavais and Lackaff (2008) quantifies topic coverage by measuring the similarity of the topical distribution of articles on Wikipedia to the topical distribution of books in print, based on established library classification systems. Wide disparity was observed, with subjects such as science and geography better represented in Wikipedia than in library books, and conversely, subjects such as medicine and literature much less represented in Wikipedia. Further work is needed to extend these findings to understand the nature of how events related to different topics (e.g., celebrity, politics, news, etc.) are reflected over time in Wikipedia.

Event Coverage

Major predictable and unpredictable events typically have their own dedicated articles (e.g. '*39th G8 Summit*', and '*2013 North India Floods*'), with the most important events covered by multiple articles discussing different aspects (e.g., timeline, comparison to similar events, or reactions). Less prominent events (including those that occurred before Wikipedia began) may appear as a sentence or section in a related article, or be mentioned in the Current Events portal, with a brief summary referencing the date of the event and links associating entities.

Many mainstream recent international events (including major sports events) appear in the Current Events portal[2], categorised by date and topic (e.g., ongoing events, deaths, conflicts, elections and trials). Archived versions provide a vast almanac of daily events since January 1900 (although earlier dates are more sparse and less structured). A separate Wikipedia for news (i.e., WikiNews[3]) is available, however is much less updated compared to the main Wikipedia projects – so I do not investigate it in this work.

Content Quality/Correctness

Many policies and guidelines govern Wikipedia editing in an effort to maintain encyclopaedic consistency[4]. In summary, article content should be written from a neutral point of view,

[1] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`, accessed: June 2013

[2] `http://en.wikipedia.org/wiki/Portal:Current_events`

[3] `http://www.wikinews.org`

[4] `http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines`

based on reliable sources and kept objective. A side-effect of Wikipedia's editing openness is that it sometimes leads to inaccurate reporting, deliberate vandalism or more subtle abuse (Potthast et al., 2008). The community reviewing mechanism often corrects obvious issues relatively quickly, with the aid of bots that watch recent changes and apply automated machine learning and heuristics to immediately flag issues (and occasionally instantly revert article revisions). High profile articles (e.g. celebrities, well-known politicians and currently prominent events) are often locked so that only administrators or established editors may change their content, reducing the volume of article edits, but ensuring accuracy. In the spirit of the fundamental policies and guidelines, article editing with current news must be backed by references, hence Wikipedia is not intended to be a primary source for news[1]. The 'Talk' page accompanying every article often contains commentary related to recent or necessary changes in the article, especially if there is concern about the content. Indeed, temporal discourse can be detected through the presence of discussion. An extensive discussion of the impact of Wikipedia editing policies on news reporting can be found in Keegan et al. (2013).

Comparison with other Event Information Sources

Twitter has become popular for monitoring real-time events because of its immediacy and volume of citizen reporting with '*tweets*' about ongoing events. Compared to Wikipedia it poses challenges, including: the scale and volume of tweets, limited document size, slang vocabulary, misspellings and spam. In fact, Twitter provides a 'soapbox' platform where user-generated content is often '*hearsay*', or non-neutral.

Twitter is undoubtedly an excellent source for detecting breaking new stories, especially instantaneous events such as spreading earthquakes (Sakaki et al., 2010). However for understanding of events, the quantity and quality of conflicting and redundant information created by users discussing and speculating event-based topics makes it difficult to monitor and organise key event detail and structure. In contrast, news wires offer a stream of professionally curated, and in most cases, high quality news stories. Consequently, news agency articles may be shortly delayed due to the fact they need to be written and edited prior to publication.

In the majority of cases, Wikipedia does not reflect events as quickly as Twitter – except perhaps for the most high profile topics (Osborne et al., 2012). However, Wikipedia trades a lag in reporting time for the sake of reporting accuracy. Likewise, in comparison to news wires, although Wikipedia does not have rigorous pre-publication editorial validation like well-respected publications, content can evolve quickly through editing. Users exposed to

[1] http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_newspaper

different media outlets aggregate and distill multiple sources of event detail into a single Wikipedia article through citations. Furthermore, Wikipedia offers unique structural characteristics for understanding many events. As it forms a linked knowledge base structure of articles, event details become encyclopaedic and hierarchically structured, as all related topics become associated.

5.3.2 Wikipedia Temporal Dynamics

Aside from content itself, the impact of past and present events and phenomena may be understood through several Wikipedia temporal dynamics. For instance, changes made in increased article editing, or raised article viewing popularity may be suggestive of related events and phenomena. Combining these temporal dynamics can provide several informative clues of temporal details, related entities, sequencing and impact.

Past and present temporal dynamics are readily available for many aspects of Wikipedia. In this section I discuss Wikipedia's data availability, and outline the page edit and page view streams which are used in the experiments for this chapter. I refer to temporal dynamics derived for the "Arab Spring"[1] event and accompanying Wikipedia article to anecdotally illustrate examples.

Data Availability

All Wikipedia article text, structure and meta data is openly available for research. A subset of data is available through real-time feeds, although completely historic archives are available to download. Wikimedia dump mirrors host many raw XML, SQL and CSV-formatted datasets[2]. Real-time monitoring of Wikipedia article creation and edit activity is possible through an Internet Relay Chat (IRC) channel and syndication feeds. Hourly page view statistics for all Wikipedia projects are available from December, 2007.

Article Revision Stream

Every Wikipedia page (including articles, talk and meta pages) has a complete full text revision history available, including reverted revisions. Several content-based temporal dynamics can be derived from the raw article revision history. A simple Unix `diff` operation between the markup of two revisions reveals edited content. In combination with the article markup, changing rich features such as article structure and links can be identified. The temporal editing activity and length for the 'Arab Spring' article is shown in Figure 5.1. The gradual

[1] http://en.wikipedia.org/wiki/Arab_Spring

[2] http://dumps.wikimedia.org

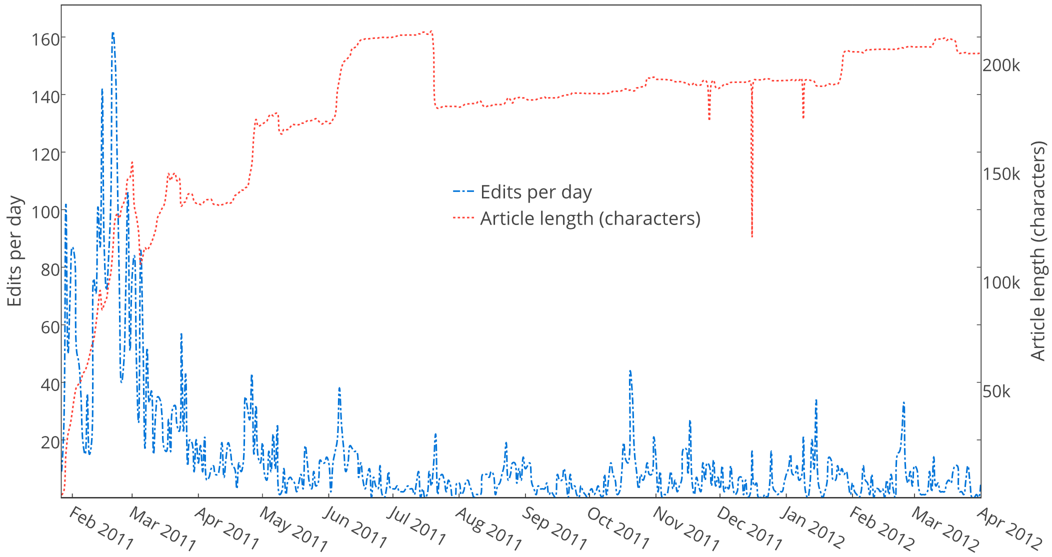


Figure 5.1: 'Arab Spring' daily article edit frequency and length (in characters) from 27th January 2011 to 23rd March 2012.

increase in article length as new events are added is prominent. The rapid decreases in article length (e.g., around August 2011) is caused by article restructuring, where bulky content related to a sub-topic is removed and placed in its own Wikipedia article. A `patch` operation resolves changes to character locations, allowing resolution to the hierarchical section in which they occurred (note that sections themselves will also evolve over time). Section change activity is illustrated for the 'Arab Spring' article in Figure 5.2. In Section 5.5, I exploit the section change stream for modelling temporal query intents of event-driven multifaceted queries.

Article View Stream

Article views are a direct visitor-driven measure of a topic's temporal popularity in Wikipedia. In Figure 5.3, I present the edits and page views per day for the 'Arab Spring' article. Interestingly, page views are not always strongly correlated with increased editing activity.

The latest data is available with a one hour lag, however there is occasionally corruption or empty data, for which smoothing or extrapolation can be helpful. Individual article views can be visualised and downloaded in JSON format using a 3rd-party tool[1]. In Section 5.4, I exploit the article view stream for modelling the temporal popularity of query intents for ambiguous queries.

[1] `http://stats.grok.se`

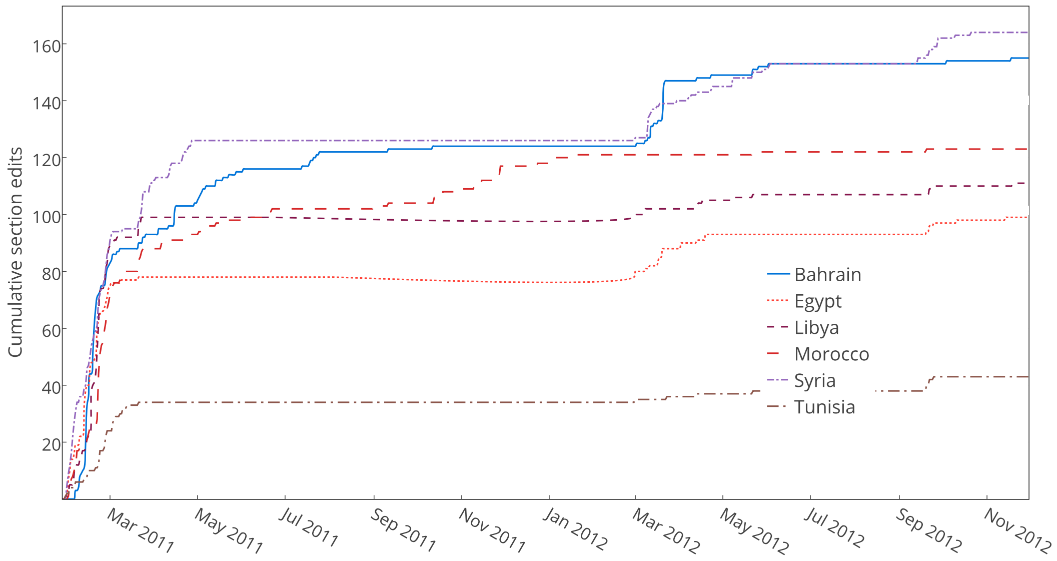


Figure 5.2: Cumulative 'Arab Spring' article section edit frequency, from 27th January 2011 to 26th November 2012.

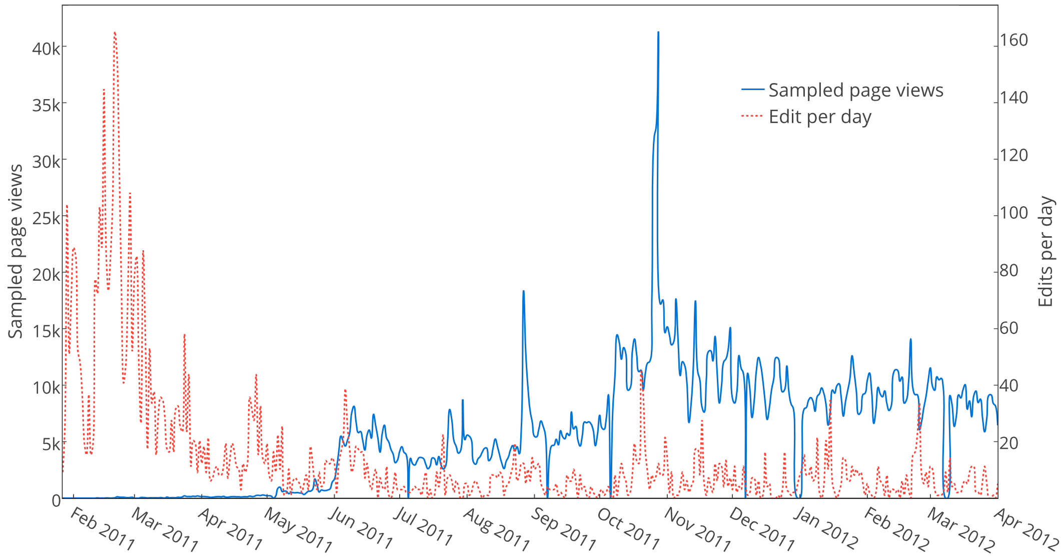


Figure 5.3: 'Arab Spring' article daily edit frequency and article views, from 27th January 2011 to 23rd March 2012.

5.3.3 Using Wikipedia as a Surrogate Query Log

In the previous sections I have discussed and illustrated how Wikipedia captures large-scale information behaviour temporal dynamics related to events and phenomena. I now present an argument that Wikipedia data offers insight into the same temporal dynamics as would be found in a long-term query log, based on the findings of several relevant works.

The first question is of course whether Wikipedia activity is closely tied to search engine activity. Safran (2013) found a relevant Wikipedia article ranked in the top 10 Google results (i.e., on the first page of search results) for around 80% of informational and 73% of transactional single word queries. This proportion reduces to a consistent 50-60% for two or more word queries. Accordingly, a Wikipedia article is among the top ranked results for the majority of search queries – and therefore, highly likely to be selected by the user. This is reflected directly in the referrer data reported for Wikipedia. In January 2015, 42% of Wikipedia's article views were directly referred from one of the three major search engines[1], i.e., Google, Bing or Yahoo. Hence, this finding shows temporal trends are driven by both search engine referral and in-site navigation behaviours.

Additionally supporting the argument that Wikipedia activity is closely tied to search engine activity, Druk (2014) studied the temporal correlation between Wikipedia article page views and related Google search query volume, as measured through Google Trends. Out of a random sample of 10,000 articles, daily temporal dynamics of query popularity were available for 8,000 articles on Google Trends (i.e., 80% of article topics are searched for relatively regularly). The average correlation co-efficient between the query popularity and article views was found to be 0.45. Furthermore, for 577 articles known to have periodic popularity (e.g., they relate to a known annual event), the correlation co-efficient rises to 0.63 – demonstrating an even stronger reflection between Wikipedia and search engines for periodic topics.

Overall, with such a large number of users turning to search engines to satisfy an information need, and subsequently selecting Wikipedia search results, many temporal dynamics on Wikipedia reflect the same information behaviour captured in a search query log over time. Thus, I posit Wikipedia temporal dynamics can be viewed as an indirect sample of search temporal dynamics. With that in mind, I argue that Wikipedia provides a reliable and timely surrogate source of the temporal dynamics found in a query log, and so use its open data extensively in the following sections to answer the research questions of this chapter. Using Wikipedia as a surrogate to the temporal information contained in a query log is clearly a limitation of some findings presented in this chapter, since there may be cases when the temporal dynamics contained in either source are influenced by factors that are not fully matched

[1] https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream_top_referrers

between them. Nevertheless, for the purpose of the research questions addressed in this chapter, I consider Wikipedia data reliable enough to make indicative temporal analysis and thus, sound findings. Future work will seek to further validate the findings of this chapter using proprietary long-term query logs, which are unavailable to researchers outside industry.

5.4 Temporal Dynamics of Ambiguous Queries

In this section I focus on quantifying the extent to which temporal dynamics affect optimal intent-aware ranking for ambiguous queries. Recall, a query is *ambiguous* if it has multiple interpretations, yet only one which would be considered relevant. For example, the user's intent behind the query "`powerplay`" may be to find information related to (i) the Power Play lottery option, (ii) IBM Cognos PowerPlay product, (iii) power play sporting term, (iv) powerplay fielding restrictions in cricket, or (v) the Power Play Star Trek episode.

Temporal dynamics often play a role in the user's query intent. Figure 5.4 shows the temporal popularity of the five most common query intents for the ambiguous query "`powerplay`". Each query intent likelihood observation corresponds to the likelihood of that query intent during a Western working week (i.e., Monday to Friday) or adjacent weekend (i.e., Saturday and Sunday). Most notably, there is a great deal of fluctuation and 'trading' of popularity over these periods of time, with several prominent temporal patterns and trends observable. The "Power Play" lottery option query intent and "IBM Cognos PowerPlay" product query intent are almost always most popular. However, there are clear periodic factors affecting their likelihood. The lottery option query intent is most likely at weekends, and conversely, the IBM product query intent during the week. However, several event-driven and so unrepeated effects are also apparent. A lottery roll-over causes the lottery intent to remain consistently most likely during week days/weekends shown between 35-45. Furthermore, the power play sporting term interpretation has a brief burst in popularity, probably because of controversy in a major sporting event bringing the term to the forefront. Overall, it is apparent the optimal query intent ranking for the ambiguous query "`powerplay`" is considerably affected by temporal dynamics.

Despite the clear impact of time, no previous work has analysed in-depth the impact of temporal dynamics on optimal intent-aware result ranking for ambiguous queries[1], and furthermore, to what extent over relatively short and long time frames (e.g., hour-to-hour and month-to-month). This insight will contribute an understanding of the level and nature of time-sensitive diversification necessary to consistently maintain optimal intent-aware ranking over time.

[1] Although I focus on ambiguous queries in this section, the proposed approach and methodology is equally applicable to multi-faceted queries.

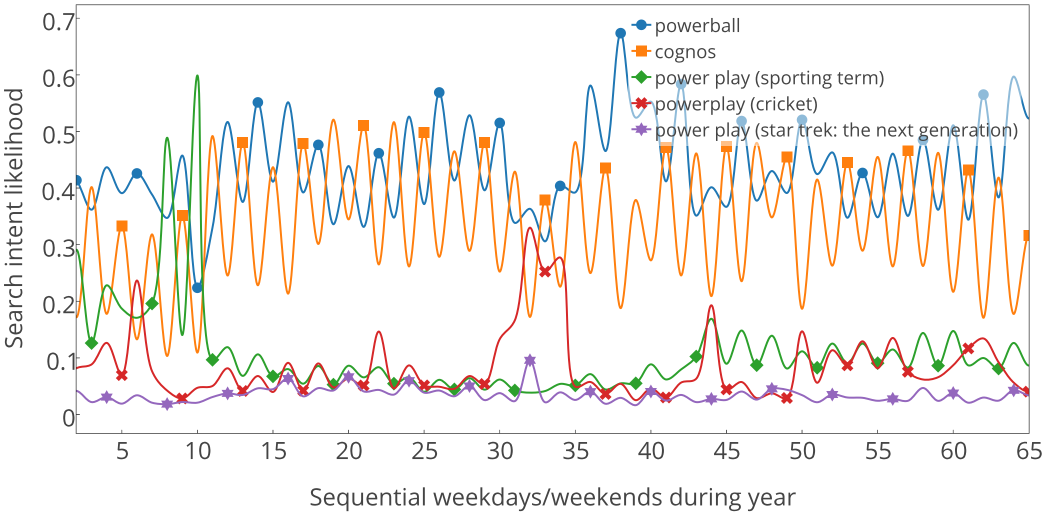


Figure 5.4: Query intent popularity over 67 sequential weekdays (i.e., Mon-Fri) and subsequent weekends (i.e., Sat-Sun), for five common query intents of the ambiguous query topic "`powerplay`" from May 21st, 2014.

Furthermore, it motivates the impact of modelling and successfully anticipating current periodic and event-driven ambiguous query intents on user satisfaction.

5.4.1 Analysing Temporal Ranking Changes

To analyse the effect of temporal dynamics on optimal intent-aware search ranking, I must study the past changes in optimal intent-aware search ranking over time. In this section I propose Time Slice Ranking Analysis (TSRA) as a general framework to: (i) identify search ranking changes, and (ii) quantify the severity of these changes over time.

Comparing search rankings between algorithms and search engines has received much attention in recent years. Established rank correlation measures such as Spearman's ρ and Kendall's τ have several shortcomings for this purpose. Predominantly, this includes their inability to deal with ranking ties, non-concordant search rankings (i.e., an item in one list but not the other) and ordinal weighting (i.e., swaps at lower ranks are less important). Several solutions to these issues have been proposed (Webber et al., 2010). Each of these IR-specific result rank correlation measures is based on various intuitions and requirements for handling ties, non-concordant items and ordinal weighting which are common in result ranking. The demands of the intent ranking change scenario I am investigating are more straight-forward, since although ranking ties can occur, the same set of query intents are present in both rankings being compared. However, the set of query intents can vary in size between ambiguous topics (i.e., two or more query intents), so, using ordinal weighting would mean the extent of temporal ranking change between different ambiguous query topics could not be compared.

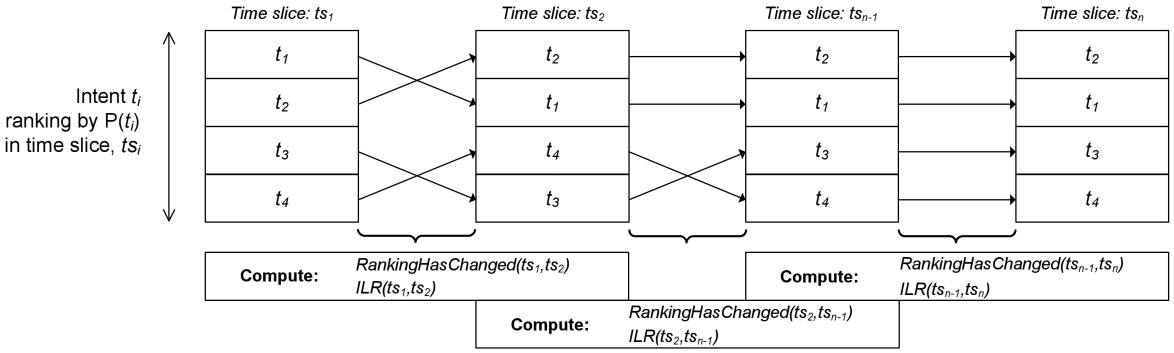

Time slice: ts_1
Time slice: ts_2
Time slice: ts_{n-1}
Time slice: ts_n
Intent t_i ranking by $P(t_i)$ in time slice, ts_i
t_1
t_2
t_3
t_4
t_2
t_1
t_4
t_3
t_2
t_1
t_3
t_4
t_2
t_1
t_3
t_4
Compute: RankingHasChanged(ts_1,ts_2) ILR(ts_1,ts_2)
Compute: RankingHasChanged(ts_2,ts_{n-1}) ILR(ts_2,ts_{n-1})
Compute: RankingHasChanged(ts_{n-1},ts_n) ILR(ts_{n-1},ts_n)

Figure 5.5: Conceptual illustration of Time Slice Ranking Analysis (TSRA) for all query intents t_i of ambiguous query topic T_i.

Consequently, as I am comparing so many rankings of between 2 and 20 query intents across over 90 thousand topics repeatedly over time, I propose TSRA as a straight-forward technique to ensure intuitive and interpretable analysis necessary to answer RQ2.1 and RQ2.2 in this chapter.

TSRA models the result ranking changes for a given query topic T over a sequence of adjacent past time frames, named *time slices*. TSRA is illustrated conceptually in Figure 5.5.

Each time slice, ts_i, represents the result ranking during a specific time frame, for instance an hour, part of the day (e.g., morning/afternoon/evening/night), day, month, and so on. For this work, each time slice comprises the set of query intents t_i, ranked by their popularity $P(t_i)$ observed during that time frame (e.g., clicks on results covering the query intent).

The result rankings contained in two adjacent time slices, i.e., a 'time slice pair', are compared to analyse ranking changes. I detail this process in the following section. There are $ts_n - 1$ time slice pairs available to compare. Hence, for 12 monthly time slices covering 1 year (as in our experiments), there are 11 time slice pairs to compute comparisons between.

The search rankings for each query topic T are analysed with TSRA based on several time slice temporal resolutions in our experiments, allowing analysis of ranking changes over both relatively fine-grained and coarse time frames. Only analysing ranking changes over a month can smooth over lesser time frame changes (e.g., day-to-day), so this approach facilitates greater insight into all temporal dynamics.

Identifying Significant Ranking Changes

In TSRA, the function $\mathrm{RankingHasChanged}(...)$ identifies a change in the result rankings between two adjacent time slices, taking ranking ties into account. However, a temporal ranking change can only be considered *significant* if it is accompanied by a relatively severe shift in query intent popularity – that is, a substantial change in the users' temporal focus. Greater query intent popularity changes will have most impact on user satisfaction.

To quantify ranking change severity – which I use with a minimum threshold to filter only significant ranking changes in experiments – I propose the 'Intent Likelihood Reallocation' (ILR) measure, based on the L1 distance metric. ILR computes the reallocation (or, *trading*) of popularity among query intents, based on the past preferences shown towards each intent, between a pair of time slices. This reallocation is responsible for any intent ranking change that occurs between time slices. ILR is defined as follows, where $P(t_i, ts_i)$ is defined as the likelihood of intent t_i being that which the user is interested in, in time slice ts_i:

$$ILR(ts_i, ts_{i+1}) = \Big(\sum_{i=1}^{|ts_i|} |P(t_i, ts_{i+1}) - P(t_i, ts_i)|\Big) \times 0.5 \tag{5.1}$$

Since ILR is based on the reallocation of likelihood, it can be at most 1. For a topic with two query intents, t_1 and t_2, an ILR of 1 indicates a 0% and 100% likelihood of t_1 and t_2, respectively, in the first time slice that changed to 100% and 0% likelihood of t_1 and t_2, respectively, in the second time slice (or vice-versa). As such, in combination with identifying a ranking change, ILR quantifies the relative severity of intent popularity change behind the ranking change, and hence is useful for discerning only the most impactful ranking changes over time.

5.4.2 Experimental Setup

I conduct several experiments with ambiguous queries derived from past user interaction to answer the RQs set forth in this paper. As I explored in Section 5.3, I use Wikipedia English as source of ambiguous queries, and a surrogate to the temporal dynamics of query intent popularity that would likely be observed in a real long-term query log.

Ambiguous Query Dataset

Over 255K topic disambiguation pages are present on Wikipedia to assist users in selecting their expected article – I consider each representative of a possible ambiguous query topic T. I treat each linked article from disambiguation page as a possible query intent t_i of T. In some cases, articles on disambiguation pages might mismatch what would be expected for information need disambiguation in a search engine. For instance, the disambiguation page “St. Mary’s Church” (a very common Christian church name) includes several links to place names (e.g., cities) which have a church of that name. These are in the minority, and typically disambiguation pages link directly to the disambiguated articles, and so this issue is unlikely to considerably impact the findings of this work.

The popularity of each query intent over different time frames (during 2014) is computed by aggregating hourly article view counts (as described in Section 5.3). Future work will study multiple years to reliably characterise broader temporal change, e.g., seasonal and year-to-year. Furthermore, additional work will explore how rankings change temporally on desktop versus mobile devices.

Wikipedia article viewing follows a characteristic power-law distribution. Accordingly, many articles are seldom visited. These articles offer little meaningful insight for our analysis since any changes are likely to be insignificant. To remove this noise, I discard all query intents represented by articles with less than 10,000 total views over 2014. Following this filtering, ambiguous queries left with one or no query intents remaining are discarded. The final dataset used in our experiments consists of 90,161 ambiguous query topics, with an average of 5.8 (± 6.2) query intents each[1].

5.4.3 Experimental Procedure

For each ambiguous query topic, I apply TSRA to analyse intent ranking changes over four time frames: (i) hour-to-hour, (ii) part-of-day to part-of-day (i.e., morning / afternoon / evening / night), (iii) weekdays-to-weekends, and (iv) month-to-month (during 2014). For each ambiguous query topic I count the number of ranking changes using TSRA over each time frame, subject to the following additional data filtering. If a query intent has very low popularity (i.e., < 10 article views) in either time slice of the time slice pair, I exclude it from the ranking comparison (note, this excludes a lot of long-tail query intents especially for intra-day analysis). Further, I only consider significant ranking changes, that is, with a relatively severe ILR > 0.1 (i.e., 10% of overall query intent popularity has been traded).

Some of the observed query intent ranking changes may be the result of random intent popularity changes over time – rather than the effect of non-random and significant temporal dynamics. Hence, in addition to analysing the temporal ranking changes produced by the respective intent popularity observed through Wikipedia article page view activity, I also analyse randomly generated intent popularity ranking for comparison. This random comparison

[1] For comparison, the NTCIR-10 Intent Mining and TREC Web diversity tasks use on average 7.7 and 3.7 intents per query topic, respectively (Sakai et al., 2013). Hence, the average number of intents per topic I analyse reflects previous intent mining and retrieval evaluation methodologies.

is generated by taking Wikipedia derived ambiguous query and intent dataset, but simulating a random popularity for each query intent in each time frame, and applying TSRA in the same fashion as for the Wikipedia observed intent popularities. Divergence between the observed and randomly generated intent ranking changes is indicative of non-random activity over time.

5.4.4 Results

In this section I present overall results of temporal ranking changes for 90K ambiguous queries, computed using TSRA over four time frame resolutions: (i) hour-to-hour, (ii) part-of-day to part-of-day (i.e., morning/afternoon/evening/night), (iii) weekdays-to-weekends (and vice versa), and (iv) month-to-month during 2014.

For all four time frame resolutions, I present the frequency distribution (which presents as a power law) of ambiguous query topics with n ranking changes (as defined by the constraints in the previous section) over that time frame. Logarithmic ($\log_{10}$) scaling is used on the result graph y-axis, i.e., the number of topics with n ranking changes. Each histogram distribution characterises the overall number of ambiguous topics with no or few temporal ranking changes through to constant ranking changes (i.e., persistently periodic over the time frame).

Random Intent Ranking Comparison

The first question that must be addressed is whether the temporal intent ranking changes I observe in this study are randomly occurring over time, or indeed non-random temporal dynamics. To this end, in the TSRA intent change results presented in Figures 5.7, 5.7, 5.8 and 5.9, I plot randomly generated temporal intent popularity ranking changes for comparison with the observed temporal intent popularity ranking changes.

Noteworthy in results presented for all time frames is that the number of randomly generated intent ranking changes follow a distinctly different distribution to the observed intent ranking changes, especially for larger time frames (e.g., month-to-month, weekdays-to-weekends and day-to-day). For shorter time frames (e.g., part of day-to-part of day), more random ranking changes are suggested by the closer distribution of observed and random intent ranking changes – however, there is still a considerable difference between the random and observed ranking changes. Of course, while a proportion of the intent ranking changes I observe may be the result of random activity, the divergence of observed and random ranking change suggests a the majority of the temporal dynamics I observe are not a random effect, and thus the findings of this study are indicative of non-random temporal information behaviour.

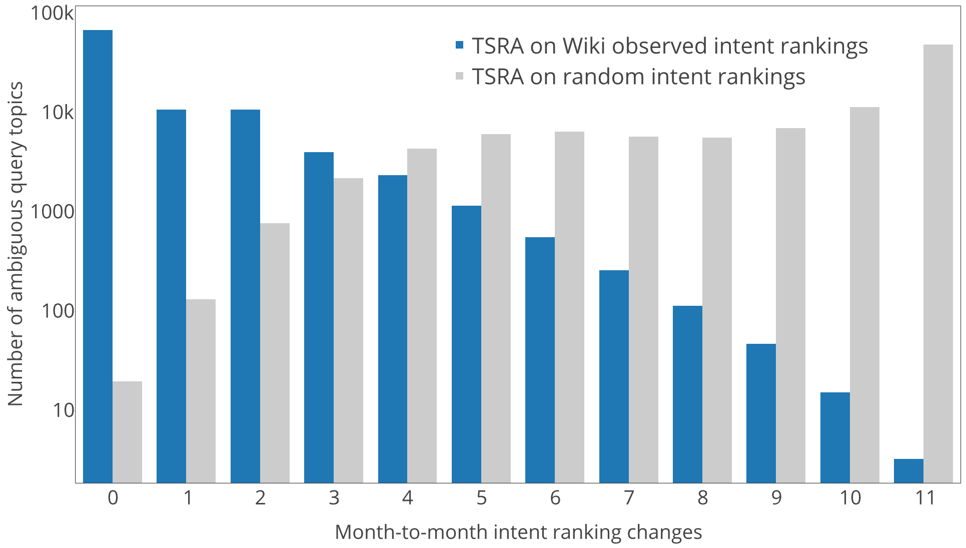


Figure 5.6: Histogram showing the number of ambiguous queries with *n* month-to-month intent ranking changes (0 = no ranking changes).

Month-to-Month Temporal Dynamics

In Figure 5.6, I present the number of ambiguous query topics affected by month-to-month temporal dynamics. Note there are at most 11 month-to-month ranking changes possible in the 1 year (2014) experiment duration. The results graph shows a characteristic power-law distribution, with the majority (i.e., 70%) of ambiguous topics having no month-to-month ranking changes in 2014. A medium number of ambiguous topics have a few ranking changes, i.e., 30% have between 1 and 6 month-to-month ranking changes over the year. Only a very small number of ambiguous topics have ranking changes occurring almost every month. Anecdotal analysis of the results shows these topics are mostly composed of major annual events with ambiguous aspects, such as the "`Battle for Gaza`" and "`2014 World Cup`". Since these queries make up a considerable proportion of web search query volume, in Section 5.5, I further study these event-driven query topics, and propose methods to better support them over time.

Weekdays-to-Weekends Temporal Dynamics

In Figure 5.7, I present the number of ambiguous query topics affected by weekdays-to-weekends temporal dynamics. Similar to month-to-month ranking changes, weekdays-to-weekends ranking changes also exhibit a power-law distribution, except, the majority of ambiguous query topics have one or more weekday-to-weekend change. Analysis of the results shows 3-4 letter acronyms, such as "`OSD`", "`DAT`" and "`TCPA`" mainly comprise the queries with the most persistently changing rankings over this time frame.

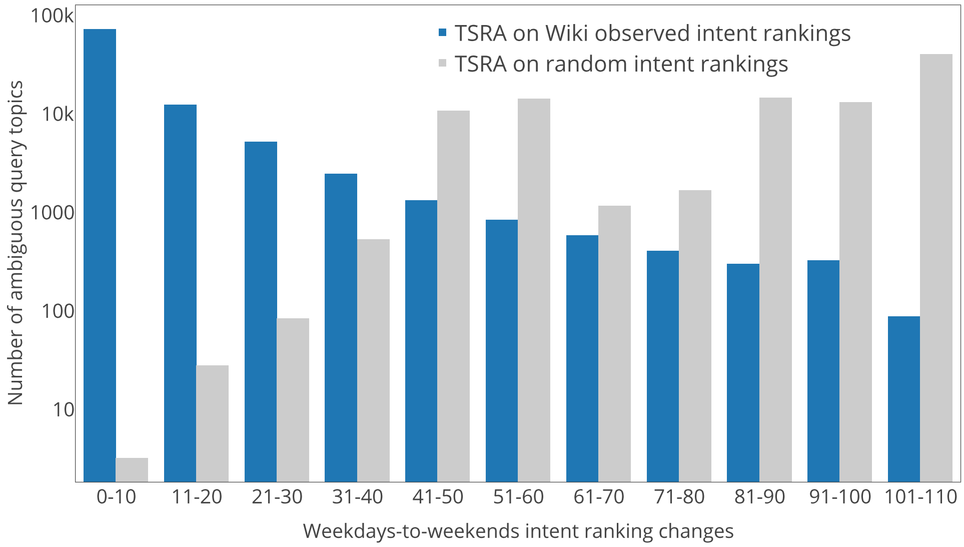


Figure 5.7: Histogram showing the number of ambiguous queries with n weekdays-to-weekends intent ranking changes (0 = no ranking changes).

Day-to-Day Temporal Dynamics

In Figure 5.8, I present the number of ambiguous query topics affected by day-to-day temporal dynamics. Few ambiguous queries have ranking changes over more than 200 days a year, however, the vast majority have between 1 and 150 daily ranking changes. Analysis of the results shows ambiguous entities such as names, films and TV shows comprise the most persistently changing rankings over this time frame.

Part-of-Day to Part-of-Day Temporal Dynamics

In Figure 5.9, I present the number of ambiguous query topics affected by part-of-day to part-of-day temporal dynamics. With a similar power-law trend evident as for other time frames, a large number of ambiguous query rankings are changing throughout the day. Analysis of the results shows ambiguous location and place names feature prominently among the most persistently changing rankings over this time frame, and even more so when examined hourly.

5.4.5 Discussion

In this section I discuss the results presented in the previous section with respect to the research sub-questions in this chapter related to ambiguous queries (i.e., RQ2.1 and RQ2.2).

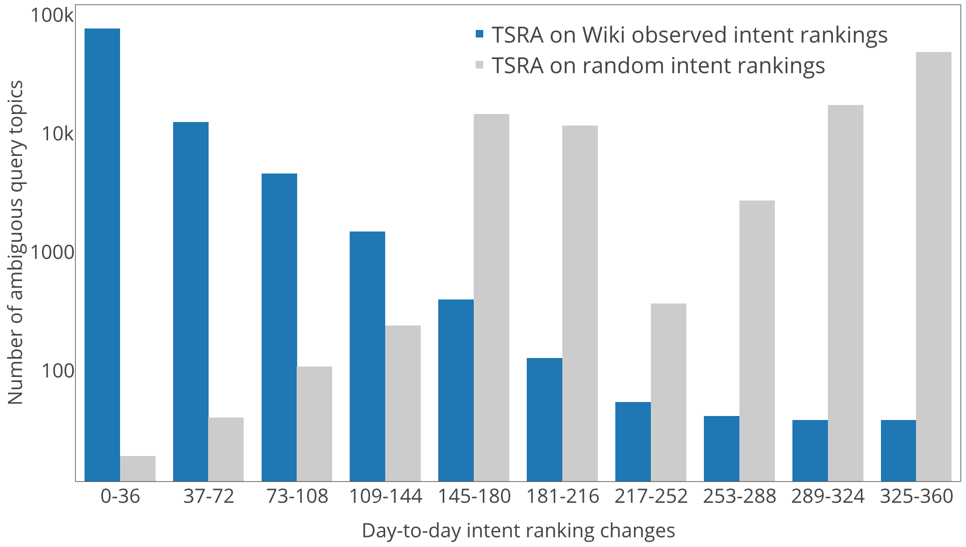


Figure 5.8: Histogram showing the number of ambiguous queries with *n* day-to-day intent ranking changes (0 = no ranking changes).

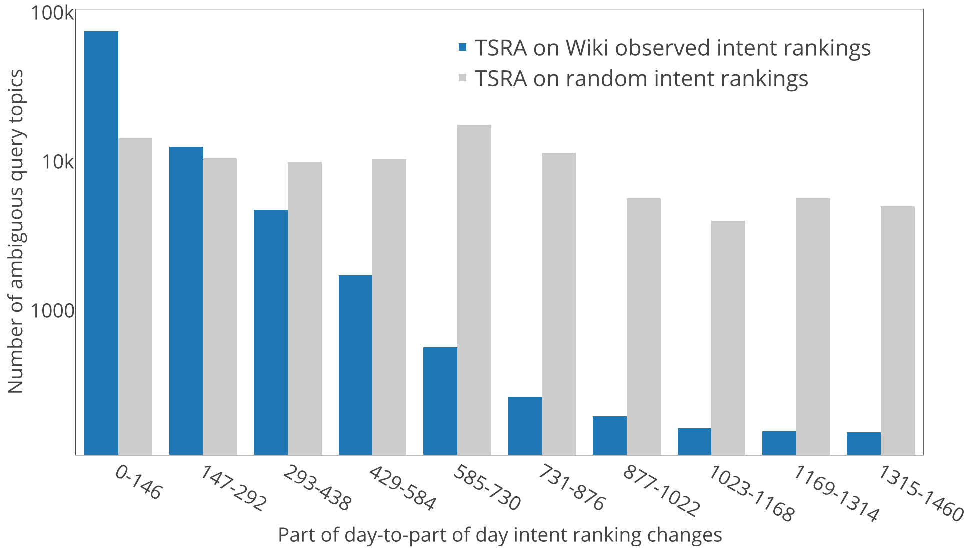


Figure 5.9: Histogram showing the number of ambiguous queries with *n* part-of-day to part-of-day intent ranking changes (0 = no ranking changes).

RQ2.1: To what extent are ambiguous queries affected by occasional (i.e., event-driven) ranking changes?

Occasional ranking changes are characterised by ambiguous topics with one or few ranking changes over time, but, which do not have ranking changes persistently occurring *every* time slice. The leftward skew of all frequency distributions in Figures 5.7, 5.7, 5.8 and 5.9 show that occasional ranking changes, caused by events and phenomena, are common for the majority of ambiguous queries over all the short and longer time frames I studied (i.e., hours to months). Long-term major news events are strongly represented in longer-term optimal ranking changes, e.g., month-to-month. In contrast, more 'temporary' events such as TV, news, sport and politics only occupy the user's interest for shorter-term time frames, and so only change optimal ranking day-to-day. Despite these ambiguous queries changing only occasionally, together these types of queries are likely to account for a relatively large proportion of query volume, and so the impact of supporting their temporal dynamics is valuable for increasing user satisfaction over time.

RQ2.2: To what extent are ambiguous queries affected by consistently periodic ranking changes?

Consistently periodic ranking changes are characterised by ambiguous topics with ranking changes occurring persistently every time frame. While consistently periodic ranking changes are common, the leftward skew of all frequency distributions in Figures 5.7, 5.7, 5.8 and 5.9 show far fewer ambiguous queries are affected by persistent ranking changes over any time. Most interestingly, I found for short time frames (e.g., hours to days), the most persistent ranking change for ambiguous queries tends to occur for acronyms, entities and places. As shown in Chapter 3.4 (Section 3.9), transition from work to leisure activities in the evening and weekend, and conversely, will likely account for some of this natural ranking change. Moreover, location and time are closely related – international users will be searching for different places at different times, given their time zone. Related effects are probably due to TV shows and films etc. being aired on major media networks in different time zones.

In the next section I turn my attention to event-driven multi-faceted queries, and consider how to model emerging query intents and their popularity over time.

5.5 Temporal Dynamics of Multi-faceted Queries

In this section I turn my attention to multi-faceted queries. Recall, a query is multi-faceted if it describes a broad topic consisting of multiple aspects. The user may be most interested in information related to one or more of these possible facets.

I focus on the temporal dynamics of multi-faceted queries, in particular, those related to recent and ongoing events – i.e., *event-driven* multi-faceted queries where temporal change is prominent. These types of query are responsible for a considerable proportion of daily web search engine traffic (Adar et al., 2007; Kairam et al., 2013; Kulkarni et al., 2011). However, providing effective results for these queries is problematic since users expectations can change very quickly (Dong et al., 2010a,b; Kairam et al., 2013). For example, consider the 2011 'Libya Intervention', for which key query intents are shown in Figure 5.10. Since the event began, several new query intents have emerged, such as NATO responsibility, opinion polling, political debate, oil controversy, Gaddafi's death and the later US Embassy attacks. Although all intents are now present, during the event each intent has temporally become the focus of attention. During the period of military action, query intent was most likely related to the air strikes and military movements. Once the country had been liberated, focus instead turned to how to rebuild the country as well as controversy surrounding oil. Similarly, debates before and after action held different meanings – earlier debate was related to the legality of the action, whereas latter debate was related to the political debate surrounding Libya during the US Elections.

Intent-aware search result diversification approaches typically rely on past query log mining to discover and model the popularity of query intents. However, for event-driven queries the past query log can be either outdated or unreliable, making it difficult to characterise real-time user expectations and consistently satisfy them (Dong et al., 2010b). This problem motivates the need to look at methods of characterising temporal query intents and their popularity *without* relying on query logs. Henceforth, I look towards changes in relevant content for clues on what the user is likely expecting. To that end, I use content dynamics in related information – in particular, the large-scale real-time collaboratively edited Wikipedia encyclopaedia – for clues to characterise multi-faceted query intents and their temporal popularity, in turn facilitating effective time-sensitive search result diversification for event-driven multi-faceted queries.

As I explored in Section 5.3, with users collaborating globally, major events and phenomena are often described shortly after their occurrence. Previous work has exploited Wikipedia taxonomy structure and article linking for representing non-temporal intents in intent-aware diversification (Hu et al., 2009). However, to model finer-grained intents of event-driven information needs, I propose to exploit the content structure of articles related to the information need. Furthermore, based on the assumption that content which changes regularly is popular (i.e., users edit what is currently of interest), I hypothesise that the editing activity of the content structure in turn reflects its popularity as a query intent. Accordingly, I investigate this hypothesis as RQ2.3 and RQ2.4 in this chapter.

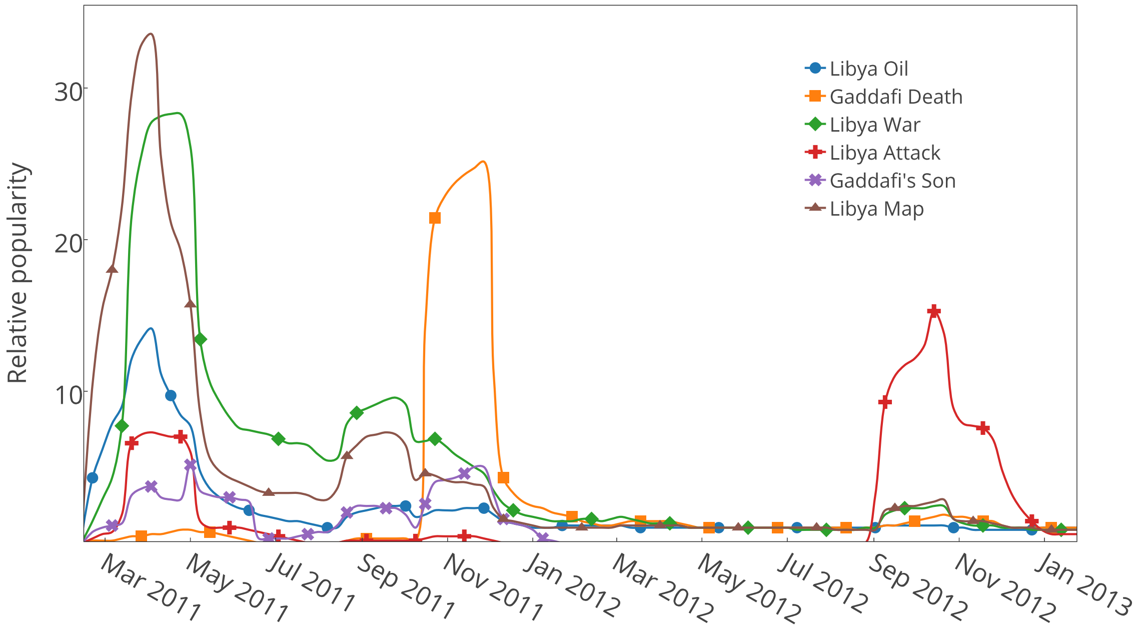


Figure 5.10: Temporal dynamics of popularity from January 2011 for six query intents of the Libya Intervention. Each time series was normalised (maintaining magnitude) and exponentially smoothed for clarity ($\alpha = 0.35$).

5.5.1 Deriving Query Intents from Content Structure

To answer RQ2.3, in this section I evaluate the effectiveness of deriving query intents from Wikipedia article content structure (i.e., article sections), compared to query intents obtained from ground-truth query log. As I am studying events following their occurrence, for simplicity at this stage I assume that all ground-truth query intents are present during the event.

The experimental procedure conducted is outlined in the following sections. I first select a set of 20 event-driven queries/topics to build a test set for answering RQ2.3 and RQ2.4. Following this, I describe how to obtain the ground-truth of possible query intents, and describe the methodology for assessing the matches between ground-truth query intents and Wikipedia article sections. Finally, I present evaluation results of deriving event-driven query intents using Wikipedia article structure.

Queries

I select two categories of event-driven queries for evaluation, related to significant events between January 2010 and December 2012. Four international graduate students were asked to collectively identify major events in the Wikipedia 2011-2012 News pages[1], and provide the query they would have used to find general information on the event. From this pool of events

[1] `http://en.wikipedia.org/wiki/[2010|2011|2012]`

Table 5.1: Example short- and long-term event-driven queries, along with their multiple query intents (obtained from Google Related Searches).

Query (topic)	Query intents (from query-log)
`Eyjafjallajokull` *Short-term* *(Topics 1-10)*	eyjafjallajokull **effects**, eyjafjallajokull **facts**, eyjafjallajokull **volcano webcam**, how to **pronounce** eyjafjallajokull, eyjafjallajokull **bbc**, eyjafjallajokull **case study**
`Libya Intervention` *Long-term* *(Topics 11-20)*	libya intervention **responsibility to protect**, libya intervention **poll**, libya intervention **debate**, libya intervention **timeline**, libya intervention **nato**, libya intervention **legality**, libya intervention **oil**, libya intervention **success**

and queries I selected two sets of 10 topics based on the following characteristics.

All topics are themselves an event, with each having a central descriptive Wikipedia article. Topics 1-10 are relatively short events (e.g., severe weather or a shooting), which have most temporal interest between 1 to 14 days. In contrast, topics 11-20 are prolonged events which happen over many weeks, months or even years (e.g., the Libya Intervention). Often these longer events are composed of many facets, concerning different people, places and interaction over time[1]. Two example categorised queries with their multi-faceted intents are presented in Table 5.1. The motivation for choosing these two categories of queries is to reflect events with diverse temporal characteristics, evolution and interest. A full listing of the events/queries used is provided in Appendix B.

Query Intents

I first obtain the ground-truth of query intents for each event-driven query. For large-scale commercial search engines, the ground-truth of intents should be based on a large number of users. Since I do not have a query log covering the event periods, I instead propose an approach to derive intent ground-truth using features provided by a commercial search engine. Since Google is the most universally used commercial web search engine, I examined the suggestions provided by Google Query Auto-Completion, Google Related Searches and Google Trends Related Searches. To select the best source, I define the criteria as follows: (i) the source should cover a variety of diverse intents/facets of an event, and (ii) it should cover the most popular query intents so that temporal statistics can be obtained. Based on my observations, I believe that the queries suggested by Google Related Search met the criteria and therefore I selected it as the ground-truth for this study. Google Auto-Completion and

[1] In this work I do not investigate seasonal event-driven queries such as Christmas. Instead, I am interested in events that are much more difficult to anticipate given a lack of past evidence and regularity.

Google Trends Related Searches data either over-reward tail queries or do not cover multiple diverse query intents. An example of ground-truth query intents obtained in this way is shown in Table 5.1.

Intent Matching and Assessments

To establish which ground-truth query intents are reflected by Wikipedia article sections, I attempt to match each ground-truth intent to a possible section and furthermore, assess the match strength. This consists of several steps: (i) *event article identification*: identifying multiple Wikipedia articles that are most related to event-driven topics, (ii) *section-intent automatic matching*: retrieving sections from the articles identified above, that might match the intents (for further assessments), and (iii) *match assessments*: assessing match strength between retrieved sections and query intents. I illustrate each step in detail below.

1. **Article Identification.** Before listing all the candidate sections that can be potentially matched to the query intents, the set of Wikipedia articles most related to each event-driven topic, $\{A_{topic}\}$, must be identified. Major events are typically represented by a central article (e.g. "Occupy Movement"), with related articles detailing substantial aspects such as "Reactions to the Occupy Movement", "Occupy Movement in the United States" and "Occupy Canada". As this work concentrates on a small number of topics I manually identified related articles as those linked from the central article via "See also:" and "Main article:" references, although past work has proposed automatic methods (Hu et al., 2009).

2. **Section-Intent Auto-Matching.** I posit that a query intent is reflected by one or more sections (or, subsections) contained in $\{A_{topic}\}$. For example, the 'Occupy Movement' article has sections including: 'Background', 'We are the 99%', 'Goals', 'Methods', and 'Protests' (with a subsection for each participating country). Despite the hierarchical nested structure of Wikipedia sections, to avoid complexity I employ a flat section structure. Hierarchy is particularly challenging for Wikipedia articles as it will change dramatically over time, so is left for later work. Matching between query intents and article sections was performed semi-automatically. To begin, I took each ground-truth query intent and extracted the intent key terms and automatically retrieved up to three sections from $\{A_{topic}\}$ which most contained the term in their header title or text. For example, for the query intent 'libya intervention oil', I identified article sections containing the term 'oil', such as 'Controversy' and 'Oil Supply Disruption'.

3. **Match Assessments** With the large pool of potentially matched sections retrieved by the system described above, two separate individuals were asked to annotate the extent

to which each section reflected the intent. Assessments were made in three grades: either a *strong* match (i.e., section is entirely about the intent), *weak* match (i.e., loosely related) or *no* match. For 92% of intents, the annotators were in agreement of match grade. For the remainder with labelling conflicts, where a no match label was present, it was selected by default. Similarly, where a 'strong' and 'weak' label were selected for an intent, they were resolved by defaulting to a 'weak' label. This annotated dataset provided the final ground-truth intent and Wikipedia section matches for study.

5.5.2 Deriving Query Intent Popularity from Content Dynamics

In the previous I proposed an approach and methodology to evaluate the effectiveness of deriving multi-faceted query intents from Wikipedia article content structure. In this section, to answer RQ2.4 I observe how the temporal dynamics of query intent popularity is reflected by content dynamics, that is, the frequency of section change activity in Wikipedia.

With that goal in mind, I compare the ground-truth temporal popularity of a query intent (obtained from a query-log) to the temporal dynamics of changes made to the intent representation in Wikipedia, i.e., the frequency of changes made to the informational content in the particular article section.

Experimental Procedure

Comparing the temporal dynamics of ground-truth query intents and Wikipedia article sections requires comparable time series of query volume and section change activity.

Previous user behaviour contained in query logs captures the past popularity of query intents over time. As I do not have access to a suitably large-scale and long-term query log, I rely on the temporal query volume data provided by Google Trends[1] as the ground-truth popularity temporal dynamics for each query intent. An example illustrating the comparison of these temporal dynamics is shown in Figure 5.11 for two query intents of the 2011-2012 Thailand Floods. The significant correlation between query intent ground-truth popularity and related Wikipedia section changes is prominent.

Temporal changes of Wikipedia article sections can be obtained by comparing contiguous Wikipedia article revisions. The stream of changed sections can be mined from the stream of article revision text. Standard `diff` and `patch` operations identify locations of text changes between adjacent revisions. Each change can in turn be resolved to a specific section by seeking its nearest parent section title header.

[1] `http://www.google.com/trends`

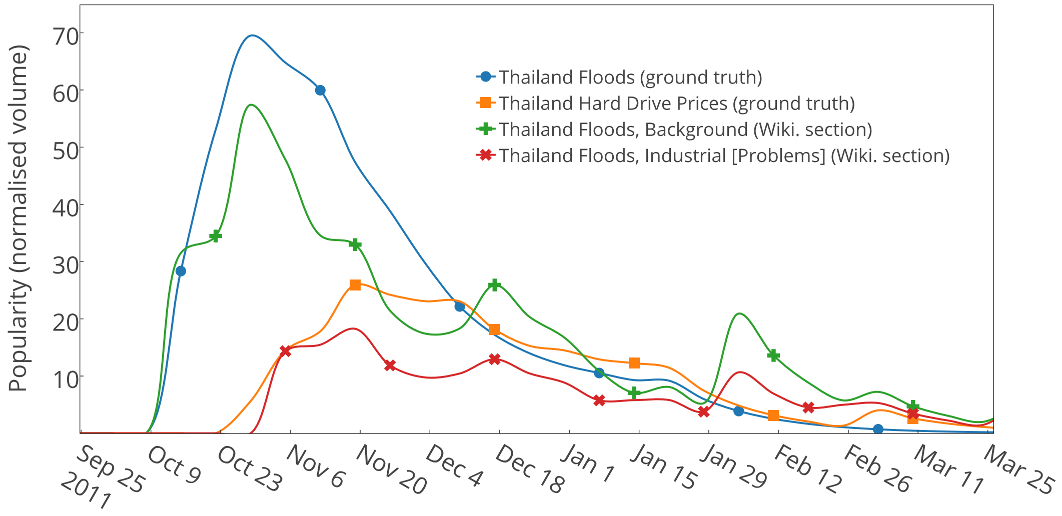


Figure 5.11: Temporal dynamics of popularity from 2011-09-18 for two query intents of the 2011-2012 Thailand Floods. Each time series was normalised (maintaining magnitude) and exponentially smoothed for clarity ($\alpha = 0.35$).

Section structure (e.g., section presence and hierarchy) is constantly evolving during collaborative editing. In some cases, this poses challenges for identifying the relevant section after sequential article revisions. In many cases, a re-organisation leaves the original content, yet provides new structuring.

Evaluation

To compare the ground-truth query intent popularity and related Wikipedia article section change activity, I aggregate observations to represent their temporal dynamics as time series of differing temporal resolutions. Using down-sampling, I experiment with three temporal resolutions: 1, 7 and 14 days. A lower resolution (i.e., 14 days) acts to smooth erratic temporal variation and noise which would be more apparent at higher resolutions, where short-term temporal factors may also effect comparison (e.g., weekday-to-weekend natural variance). Pearson's correlation co-efficient, r, is used to measure the temporal correlation between the temporal dynamics. Time series representing the temporal dynamics were constructed and down-sampled as outlined in Section 3.13 on page 53.

5.5.3 Results

In this section I report results from the experiments conducted for RQ2.3 and RQ2.4, using the approaches and methodology outlined in the previous sections.

Table 5.2: $Recall_{wiki}$ of weak and strong matched intents, for topics 1-20 (All), 1-10, and 11-20.

	Weak	Strong
Topics	$Recall_{wiki}$	$Recall_{wiki}$
All	0.89	0.68
1-10	0.87	0.64
11-20	0.92	0.72

Table 5.3: Average temporal correlation between Wikipedia section change activity and ground-truth query popularity.

	Average Pearson r		
	Temporal resolution		
Topics	1 day	7 days	14 days
1-10	0.32	0.49	0.58
11-20	0.15	0.25	0.33

Deriving Query Intents from Content Structure

In Table 5.2 I report $Recall_{wiki}$ to evaluate the effectiveness of query intent representation in Wikipedia article sections against the ground-truth, (i.e., # intents that Wikipedia covers / # total intents). The recall is initially computed per-topic, however I report the average computed over all, short- and long-term topics for comparison.

In all cases, recall of query intents by Wikipedia sections is medium to high. Around 10% to 20% more query intents are matched with weak matching rather strong matching. Weak matching yields the highest recall of query intents with 0.89 for all topics.

Deriving Query Intent Popularity from Content Dynamics

In Table 5.3, I report the average correlation r between the ground-truth and Wikipedia intent representation popularity, for each temporal dynamic time series resolution, and topics 1-10 and 11-20. Short-term events have the greatest temporal correlation correlation, particularly at larger temporal resolutions.

5.5.4 Discussion

In this section I discuss the results presented in the previous section with respect to the research sub-questions in this chapter related to event-driven multi-faceted queries (i.e., RQ2.3 and RQ2.4).

RQ2.3: Does content structure of related information reflect query intents for event-driven multi-faceted queries?

In Table 5.2, I observe that Wikipedia sections do substantially reflect user's query intents as there is a relatively high $Recall_{wiki}$ for both strong (0.68) and weak (0.87) match assessments, thus answering RQ2.3. From closer examination, it is apparent the query intents without coverage are generally those related to a specific resource (e.g., "bbc"), or generic type of information (e.g., jokes or videos), and so are missing from Wikipedia. These query intents are less likely to change over time as they refer to generic facets common for many event-driven topics, and so, their lack of presence in Wikipedia does not harm the general applicability of this approach.

RQ2.4: Do content dynamics of this structured information correlate with query intent popularity for event-driven multi-faceted queries?

In Table 5.3, I observe the temporal correlation between query intent popularity and Wikipedia section editing temporal dynamics. Considering the raw Wikipedia article revision stream is relatively noisy (e.g. if one editor repeatedly commits tiny changes), short-term topics have a relatively strong correlation at all temporal temporal resolutions. Evident is that as the temporal resolution increases, correlation is increased as daily noise is aggregated and smoothed. For long-term events, correlation increases with a larger temporal resolution, e.g., 14 days. This is likely caused not only by noise smoothing, but also the fact that longer events may consist of aspects which develop more slowly over many weeks, rather than just days.

5.6 Chapter Conclusions

Short and non-specific queries are a common problem for search engines. Such queries lead to uncertainty about the user's information need, and hence, what information is relevant. Addressing this problem, intent-aware search result diversification approaches interleave search results covering the most popular interpretations (or, query intents) to satisfy as many users possible. Since uncertainty changes over time, in this chapter I considered time-sensitive search result diversification as a practical means to integrate temporal relevance into time-aware IR.

In this chapter I explored the temporal dynamics of query intents for both ambiguous and multi-faceted queries, with a view to facilitating temporal relevance in time-aware IR. Given that an adequately large-scale and long-term query log is unavailable for time-aware IR research, I first argue the suitability of Wikipedia as a surrogate source of temporal dynamics. This argument is based on the findings of several studies which have examined topical, coverage, real-time and behavioural aspects of Wikipedia from different perspectives. Using temporal dynamics sourced from Wikipedia, I have analysed the temporal dynamics evident in ambiguous queries, and developed techniques to support the temporal dynamics in event-driven multi-faceted queries. The findings made in this work provide a foundation to future work in modelling temporal relevance as time-sensitive search result diversification. In the following two sections I outline the conclusions specific to ambiguous and multi-faceted queries.

Temporal Dynamics of Ambiguous Queries

Temporal dynamics play a central role in the intended intent of many ambiguous queries. Optimal intent-aware result ranking is far from stationary for the overwhelming majority of ambiguous queries. Indeed, ranking changes frequently over hours, days and months because of both random effects, and to a much greater extent, temporal dynamics. Importantly, differing types of ambiguous queries exhibit varying temporal dynamics. Person/place/film entities change hourly throughout the day, in contrast to acronyms which change daily. Neither periodicity nor event-driven ranking changes are exclusive. As illustrated in Figure 5.4, and indicated by the results I presented in Section 5.4.4, many distinct temporal dynamics interact to produce complex temporal ranking effects. Modelling these compound effects is a considerable challenge. Established time series modelling approaches (Radinsky et al., 2013b) will likely need to be adapted to provide truly time-sensitive ranking, which is proactive to the changing influences and needs of web search users – which is often not predictable solely from past popularity evidence. Overall, the findings of this work hold several implications for developing temporal relevance approaches to support real-time information seeking behaviour, including: (1) the need to support varying temporal dynamics for intrinsically different types of queries over hours, days and months, (2) past query intent popularity evidence is important, but cannot be considered in isolation, (3) external sources may offer insight for anticipating upcoming ranking changes caused by unpredictable event-driven influences. Moreover, with sufficiently large-scale ground truth query intent ranking provided by Wikipedia, "back-testing"[1] methodologies can be employed to evaluate future temporal relevance models based on past temporal query intent popularity and social signals.

One clear limitation of these findings that they are based on Wikipedia data, rather than real long-term query log past interaction data – which is not available for open research. While I argued that Wikipedia temporal dynamics reflect those found in query logs in many cases, there may be many instances where that is not the case. Consequently, while these findings are suggestive of the temporal dynamics found in ambiguous queries, further work is necessary to validate these findings with a real large-scale proprietary query log. By analysing a full

[1] *Back-testing* refers to evaluating a model on past time periods, as is commonly employed to validate quantitative finance models in diverse market conditions.

query log, as well as understanding what percentage or ambiguous queries are affected by temporal dynamics, further analysis will highlight the percentage of overall query volume that is affected – and hence, the satisfaction impact possible.

Temporal Dynamics of Multi-faceted Queries

Temporal dynamics play a central role in the query intent popularity of event-driven multi-faceted queries. Without knowledge of the emerging query intents, and their temporal popularity, applying effective intent-aware query ranking is problematic. Reducing the reliance on past query logs to mine query intents, I have found related content structure and dynamics can be used to derive multi-faceted query intents for event-driven topics (e.g., long-running news events). In particular, I have shown that the majority of major event-driven query intents can be represented by Wikipedia article sections and subsections. Moreover, the popularity of each of these intents over time is equally reflected by the editing activity of informational content in each section. Consequently, Wikipedia article structure offers a means to understand (i) the query intents currently present and emerging, and (ii) the temporal popularity of each intent. Overall these findings lead to new methods for modelling query intents to support intent-aware search result diversification in scenarios where there is insufficient, or outdated past query log evidence to adequately rank query intents in real-time.

Future work on this theme will investigate how this approach can be extended to more general content (e.g., all web pages), and less quickly evolving multi-faceted queries which may still have temporal elements. Moreover, given the manual data collection and cleaning necessary to evaluate the proposed approach, I employed a relatively small sample set of event-drive queries which may not be fully representative of all types of event-driven queries (Kairam et al., 2013). Accordingly, future work needs to characterise differing types of events (e.g., sports, natural disasters, politics, and so on) over more periods of time, to ensure any underlying temporal event factors such as the impact and social interest are better captured to validate these findings, and highlight any types of events that may be problematic (e.g., users are less prone to update the content in real-time in Wikipedia, or alternatively, or more likely to falsify content on Wikipedia).

Part III

Exploiting Temporal Dynamics in Collections

Chapter 6

Temporal Semantic Query Expansion

In Chapter 3, I explored several temporal dynamics evident in information collections, including patterns and trends in word and phrase use over time. Despite the inherent temporal dimension of many collections – such as the web, news and tweets – the majority of information retrieval research has concentrated on developing models based on a static view of the collection as a whole. In particular, the statistical measures and distributions used to characterise information, relationships and topics by conventional retrieval models are often assumed to be stationary over time. In this chapter, I explore methods to exploit the temporal dynamics of index term use in time-based collections to improve retrieval performance. Based on the epiphenomenon that semantically similar index terms have highly similar temporal dynamics, I propose a novel approach for identifying a topic's *chronotype* terms – that is, the cluster of consistently temporally related words which comprise the topic. I exploit the terms uncovered by this method in automatic query expansion to improve IR system effectiveness for diverse time-based collections.*

6.1 Introduction

The majority of past information retrieval (IR) research has typically centred around developing and evaluating approaches using relatively short-term or static snapshots of real-world time-based collections, including web documents, news stories and user-generated content such as blogs and tweets. As a result, such approaches disregard the underlying temporal dynamics captured in many time-based collections as they change over time.

This has led to a prevalence of retrieval models and related approaches developed to perform well empirically for static and stable collections. However, the majority of real-world collections evolve as new items are added incrementally over time. This means many of

*Research presented in this chapter is published in Whiting et al. (2011) and Whiting et al. (2012a).

the statistical measures relied upon by IR approaches – such as term frequency, specificity and semantic relationships – are far from stationary over time, as I showed previously in Chapter 3. Evolution in the composition of the collection may mean many IR approaches are not optimal over time. Nevertheless, static IR approaches fail to exploit the potentially rich insight in terms of temporal structure and meaning afforded by the underlying temporal dynamics of the collection. In particular, relatively little work has exploited the temporal dynamics of term popularity and semantic similarity in IR. In this chapter I look to exploit the temporal dynamics of term popularity in a collection for improving retrieval effectiveness through query expansion (QE).

QE is a technique commonly employed to augment a user's query with additional highly related terms to improve retrieval performance. It is motivated by the fact users often struggle to formulate specific and descriptive queries for their information needs. Indeed, Teevan et al. (2011) recently found an average of 3.08 and 1.64 words used for web and Twitter queries, respectively. More specifically, QE aims to improve the document matching capability and topic coverage afforded by a query, and thus ultimately improve relevance ranking. Pseudo-relevance feedback (PRF)[1] is a common approach to query expansion. PRF assumes the top-k ranked documents retrieved by the original user-provided query are relevant. Subsequently, by identifying distinctive terms contained within these documents, the original query can be expanded with further terms descriptive of the query's topic to improve retrieval effectiveness by better matching further relevant documents (i.e., increasing recall, and if possible, maintaining precision). Since the seminal work of Rocchio (1971), PRF has become established as an effective method to improve retrieval system performance on average (Carpineto and Romano, 2012). However, PRF is known to be problematic in many scenarios since it can have an erratic effect on retrieval performance as it can both substantially harm, as well as dramatically help individual query performance. It would be interesting to study how temporal dynamics can be exploited to improve the effectiveness of PRF techniques.

At the heart of any PRF approach is the method used for distinguishing the most distinctive terms from the PRF (or, *local*) documents. Distinctive terms will ideally only be found in relevant documents, and thus will discern relevant from non-relevant documents following PRF. The majority of established approaches (Lavrenko and Croft, 2001; Rocchio, 1971; Zhai and Lafferty, 2001) rely on non-temporal local term importance and global (or, collection-based) term discriminability statistical measures. In general, temporal dynamics have seen little consideration in QE and PRF approaches.

In this chapter, I posit the temporal dynamics of term popularity are valuable for identifying highly related (or, *semantically similar*) terms suitable for QE in time-based collections. Ac-

[1] Often also referred to as *blind feedback* or *automatic query expansion*.

cordingly, based on the epiphenomenon that semantically similar terms typically have very similar temporal dynamics (Alfonseca et al., 2009; Chien and Immorlica, 2005; Radinsky et al., 2011), in this chapter I hypothesise that the most effective terms for QE in a time-based collection are not only those which are distinctive in PRF documents (i.e., non-temporal evidence), but also those with a consistently high degree of temporal dynamic similarity with one another (i.e., temporal evidence). I consider these terms to comprise the query topic's *chronotype*. Importantly for QE, chronotype terms appear together in documents persistently over time – and thus, have a stable, or at least temporally significant semantic similarity over the collection time period. These terms are therefore optimal candidates for QE since they can distinguish relevant from non-relevant documents consistently over time in a time-based collection which contains documents covering evolving topics.

To examine the aforementioned hypothesis, I propose *Temporal Semantic Query Expansion* (TSQE). TSQE relies on a Temporal Semantic Network (TSN) to capture non-temporal (i.e., term frequency in PRF documents) and temporal (i.e, temporal semantic similarity between terms) evidence available for candidate QE term selection. Combining temporal and non-temporal evidence, network analysis of the TSN is employed to score the candidate QE terms and determine those which are most valuable.

6.1.1 Motivation

Despite the prevailing assumption in much of conventional time-insensitive IR, term semantics are rarely stationary over time. Of course, the period over which change occurs is dependent on the collection. For example, in rapidly changing real-time user-generated content (e.g. Twitter), words and phrase semantic relations could change in minutes as unfolding events take precedence. In contrast, books may take many years – even decades to centuries to reflect shifts in semantics (Michel et al., 2010). To consistently satisfy users over time, IR should take into account the inherent change in collections such as these.

Radinsky et al. (2011) establishes that temporal dynamic similarity is a strong indicator of semantic similarity between terms, however this finding has never been operationalised to improve IR system effectiveness. In an early preliminary study, I employed a naive independent term model for QE based on temporal dynamic similarity and found small but significant improvements in retrieval effectiveness. The approach and results of this study are summarised in Appendix C. In this chapter, I present the subsequent refined approach for QE based on exploiting temporal dynamics.

Isolated temporal dynamics of individual index term frequency have been explored for term weighting in IR. Liebscher and Belew (2003) argue that words whose frequency is high early in the period covered by a collection should be more favoured than more recently popular terms, however perform no IR experiments to conclude on their assumption. Although fundamentally different in the motivation and approach, Efron (2010) posits that temporal dynamics of a word's collection frequency can be used to better measure word specificity than non-temporal approaches (e.g. TF-IDF). More specifically, he finds that words with a future temporal dynamic predictable from prior temporal dynamics tend to be less discriminative for retrieval, and vice-versa. In this work, I am concerned with modelling temporal semantic relationships between topical terms, rather than their individual importance at the time of querying, for the purpose of QE.

Many established QE and PRF techniques (detailed in Section 6.2) rely upon statistical measures and distributions (e.g., inverse document frequency, or IDF) computed over the duration of the collection to calculate term discriminability, and therefore term suitability for QE. However, in the case of time-based collections these measures do not remain stable while the collection grows to include documents reflecting new and evolving topics, with language in changing distributions. In Topic Detection and Tracking (Allan, 2002), such changes are commonly used in tasks such as first story detection and story segmentation, yet have seen little application in IR approaches such as QE. Most often these temporally dynamic distributions are collapsed into single non-temporal global statistics, such as inverse document frequency (IDF) over the entire past collection.

Recent work has studied temporal QE term selection approaches using Temporal Query Modelling (TQM) (Choi and Croft, 2012; Peetz et al., 2014). TQM-based QE assumes any temporal aspects of a query will be reflected by temporal trends (e.g., bursts) in the time-stamp distribution of the PRF documents; and therefore selects expansion terms appearing primarily during the temporal bursts. However, not all topics may have strong temporal patterns, and indeed, they may not be robustly observed from the timestamp distribution of PRF documents – especially if the initial retrieval is poor. Accordingly, in an attempt to achieve an approach which is less sensitive to initial retrieval quality, I propose to model the lower-level artefacts of time in the collection, i.e., word and phrase popularity over time.

6.1.2 Research Questions

In this chapter I investigate the following specific research sub-questions, with regard to **RQ3** of this thesis: "Term popularity exhibits many patterns and trends over time in a time-based document collection. *Can these temporal dynamics be exploited to improve IR system effectiveness over time?*".

- **RQ3.1:** Can temporal semantic QE, combining non-temporal (i.e., term frequency) and temporal (in the form of temporal semantic similarity) evidence improve overall retrieval effectiveness?

- **RQ3.2:** QE approaches can yield erratic per-query performance. How does the proposed QE approach perform for individual queries?

- **RQ3.3:** What PRF conditions and QE approach parameters are necessary to maximise IR system effectiveness in diverse time-based scenarios?

- **RQ3.4:** If optimal PRF and QE approach parameters can be predicted per-topic, what is the expected impact on retrieval effectiveness?

6.1.3 Chapter Outline

This chapter is organised as follows:

- Section 6.2 presents related work from the large body of research on QE and PRF approaches.

- Section 6.3 details the proposed Temporal Semantic Query Expansion (TSQE) approach.

- Section 6.4 outlines the experimental setup and implementation, including collections and experimental baselines.

- Section 6.5 reports the evaluation results of the proposed QE approach in diverse test time-based collections, and discusses results with regard to the four research subquestions outlined in the previous section.

- Section 6.6 makes conclusions on the work and findings presented in this chapter.

6.2 Related Work

Presented below is a review of relevant work related to QE and PRF approaches. Work related to general time-aware IR is presented in Chapter 3.

For the interested reader, Carpineto and Romano (2012) provides an exhaustive survey on the long history of PRF approaches. In the following I outline influential approaches, and those most related to the approach proposed in this work. Regardless of implementation, the underlying premise of PRF is the same. It is assumed that the initial retrieval contains at least some documents relevant to the query. Hence, by sampling distinct features from these, the original query can be expanded to include terms likely to appear in relevant documents, and resubmitted with the intention of retrieve more similar and therefore also relevant documents.

PRF has long been established as an effective mechanism for improving retrieval performance through automated query expansion for often short or over-specific queries. For vector-space retrieval models, many variants of the original algorithm propose by Rocchio (1971) for query vector modification using TF-IDF exist. More recent state-of-the-art approaches in probabilistic language-modelling approaches are based on the Relevance Model (RM) (Lavrenko and Croft, 2001) and Mixture Model (Zhai and Lafferty, 2001). Principles of the RM approach are detailed in the background (i.e., Chapter 2) of this thesis.

Based on the influential conjecture by van Rijsbergen (1979) that relevant retrieved documents are closely clustered (that is, principle that relevant documents are similar to one another), clustering approaches have been used to sample related, and to discard non-related terms during PRF (Lee et al., 2008).

Graphical models and related graph-theoretic algorithms (such as random walks, clique or community detection and path finding, etc.) have been used extensively in natural language processing and IR approaches (Blanco and Lioma, 2012). Their ability to model complex network structure and interdependence has proven highly effective in many applications. However, no existing graph-based approach has employed temporal evidence provided by temporal semantic similarity between index terms for QE.

For query expansion, van Rijsbergen (1977) and Harper and van Rijsbergen (1978) employ a word dependence tree based on word co-occurrence, and use the maximum spanning tree to identify term dependencies to improve retrieval effectiveness. Collins-Thompson and Callan (2005) employ a Markov chain random walk model to a graph to combine multiple sources of lexical and semantic evidence of term association. Yin et al. (2009) select query expansion terms by using a random walk model from a graph of queries and associated result clicks, extracted from a query-log. Erkan and Radev (2004) and Mihalcea and Tarau (2004) use random walks on a graph-based model to select optimal terms for keyword and sentence extraction. Blanco and Lioma (2012) propose various IR ranking models based on graph-based term weighting, thereby modelling term relationships. In natural language processing, Siblini and Kosseim (2013) propose an approach based on a weighted graph of semantic relationships between terms (e.g., the lexicon, semantic relation types and definitions) to measure semantic similarity between terms. Since none of the aforementioned techniques are based on temporal evidence, I propose constructing a temporal semantic network to select QE terms based on a combination of temporal and non-temporal evidence – which is a novel approach for integrating temporal evidence in a retrieval model. An existing temporal QE approach is proposed by Peetz et al. (2014), who find QE based on terms in temporal bursts of pseudo-relevant documents significantly improves retrieval performance. Accordingly, I employ this approach as a state-of-the-art baseline for experiments reported in Section 6.5.

6.3 Temporal Semantic Query Expansion Approach

In this section, I propose an approach for Temporal Semantic Query Expansion (TSQE), which combines non-temporal and temporal evidence for QE. Non-temporal evidence refers to conventional index term distribution statistics, such as term frequency in PRF documents. Such statistics are commonly employed in existing PRF approaches to distinguish effective terms for QE. Temporal evidence refers to the temporal dynamics of index term popularity captured in the collection over time. Temporal semantic relationships between terms are implied by their temporal dynamic similarity (Radinsky et al., 2011).

The proposed TSQE approach consists of the steps summarised below, and detailed in the following sections:

1. Extracting all possible (i.e., *candidate*) QE terms from PRF documents.

2. Obtaining the term frequency temporal dynamics for these terms in the collection.

3. Measuring temporal semantic similarity between all candidate QE terms to characterise temporal semantic relationships.

4. Combining non-temporal evidence (i.e., candidate QE term frequency in PRF documents) and temporal evidence (i.e., relationships between terms implied by high temporal semantic similarity) to construct a topic-specific temporal semantic network (TSN).

5. Discovering the topic's *chronotype* terms – that is, the cluster of terms that temporally relate most consistently with one another, and further, are distinctive of the query topic (that is, they are distinctive in the feedback documents). Accordingly, chronotype terms are selected for QE.

6. Expanding the user's original query with chronotype terms.

6.3.1 Extracting Candidate QE Terms

Initially, all terms found in the top-k PRF documents are considered to be possible candidate QE terms. In this approach, I first filter these terms, then use non-temporal and temporal evidence to determine which of these terms are expected to be most valuable for QE.

Candidate QE terms comprise of single word terms (or, *unigrams*, e.g., "`market`") and composite two word terms (or, *bigrams*, e.g., "`stock market`") extracted using a sliding window algorithm over PRF document full text. Linguistic boundaries such as full stops, commas and colons are observed such that bigrams are formed only of syntactically adjacent unigrams.

Bigrams are necessary for this approach since many unigrams are ambiguous. For example, the term "`market`" might relate to a "`stock market`", "`open market`", "`market forces`", "`cattle market`", etc. Indeed, the document frequency temporal dynamics for "`market`" are likely to be very different to any of the more specific interpretations. As such, the bigram implies context of use, and ensures the relevant temporal dynamics are included in the Temporal Semantic Network (TSN). Note the retrieval model used in experiments reported later in this chapter are based on unigrams, so candidate QE terms are ultimately treated as individual unigrams, regardless of their composition in the TSN.

A large proportion of the possible terms consist of stop words, for example: "`and`", "`the`", etc. As these are extremely common terms, they are of little value to QE because they are unable to discriminate relevant documents, so unigrams or bigrams comprising any common stop word are discarded to reduce later processing. In order to maintain precise term meaning during TSN construction, no term stemming is applied. However, stemming is later applied by the unigram retrieval model in experiments. All terms are made case-insensitive, so upper- and lower-case term variants are treated equally. For the majority of cases this is acceptable, however, there will likely be rare exceptions where the context of the upper-case character is lost (e.g., "`CAT`" the construction machinery manufacturer[1] versus "`cat`" the animal). Future work will examine the effectiveness of extracting only key words and phrases, such as entities (e.g., people, places and companies) from PRF documents.

In the following section I outline obtaining term popularity temporal dynamics for the candidate QE terms identified by this process.

6.3.2 Obtaining Temporal Dynamics of Candidate QE Terms

The proposed approach relies on document frequency (DF) temporal dynamics[2] to measure temporal semantic relationships between terms. Consequently, the time series of document frequency over the duration of the collection, X, is obtained for each candidate QE term,

[1] Caterpillar, Inc. `http://www.cat.com`

[2] Document frequency is the number of distinct documents containing the term in each time frame covered by the collection.

and normalised to X' (using the number of documents present in each time frame), based on the procedure outlined in Section 3.13 on page 53. In an existing time-based collection (such as those used in the experiments for this work), the time series for each term is mined retrospectively. Alternatively, it can be sampled from the stream at index-time, or from inverted indexes split into time windows[1] (e.g., one index per day), as is commonly employed for real-time incremental indexing of streaming collections (Zhuang, 2014).

The time series resolution used for temporal dynamics varies between experimental test collections because of their overall duration and expected temporal variability, which I discuss later in Section 6.4.

Temporally Insignificant Term Filtering

Given the large number of index terms in PRF documents, I employ temporal filtering to reduce the number of terms needed to be modelled in the more computationally complex latter stages of the approach. Terms which do not exhibit any significant document frequency temporal dynamic patterns or trends are of little value for QE[2], and therefore can be discarded prior to TSN construction. This includes terms which exhibit relatively little variation over time either because they are extremely uncommon, and thus unlikely to be used in a the language of a query, cf. Luhn (1958). It also includes terms which are consistently popular and therefore, not temporally discriminative (e.g., stop words).

Such undesirable terms are identified by measuring the excess kurtosis of their DF time series. Kurtosis is a common descriptive statistical measure of the ‘burstiness’ of time series data (Jones and Diaz, 2007). Terms with a time series kurtosis < 5 (i.e., a very flat temporal dynamic with little temporal discriminability) are discarded at this stage, thereby reducing the size and complexity of the TSN to be constructed.

In the following section I outline measuring temporal semantic relatedness between index terms based on their DF temporal dynamics.

6.3.3 Measuring Temporal Index Term Semantic Similarity

A strong semantic similarity between two terms implies they are highly related in some way, and therefore likely to appear together in documents. As such, for the purposes of QE, a query containing one term would benefit from including the other highly related term to increase the likelihood of distinguishing relevant from non-relevant documents.

[1] Also known as *time-sharded* indexes.

[2] This finding was made in preliminary experiments, reported in Appendix C.

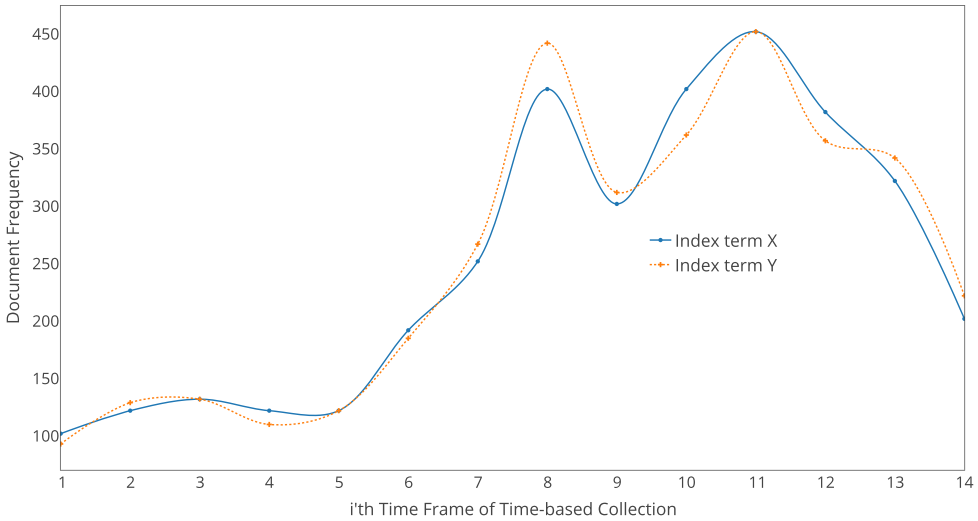


Figure 6.1: Example document frequency temporal dynamics of two semantically similar index terms, X and Y. The strong temporal correlation between their temporal dynamics reflects their temporal semantic similarity over the 14 time frame duration of the collection.

Two semantically similar terms are likely to appear in the same documents over time, so, their document frequency temporal dynamics can be expected to exhibit similar trends and patterns over time. This epiphenomenon is illustrated for two semantically similar index terms, say X and Y, in Figure 6.1. Note the high temporal correlation between the temporal dynamics, thus indicating a strong temporal semantic similarity. Indeed, as I considered in Chapter 3 (in Section 3.11.1), the semantic similarity between index terms is itself temporal, given new events and phenomena changing meaning over time. As such, temporal semantic similarity estimates the degree to which two index terms may be semantically related over any period of time in the collection by measuring corresponding (or, *correlated*) temporal dynamic patterns and trends. In this work I measure temporal semantic similarity over the duration of the collection prior to the query time (if available).

Defining a reliable function of temporal semantic similarity between index terms is important for effectively extracting a topic's chronotype, which is dependent on capturing the strength of temporal semantic similarity between terms. Two approaches have been proposed and studied. Radinsky et al. (2011) experiment with cross-correlation (since there is no time shift, it is in essence Pearson's product-moment correlation coefficient, r) and Dynamic Time Warping (DTW) as part of Temporal Semantic Analysis (TSA). They find that Pearson's r is most effective for measuring correlation between temporal dynamics. Similarly, Chien and Immorlica (2005) find that Pearson's r is effective for identifying many semantically similar queries based on the correlation of their popularity temporal dynamics. Derrick et al. (1994) observe that a high Pearson's r is always indicative of a high correlation between time series in the study of noisy human motion time series data. In early exploratory work prior to the TSN model I present here (cf. Appendix C, I found Pearson's r was found to provide small but statistically significantly improvement when used for measuring temporal semantic similarity in a basic QE approach. Hence, I am motivated to employ Pearson's r in this approach.

Given the evidence supporting Pearson's r for measuring semantic similarity with temporal dynamics, the approach presented here also employs Pearson's r correlation co-efficient as an estimate of semantic similarity. Pearson's (sample-based) product-moment correlation coefficient, r, is formally defined for two document frequency time series, X and Y, as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{6.1}$$

Where X_i and Y_i is the document frequency of term X or Y in time frame i of the time-based collection. Pearson's r symmetrically measures the strength and direction of a linear relationship between two variables. Pearson's r is normalised in the range $[-1, 1]$, with $r = -1$ and $r = +1$ indicating a 'perfect' negative and positive linear relationship, respectively. Meanwhile, $r = 0$ indicates no relationship between the variables.

Dealing with Spurious Correlation

Of course, measuring a high correlation co-efficient between two temporal dynamics does *not* imply causation – that is, a strong semantic similarity between the two terms. A high correlation co-efficient (e.g. $r > 0.7$) may be the result of a confounding variable or not statistically significant, and in the context of this approach, the supposed relationship is spurious.

This problem is apparent in a number of examples. In TREC-1 topic 70 (i.e., "surrogate motherhood"), temporal dynamics of the terms "`visitation rights`" and "`country music star`" correlate highly with $r = 0.92$, yet they seem to have no intuitive relationship between them. Constructing a TSN attempts to alleviate this issue by modelling correlation among all other terms, and attempting to outweigh the effect of inevitable spurious correlations by incorporating topic-specific non-temporal PRF term frequency evidence. Furthermore, since terms in the TSN have been obtained from PRF documents, there has already been some pre-filtering to terms related to the query topic by the retrieval model.

6.3.4 Topic-specific Temporal Semantic Network Construction

A temporal semantic network (TSN) captures the non-temporal (that is, *static*) statistical and temporal semantic evidence relating a set of terms. In this work, the TSN is considered *topic-*

specific since it is constructed from terms found in the PRF documents for a given query topic. As such, non-temporal statistical evidence is the term frequency in PRF documents. Temporal evidence is the temporal semantic similarity among terms. Together, the non-temporal and temporal evidence contained in the TSN is used for selecting effective QE terms with network analysis, described in the following section. In this section I motivate my TSN approach, and detail construction of it.

Inter-relationships between entities, concepts and phenomena have been modelled with networks in several domains. TSNs draw inspiration from two existing approaches proposed in other domains. Generic co-occurrence networks have typically been used to capture relationships between elements in diverse applications including text mining, linguistics, biology and ecology. Financial asset and stock correlation networks have attracted extensive attention in quantitative finance modelling. Network analysis is used to highlight tacit structure and predict correlated value movements over windows of time (Mantegna, 1999; Onnela et al., 2003). Such correlation networks capture the temporal associations emerging between financial instruments over time, often driven by transient business relationships and events. This is an analogous intuition to capturing temporal relationships emerging between terms in a time-based collection.

Formally, a network is represented as a *graph* – a formal mathematical object composed of vertices (or, nodes) and edges (or, connections). A TSN is formed as an undirected weighted graph, $G = \{V, E\}$. The subsequent process of QE term selection thus becomes a network analysis problem, which I discuss in the next section. The construction and characteristics of the graph are as follows:

- Each vertex v signifies a QE candidate term, extracted from the PRF documents.
- Assigned to each vertex v is a weight w_v corresponding to the frequency of the term in the PRF documents.
- Assigned to each edge e is a weight w_e corresponding to the measure of temporal semantic similarity (i.e., Pearson's r) between the two connected terms.

An example TSN for nine terms (t_1 to t_9) is illustrated conceptually in Figure 6.2. The size of each vertex (i.e., *node*) corresponds to the non-temporal term frequency of the term in PRF documents. Similarly, the thickness of each edge (i.e., *connection*) refers to the temporal semantic similarity measured between the connected index terms, based on their temporal dynamics.

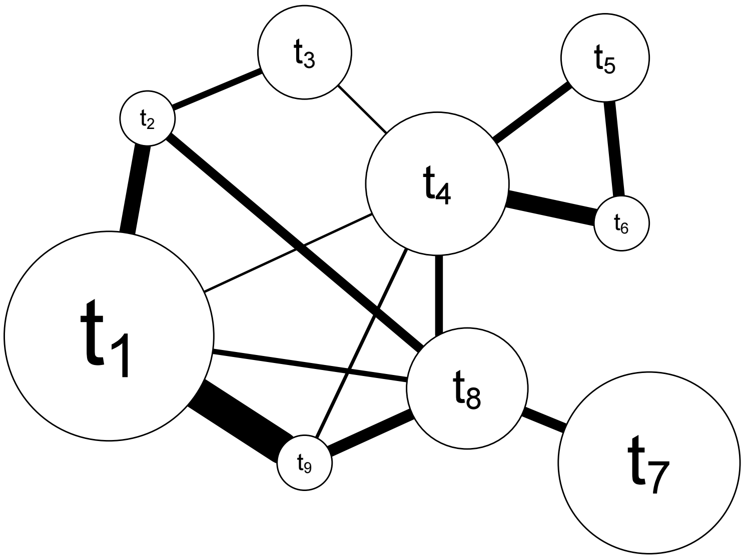


Figure 6.2: Example Temporal Semantic Network (TSN) for nine terms, with document frequency denoted by vertex size, and temporal semantic similarity by edge thickness. Edges representing very low correlation are discarded for illustration clarity.

Graph Pruning

With every QE candidate term having a temporal semantic similarity with every other term, the number of edges in the TSN will grow exponentially (i.e., N^2) with the number of terms. This complexity is alleviated by applying some straight-forward graph pruning heuristics which were found to have negligible impact on performance.

Firstly, since term frequency exhibits a characteristic long tail distribution (Zipf, 1949), there are a large number of QE candidate terms with very low frequency in the feedback documents. Terms seen only once in feedback documents are assumed unlikely to be valuable in QE, and so removed from the TSN.

Secondly, the temporal dynamic correlation measured by Pearson's r can be positive (i.e., > 0), non-existent (i.e., 0) or negative (i.e., < 1). Since this approach is concerned with positive temporal dynamics correlations which might indicate a medium to strong index term temporal relationship, edges that represent a Pearson's $r < 0.2$ are discarded. Hence, this step removes information in the TSN that is superfluous to the approach, and thus speeds up later processing of the TSN.

6.3.5 Chronotype Discovery through Network Analysis

The previous section details the construction of a topic-specific TSN, based on QE candidate terms found in PRF documents. This section outlines the final stage for identifying optimal candidate QE terms – by discovering the topic chronotype, and weighting chronotype term importance. Recall, the topic chronotype is the cluster of terms that temporally relate most consistently with one another, and further, are distinctive of the query topic (that is, they are distinctive in the feedback documents). I posit that it is these terms which will lead to optimal QE performance in a time-based collection.

Chronotype discovery is achieved by network analysis of the TSN, determining which candidate QE terms are most valuable based on temporal and non-temporal evidence. For this purpose, I employ PageRank with Priors (PRwP) as proposed by White and Smyth (2003) – a random walk network analysis algorithm – to measure the importance of candidate QE terms based on both non-temporal and temporal evidence captured in the topic-specific TSN. The characteristics and flexible behaviour of this established network analysis algorithm which make it suitable for combining evidence are discussed. Furthermore, the rank assigned to each term by PRwP is employed for weighting the importance of each term in subsequent QE.

A random walk model refers to the mathematical formalisation of a succession of random steps taken by a stochastic process. In this work I am most concerned with PageRank (Page et al., 1999), and its derivatives. Although initially designed to distinguish authoritative web pages based on web link analysis, PageRank is often used more generally to highlight the relative *importance* of vertices in a graph, based on their connectedness with other nodes.

Intuitively, the PageRank model is interpreted as a web surfer infinitely and randomly moving through the web graph, subject to occasional jumps to other randomly selected web pages (depending on the “damping” factor), thus stopping it becoming permanently stuck in dead ends. Consequently, it models the likelihood that the random surfer might visit any of the web pages (or, graph vertices) at any given moment in time. Hence, vertices are ranked according to a probabilistic distribution. With an intuitive measure of vertex association in place, PageRank can be used to estimate the relative importance of vertices as a function of their connectedness with other vertices with similarly defined importance.

The rationale for employing a random walk model for network analysis in this scenario, and in particular one derived from PageRank (Page et al., 1999), is that it is suitable for identifying the core of strongly interconnected QE terms contained the TSN, i.e., those terms with a relatively high temporal semantic similarity among themselves. Furthermore, with some modification, PageRank can be biased towards terms that are most distinctive of the topic in the TSN, i.e., terms with high frequency in the PRF documents.

PageRank with Priors

Temporal evidence is important, but so too is non-temporal evidence determining valuable QE terms. Neither should be used exclusively, and indeed both must be combined for optimal retrieval performance (cf. Appendix C). It is therefore important to develop an approach that supports setting the precedence of non-temporal versus temporal evidence contained in the TSN for selecting QE terms.

The PageRank with Priors (PRwP) algorithm (White and Smyth, 2003) is used to combine non-temporal and temporal evidence contained in the TSN in a intuitive manner. The PRwP algorithm extends the original PageRank algorithm, such that (i) it integrates edge weights representing the probability of the walker following an edge from a vertex, and (ii) it supports prior probabilities (i.e., biases) representing the relative importance of each vertex in the graph during a random jump. Additionally, rather than a damping factor, it has a back probability, β, where $0 \leq \beta \leq 1$ controls how frequently the random surfer jumps to another vertex (the selection of which is biased by the vertex priors).

As PRwP requires directed edges, the undirected TSN graph is adapted into a directed graph by replacing each undirected edge with two opposing same-weight directed edges. Normalized vertex priors are computed as:

$$p(v) = \frac{w_v}{\sum_{v \in V} w_v} \tag{6.2}$$

and, the probability of following an edge e from vertex v as:

$$p_{out}(e, v) = \frac{w_e}{\sum_{e \in e_{out}(v)} w_e} \tag{6.3}$$

I use the PRwP implementation as described in White and Smyth (2003), however several other more deterministic and efficient approximate variants exist, e.g., Rodriguez and Bollen (2006). The PageRank, π, for an index term represented by vertex v at iteration i is computed as:

$$\pi(v)^{(i+1)} = (1 - \beta) \left(\sum_{u \in d_{in}(v)} p(v|u) \pi^{(i)}(u) \right) + \beta p_v \tag{6.4}$$

Where β is the aforementioned parameter governing random jumps (i.e., the mix between temporal and non-temporal evidence). $d_{in}(v)$ is the set of vertices connected to v (i.e., other

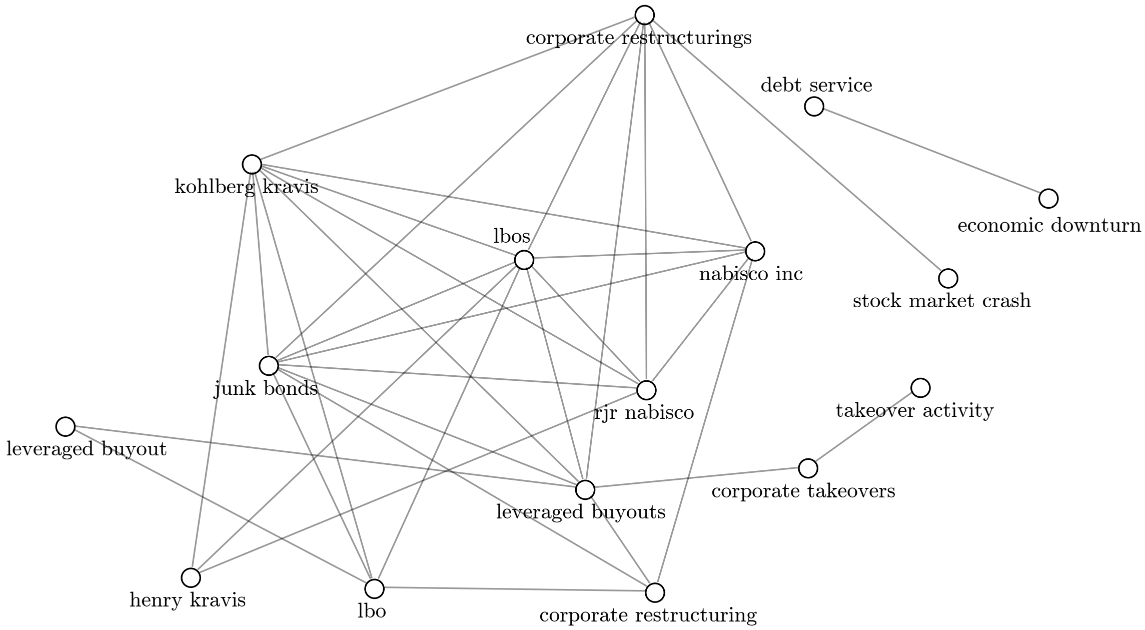


Figure 6.3: Chronotype terms identified through network analysis of the topic-specific temporal semantic network for TREC-1 topic 53 ("`leveraged buyouts`"). Filtered to edges with a high temporal semantic similarity of > 0.9 between terms for clarity.

index terms still deemed semantically following the graph pruning step). $p(v|u)$ is the probability of the random walker following the edge from vertex u to vertex v (i.e., the strength of the semantic similarity between the index terms). And finally, $\pi^{(i)}(u)$ is the PageRank, at the last iteration, of the other semantically similar index term vertex u.

The number of iterations necessary for the PageRank scores to stabilize varies depending on graph characteristics (Page et al., 1999). Since this is a relatively small and non-complex graph (at least when compared to web-scale graphs), I iterate 40 times to guarantee convergence.

Combining Temporal and Non-temporal Evidence

Adapting the degree to which QE term selection relies on non-temporal and temporal evidence is a desirable feature, since individual queries and collections are likely to have differing temporal characteristics.

The behaviour of PRwP can be adjusted by setting the back probability parameter, denoted as β. When $\beta = 1$, the random surfer will always jump randomly, therefore PageRank scores for terms will follow the same distribution as the vertex priors, i.e., using only non-temporal evidence. Conversely, when β = 0, the random surfer will never jump and so it will move using edge transition probabilities only, i.e., only temporal evidence. With $0 <$

$\beta < 1$, a mixture of both non-temporal and temporal evidence will be combined in PageRank computation. Experiments explore a range of β settings to determine the optimal mix of temporal and non-temporal evidence for effective QE.

Figure 6.3 shows the strongest chronotype terms identified in the topic-specific TSN for TREC-1 topic 53 (“`leveraged buyouts`”). PRwP was computed with $\beta = 0.1$ (i.e., mixing temporal with a little non-temporal evidence). The resulting 16 highest ranking (i.e., most important) candidate QE terms are shown. Only edges representing a temporal semantic similarity of > 0.9 between terms are included. Bear in mind that this topic-specific TSN covers 1988-89. It would likely look very different if constructed using a more recent collection where new entities, terminology and semantic similarities have emerged, thus supporting the case for using temporal semantic similarity in QE for time-based evolving collections. The highest ranking topic chronotype terms identified by this method are those used for QE in experiments presented in the following section.

6.4 Experimental Setup

I conduct comprehensive retrieval experiments using four time-based document test collections with diverse characteristics. Table 6.1 outlines the characteristics of each collection, such as the duration they cover and the time series temporal resolution used for representing index term document frequency temporal dynamics.

For the news wire collections, AP and WSJ, I use both TREC-1 ad-hoc topics 51-100 and TREC-2 ad-hoc topics 101-150. For the Twitter collection, MB, tweets by users with a non-English default language are discarded. TREC Microblogging Track 2011 topics MB001-MB050 are used. For the blog collection, Blogs06, I use the TREC Blog Track 2006 ad-hoc topics 851-900.

The diverse characteristics of each time-based test collection allow evaluation of the proposed TSQE approach under varying conditions. AP and WSJ are relatively small and clean collections, containing 164,597 and 173,252 documents respectively. MB contains approximately 16 million tweets of up to 140 characters. While tweets have a strong temporal dimension (Teevan et al., 2011), limited document length and noise poses several issues to traditional IR approaches. Blogs06 is a spam prone web collection containing approximately 3.2 million *permalink* documents and their associated blog feeds, with no editorial control over content or structure.

The temporal resolutions used to represent temporal dynamics were selected with the volume and velocity of the collections in mind. Twitter changes very rapidly, so a relatively short

Table 6.1: Details of experimental test collections, including the period of time the collection covers and the temporal resolution used for representing temporal dynamics.

Collection	Period	Temporal Resolution
Associated Press (**AP**)	12-Feb-1988 to 31-Dec-1989	7 days
Wall Street Journal (**WSJ**)	1-Dec-1986 to 24-Mar-1992	14 days
Microblogging Track (**MB**)	23-Jan-2011 to 8-Feb-2011	4 hours
Blog Track (**Blogs06**)	6-Dec-2005 to 21-Feb-2006	1 day

temporal resolution (i.e., 4 hours) is necessary to reflect change occurring throughout the day. In contrast, the AP and WSJ collections are based on daily reporting of short- and long-term news events, and so, I choose a larger temporal resolution (i.e., 7-14 days) which can represent the expected change over the multiple years spanned by both collections.

6.4.1 Obtaining Temporal Dynamics of Candidate QE Terms

All documents in AP, WSJ and MB are used to derive document frequency temporal dynamics of each QE candidate term. However for Blogs06, the provided RSS/ATOM syndication feeds are mined instead. As the permalinks were crawled 2 weeks after initial creation (to capture reader comments), they then contained future content. Unfortunately many syndication feeds contained only limited excerpts of the permalinks and so provided restricted insight into term use. AP and WSJ do not contain continuous streams of documents. Both had numerous gaps ranging from days to weeks, which were reflected in all temporal dynamics. Since this issue is consistent for all terms, the semantic similarity method employed is not adversely affected.

6.4.2 Query Timestamps

Only TREC MB topics include timestamps. As such, for TREC MB I use only temporal dynamics up to the query time, thus maintaining a realistic real-time search scenario (i.e., avoiding the use of future evidence). Similarly, retrieval for TREC MB is performed on indexes composed of only the tweets up until the query time. Using future evidence has been shown to invalidate retrieval experiments (Wang and Lin, 2014). For all other collections I assume that the query was performed at the end of the collection period and therefore used the complete temporal dynamics obtained for the duration of the respective collection.

6.4.3 Exploring β Parameter Settings

It is important to study the β parameter since it offers a means to influence the mix of temporal and non-temporal evidence during QE. Indeed, the retrieval performance of individual query topics is likely to be affected by the β parameter setting, depending on their initial retrieval performance, and temporal nature. Hence, to explore collection average and per-topic performance at each β parameter level, I sweep β parameter settings between $[0, 0.1, 0.2 \ldots 1]$ (i.e., from using only temporal evidence, to mixing evidence, to using only non-temporal evidence). To explore the lower β performance in higher granularity I also experiment with $\beta = 0.05$ and $\beta = 0.15$.

6.4.4 PRF Parameters

As I am evaluating a novel QE approach based on PRF with collections of diverse document characteristics, I also experiment by sweeping 3 PRF parameters known to have considerable impact on QE performance (Carpineto and Romano, 2012; Ogilvie et al., 2009): (i) number of feedback documents (ii) relative weight for expansion terms and (iii) expansion term selection scheme. I experiment with 5, 10, 20 or 30 top-ranked PRF documents, and with expansion term weights of 0.4, 0.5 and 0.6. These choices and related results are explained in the following section.

Further, I experiment with two schemes for expansion term selection: *static* and *adaptive*. For the static scheme, I use the N top ranking QE candidate terms ($N = 5, 10, 15$...30), according to their chronotype index term rank (i.e., computed by PRwP). For the adaptive scheme, in line with recommendations for adaptive approaches made in Ogilvie et al. (2009), the choice of terms is based on the heavy tail distribution of QE term ranks computed the temporal semantic QE approach. The adaptive method selects highest-ranked candidate QE terms until the sum of selected terms ranks is $\geq$ to the threshold, with thresholds equal to $0.05, 0.1, 0.15...0.45$.

6.4.5 Retrieval Effectiveness Measures.

To observe the ranking and precision effects of the model I report Mean Average Precision (i.e., MAP) and Precision at 10, 20 and 30 (i.e., P@10/20/30) for experiment runs. Precision is considered the primary evaluation metric for MB.

Evaluation is performed on the top 1,000 retrieved documents for AP, WSJ and Blogs06 topics. Similar to the TREC MB 2011 track evaluation strategy, evaluation is performed on the top 30 retrieved tweets re-ordered by recency, for MB topics.

6.4.6 Experimental Procedure

Documents were indexed using Indri 5.2[1], with built-in stop-word removal and Krovetz stemming applied. For the MB collection, a separate index for each topic was created containing only tweets up to the topic query time-stamp, thereby avoiding future evidence inclusion. Retrieval for all experiments was performed using a unigram language model (LM) with Indri default Dirichlet smoothing ($\mu = 2,500$).

All PRF is performed using the top-k documents retrieved by the LM baseline, reported in Table 6.2. Accordingly, I compute statistical significance based on this initial retrieval. To compare the effectiveness of the TSQE approach I employ an appropriate state-of-the-art temporal baseline, namely LM with Temporal Query Modelling (i.e., LM+TQM) as proposed by Peetz et al. (2014). TQM uses terms found in temporal bursts of pseudo-relevant documents (as retrieved by LM) to expand the query. The TQM variant employed in this work uses step wise burst decay[2], with the temporal distribution of documents based on their relevance score, and with parameters as defined in Table 3 of Peetz et al. (2014). This approach variant was shown to consistently perform well for the collections studied with the '*all*' query test sets (i.e., both temporal and non-temporal queries), and so was chosen as the baseline for this work.

6.4.7 Query Expansion

The expansion term weight (λ) is used to weight the importance of query terms added to the original query by PRF, with the weights of expansion terms being the linear combination of their chronotype ranks. Bigrams are treated as combined unigrams during expansion, demonstrated by the following Indri Query Language snippet:

```
#weight ( 1−λ #combine([original query]) λ
 ( #weight [weight] #combine([unigram 1]) ... ) )
```

6.5 Results and Discussion

I experimented with all PRF conditions and TSQE β parameter settings for each collection and topic set. Statistical significance is measured using the non-parametric paired Wilcoxon Signed-Rank test with the LM baseline. A value of $p < 0.05$ is considered statistically significant, and denoted by * in results.

[1] http://www.lemurproject.org/indri.php

[2] Referred to as "DB0" in Peetz et al. (2014)

Table 6.2: Language model with no PRF (LM), LM with Relevance Model PRF (LM+RMPRF), and LM with Temporal Query Modelling (LM+TQM) baseline performance for each test collection/topic set. * denotes statistical significance ($p < 0.05$). LM+RMPRF and LM+TQM effectiveness measures are accompanied by their percentage change relative to LM.

Collection	MAP	P@10	P@20	P@30
LM				
AP-TREC1	0.275	0.438	0.401	0.383
AP-TREC2	0.233	0.402	0.371	0.335
WSJ-TREC1	0.262	0.45	0.412	0.389
WSJ-TREC2	0.238	0.462	0.387	0.361
MB	0.174	0.437	0.431	0.407
Blogs06	0.3	0.614	0.603	0.596
LM+RMPRF				
AP-TREC1	0.275 (+0%)	0.438 (+0%)	0.401 (+0%)	0.381 (-1%)
AP-TREC2	*0.238 (+2%)	0.400 (+0%)	0.373 (+1%)	*0.347 (+4%)
WSJ-TREC1	*0.262 (+0%)	0.444 (-1%)	0.416 (+1%)	0.389 (+0%)
WSJ-TREC2	*0.241 (+1%)	0.456 (-1%)	0.392 (+1%)	0.36 (+0%)
MB	*0.184 (+6%)	0.463 (+6%)	0.448 (+4%)	*0.422 (+4%)
Blogs06	*0.308 (+3%)	*0.67 (+9%)	0.632 (+5%)	0.609 (+2%)
LM+TQM	**(BASELINE)**			
AP-TREC1	*0.287 (+4%)	*0.457 (+4%)	*0.416 (+4%)	*0.395 (+3%)
AP-TREC2	*0.251 (+8%)	*0.419 (+4%)	*0.395 (+7%)	*0.363 (+8%)
WSJ-TREC1	*0.276 (+5%)	*0.464 (+3%)	*0.437 (+6%)	*0.397 (+2%)
WSJ-TREC2	*0.254 (+7%)	*0.473 (+2%)	*0.411 (+6%)	*0.379 (+5%)
MB	*0.194 (+12%)	*0.483 (+11%)	*0.463 (+7%)	*0.441 (+8%)
Blogs06	*0.375 (+25%)	*0.677 (+10%)	*0.637 (+6%)	*0.614 (+3%)

6.5.1 Retrieval Performance Baselines

I present the LM and LM with Temporal Query Modelling (LM+TQM) experimental baseline in Table 6.2. Additionally, for comparison I report a non-temporal PRF weak baseline – LM with Relevance Model (LM+RMPRF) – to justify the selection of LM+TQM as an appropriate baseline for the remainder of results presented in this section.

LM+RMPRF parameters were selected by splitting the available topic set 50/50 for a 2-fold train/test procedure. The retrieval performance reported is for the parameters which yielded the greatest MAP for both folds. Notably, LM+RMPRF is often unable to significantly improve on LM alone for TREC1 and TREC2 collections. Furthermore, performance increases observed for MB and Blogs06 are often not significant. In comparison, LM+TQM always significantly outperforms LM and LM+RMPRF retrieval performance for all collections and measures. Accordingly, I select LM+TQM as the appropriate state-of-the-art experimental baseline for this work. Hence, the remainder of experimental approach results presented in this section are compared with respect to this baseline.

Notable in the reported results are per-collection variances. Compared to other collections, Blogs06 precision performs relatively well without any PRF. Indeed, LM+TQM is able to considerably improve on the baseline performance. This is likely because the top-k documents for each topic in the Blogs06 are largely relevant, and so PRF has a strong sample of relevant documents from which to extract distinguishing terms – so PRF is more likely to avoid query drift and improve retrieval performance overall. For all AP and WSJ topic sets, performance metrics are much more modestly improved by LM+TQM in contrast to MB and Blogs06. This is likely because AP and WSJ are relatively small clean, and also that LM has been extensively optimised through testing on these older test collections.

6.5.2 TSQE Experimental Results

This section provides preliminary analysis of the best performing runs by examining optimal TSQE β and PRF parameter settings, considering all available topics. Analysis is based upon a single static β parameter for TSQE being set for all topics (i.e., $0 \leq \beta \leq 1$).

In Table 6.3, I select and analyse the top ten (where available) best performing and statistically significant experiment runs for each collection/topic set and measure. In Tables 6.4, 6.5, 6.6 and 6.7, I analyse the β parameter, number of feedback documents, expansion term weight and expansion term selection scheme necessary to achieve the optimal performance of the runs aggregated for each collection/topic set and measure (reported in Table 6.3). Where reported measures and parameter values are aggregated, I also report the standard deviation (±) to quantify the variance of the reported average.

The objective of this section is study the conditions needed by the novel TSQE approach to perform well, and how these differ between varying collections and topics for each metric (i.e., MAP or P@k). Following this preliminary analysis, in Section 6.5.3, I study the reliability of learning the TSQE β and PRF parameters using split train/testing topic sets. Moving beyond the assumption of a single optimal static β for all topics, in Section 6.5.4, I further analyse the outcome of setting the TSQE β parameter on a per-topic basis.

Preliminary Best TSQE Run Analysis

To understand how TSQE can affect retrieval performance, given optimal parameter settings for that particular collection, Table 6.3 presents the results for the best performing single β

Table 6.3: Aggregated *best performing static β TSQE experiment runs* for each measure. Only runs with statistical significance of $p < 0.05$ are used. Less than ten statistically significant runs being available for analysis is indicated by †.

	MAP		P@10		P@20		P@30	
Collection	Avg	±	Avg	±	Avg	±	Avg	±
AP-TREC1	0.316 (+10%)	0.001	0.466†(+2%)	0.003	0.449 (+8%)	0.003	0.426 (+8%)	0.002
AP-TREC2	0.296 (+18%)	0.002	0.419†(0%)	0.041	0.415 (+5%)	0.002	0.388 (+7%)	0.002
WSJ-TREC1	0.302 (+9%)	0.001	0.513 (+10%)	0.005	0.454 (+4%)	0.001	0.42 (+6%)	0.001
WSJ-TREC2	0.287 (+13%)	0.001	0.413 (-13%)	0.005	0.439 (+7%)	0.001	0.406 (+7%)	0.001
MB	0.209 (+7%)	0.001	0.509 (+5%)	0.003	0.488 (+5%)	0.002	0.461 (+4%)	0.001
Blogs06	0.286 (-24%)	0.001	0.648 (-4%)	0.004	0.564 (-11%)	0.002	0.561 (-9%)	0.004

runs (i.e., *static* for all topics) for each collection/topic set and evaluation measure following the PRF and TSQE parameter sweep. There are at least 10 statistically significant runs available for all collection/topic sets and measures, except for P@10 for AP-TREC1 and AP-TREC2.

MAP is substantially increased by 7-18% for AP-TREC2, WSJ-TREC2 and MB over the LM+TQM baseline. For these collections, the majority of precision measures are also enhanced. In contrast, the TSQE approach always considerably harms Blogs06 retrieval performance compared to the LM+TQM baseline. Consequently, these findings demonstrate that with the appropriate parameters TSQE is capable of significantly improving retrieval performance above the state-of-the-art baseline for many collections. Using these aggregated best performing runs as a starting point, in the following sections I provide in-depth analysis of optimal PRF and TSQE parameter settings necessary to maximise retrieval performance. In turn, this provides insight into parameter sensitivity and hence training of these parameters per-collection, and furthermore, per-topic.

Setting the β Parameter to Mix Temporal and Non-temporal Evidence

Table 6.4: Aggregated *TSQE β parameter* used in best experiment runs/measures reported in Table 6.3.

	For MAP		For P@10		For P@20		For P@30	
Collection	Avg	±	Avg	±	Avg	±	Avg	±
AP-TREC1	0.57	0.13	0.45	0.07	0.59	0.19	0.5	0.18
AP-TREC2	0.42	0.1	0.36	0.21	0.36	0.13	0.35	0.11
WSJ-TREC1	0.52	0.18	0.14	0.12	0.24	0.12	0.25	0.1
WSJ-TREC2	0.49	0.12	0.54	0.35	0.51	0.21	0.46	0.16
MB	0.46	0.32	0.4	0.19	0.2	0.1	0.21	0.15
Blogs06	0.16	0.18	0	0	0.45	0.3	0.18	0.1

Table 6.4 reports the optimal β parameter setting mixing temporal and non-temporal evidence for each collection, based on the best performing runs reported in Table 6.3. To achieve optimal MAP, in all collections except Blogs06, a β of around 0.5 (i.e., equally mixing temporal and TF evidence) was required. Although, for MB there was quite a large variance observed, as well as for WSJ-TREC2 and P@10.

MB relied more upon temporal evidence for P@10 and P@20, requiring a much lower β than needed for the best MAP. For AP-TREC1, AP-TREC2 and WSJ-TREC2 the β necessary for best performance was relatively consistent on average across all measures, albeit with occasionally higher variance. Particularly for noisier collections (e.g. MB and Blogs06), to

Table 6.5: Aggregated *number of feedback documents* used in best experiment runs/measures reported in Table 6.3.

	For MAP		For P@10		For P@20		For P@30	
Collection	Avg	±	Avg	±	Avg	±	Avg	±
AP-TREC1	5	0	5	0	5	0	5	0
AP-TREC2	10	0	9	2.2	10	0	10	0
WSJ-TREC1	5	0	5	0	5	0	5	0
WSJ-TREC2	10	0	5.5	1.6	6.5	2.4	7	2.6
MB	15	5.3	27	4.8	19	3.2	20	0
Blogs06	8.5	2.42	6.5	2.42	10	0	10	0

achieve best performance at lower precision levels such as P@20 and P@30 a greater reliance on temporal evidence is necessary, reflected by a lower β. The variance of β showcases per-topic temporality variation, especially inherent in collections such as MB.

Number of Feedback Documents

Table 6.5 reports the number of feedback documents (i.e., the top-k PRF documents) required for best performance, based on the best performing runs reported in Table 6.3. TREC1 topics for both AP and WSJ require only 5 documents, suggesting that early rank precision and document quality is high for the topics in these collections. However, for TREC 2 topics and Blogs06, up to 10 feedback documents were optimum, with the exception of some precision measures performing best with fewer feedback documents thus reducing the addition of noise. MB required far more documents, 15 for MAP and 20 to 30 for precision measures, most likely due to the very limited text in each document.

Expansion Term Weight

Table 6.6 reports the expansion term weight required for optimum performance, based on the best performing runs reported in Table 6.3. For almost all collections, topic sets and measures there is a relatively large variance for the weight. As such, there is no clearly optimal weight, suggesting optimal retrieval performance is not particularly sensitive to it, at least not at the levels with which I experimented.

Term Selection Scheme

Table 6.7 reports the term selection scheme required for optimum performance, based on the best performing runs reported in Table 6.3. I experimented with two schemes, static top N terms (‡) and adaptive threshold-based selection (†). The consensus indicates the adaptive

Table 6.6: Aggregated *expansion term weight* used in best experiment runs/measures reported in Table 6.3.

	For MAP		For P@10		For P@20		For P@30	
Collection	Avg	±	Avg	±	Avg	±	Avg	±
AP-TREC1	0.54	0.05	0.45	0.07	0.59	0.03	0.54	0.07
AP-TREC2	0.58	0.04	0.52	0.08	0.58	0.04	0.54	0.05
WSJ-TREC1	0.53	0.07	0.55	0.07	0.52	0.09	0.51	0.06
WSJ-TREC2	0.55	0.05	0.52	0.09	0.46	0.05	0.45	0.07
MB	0.55	0.05	0.49	0.07	0.54	0.07	0.56	0.05
Blogs06	0.49	0.07	0.47	0.07	0.43	0.05	0.4	0

Table 6.7: Aggregated *expansion term selection scheme* used in best experiment runs/measures reported in Table 6.3. ‡ denotes static top N expansion n-gram selection and † for adaptive (threshold-based) expansion term selection.

	For MAP		For P@10		For P@20		For P@30	
Collection	Avg	±	Avg	±	Avg	±	Avg	±
AP-TREC1	†0.19	0.02	†0.18	0.04	†0.18	0.04	†0.18	0.03
AP-TREC2	†0.17	0.03	†0.2	0	†0.16	0.05	†0.18	0.03
WSJ-TREC1	†0.17	0.03	†0.09	0.06	†0.17	0.03	†0.17	0.03
WSJ-TREC2	†0.18	0.03	†0.06	0.02	†0.18	0.04	†0.17	0.04
MB	‡20	0	‡10	0	‡15	5.27	‡20	0
Blogs06	‡15	5	†0.04	0.01	‡15.56	1.67	‡17.14	3.93

scheme is most effective for all measures for AP and WSJ on both topic sets. For MAP in these collections, a consistent threshold of around 0.2 is best, and for AP, this threshold is also suitable for all precision measures. In comparison, for WSJ, optimal P@10 requires a lower threshold of around 0.1, however P@20 and P@30 return to a threshold of around 0.2.

In contrast, both MB and Blogs06 perform best with a static term selection scheme (except P@10 for Blogs06). MAP and P@20 and P@30 for MB are most improved when 15 to 20 expansion terms are used, whereas P@10 is best with 10 terms. Fewer terms are likely to keep possible query drift to a minimum.

Further analysis of the queries generated by the adaptive threshold scheme provide an insight into the distribution of the ranks assigned to expansion terms by TSQE at $0 \leq \beta \leq 1$. With an adaptive threshold of 0.2 for AP-TREC1, at $\beta = 0, 0.2, 0.4, 0.6, 0.8, 1$ the average number of query expansion terms for each topic was: $72, 77, 65, 47, 33$ and 10 (±17, 20, 19, 16, 13 and 10 expansion terms). A power-law probability distribution becomes increasingly present towards

$\beta = 1$, causing less query expansion terms to be selected. Conversely, emphasising temporal evidence reduces the long-tail of the distribution of TSQE assigned term ranks.

6.5.3 TSQE and PRF Parameter Learning

To confirm the generalizability of learning optimal collection TSQE β and PRF parameters outlined in the previous section, I train/test on different topics and present results in Table 6.8. To this end, I separated the available test topics into training and test sets for each collection and measure (e.g., MAP and P@10/20/30).

As there are two sets of topics for the AP and WSJ collections (i.e., TREC-1 and TREC-2), I trained parameters on the 50 test topics available for TREC-1, and tested them using the 50 topics available for TREC-2, and vice-versa. However, with only 50 test topics available for MB and Blogs06, I split the topics into two sets of 25 topics (denoted as MB/Blogs06-A/B in Table 6.8).

For each training set and measure, the top ten best performing and statistically significant runs were aggregated to determine the best performing parameters, on average. For AP and WSJ, the average parameters are those provided in Tables 6.4-6.6, from the runs aggregated in Table 6.3. For MB/Blogs06-A/B the parameters are mostly similar to those reported for all 50 topics, however because of space limitations they are not included in detail here.

MAP for all collections except Blogs06 is reliably improved between 7-12%, with almost all collections statistically significant. Lack of statistical significance for MB-A may be explained by the small sample size (i.e., only 25 topics). For AP, precision improves by up to 4%, however only with statistical significance for P@20 for TREC-1. Likewise for WSJ, precision also improves by up to 7%, but with statistical significance for P@20/30. Substantial performance improvement is still observed even when the training set has different optimal PRF conditions to the test set, e.g. 5 rather than 10 feedback documents, indicating that TSQE is relatively insensitive to these parameters.

Despite only training on a small set of topics, for MB-A and MB-B all measures are reliably improved. Blogs06 proves problematic with TSQE having a considerable negative effect on almost all measures, as was also observed previously in Table 6.3.

6.5.4 Per-topic β Parameter Oracle

Assuming the average optimal PRF conditions (i.e., feedback documents, expansion term weight and term selection scheme) suggested by the previous parameter learning analysis

Table 6.8: Test performance for each collection/measure after parameter training. * denotes statistical significance ($p < 0.05$).

Collection	MAP	P@10	P@20	P@30
AP-TREC1	*0.309 (+8%)	0.452 (-1%)	*0.432 (+4%)	0.410 (+4%)
AP-TREC2	*0.281 (+12%)	0.404 (-4%)	*0.374 (-5%)	0.370 (+2%)
WSJ-TREC1	*0.300 (+9%)	0.494 (+6%)	*0.453 (+4%)	0.410 (+3%)
WSJ-TREC2	*0.283 (+11%)	*0.450 (-5%)	0.419 (+2%)	*0.405 (+7%)
MB-A	0.218 (+7%)	*0.572 (+8%)	*0.528 (+7%)	0.483 (+5%)
MB-B	*0.188 (+9%)	*0.408 (+6%)	*0.425 (+8%)	*0.394 (+8%)
Blogs06-A	*0.284 (-19%)	0.604 (-3%)	*0.526 (-10%)	*0.569 (-6%)
Blogs06-B	*0.280 (-22%)	*0.560 (+7%)	0.502 (-11%)	*0.515 (-10%)

Table 6.9: Beta oracle performance. Percentage improvement over LM+TQM baseline is reported.

Collection	MAP	P@10	P@20	P@30
AP-TREC1	0.335 (+17%)	0.506 (+11%)	0.493 (+18%)	0.458 (+16%)
AP-TREC2	0.323 (+29%)	0.484 (+15%)	0.468 (+18%)	0.421 (+16%)
WSJ-TREC1	0.322 (+17%)	0.56 (+21%)	0.49 (+12%)	0.449 (+13%)
WSJ-TREC2	0.303 (+19%)	0.518 (+9%)	0.474 (+15%)	0.429 (+13%)
MB	0.226 (+16%)	0.547 (+13%)	0.52 (+12%)	0.488 (+11%)
Blogs06	0.315 (-16%)	0.68 (0%)	0.644 (+1%)	0.632 (+3%)

for each collection/topic set and measure, in Table 6.9, I present the oracle yielding the best performing β per-topic.

Setting β on a per-topic basis greatly increases performance (up to +29% in the case of AP-TREC2 MAP) for all collections/topic sets, except Blogs06 – highlighting a more fundamental issue which I discuss in Section 6.5.5.

In order to visualise the sensitivity of per-topic β parameter setting, in Figure 6.4, I illustrate the AP-TREC1 per-topic effect of β on average precision, with reference to LM non-PRF performance (i.e., the initial top-k retrieval that TSQE is based on). For most topics, TSQE, for either some or all β levels, outperforms the LM baseline. There are relatively few topics where any β harms performance compared to the baseline. In any case, for most of these adversely affected topics, optimal β PRF performance is only marginally below LM performance (e.g. topics 51 and 52). β is therefore shown, in most cases, a relatively insensitive parameter, at least in this instance.

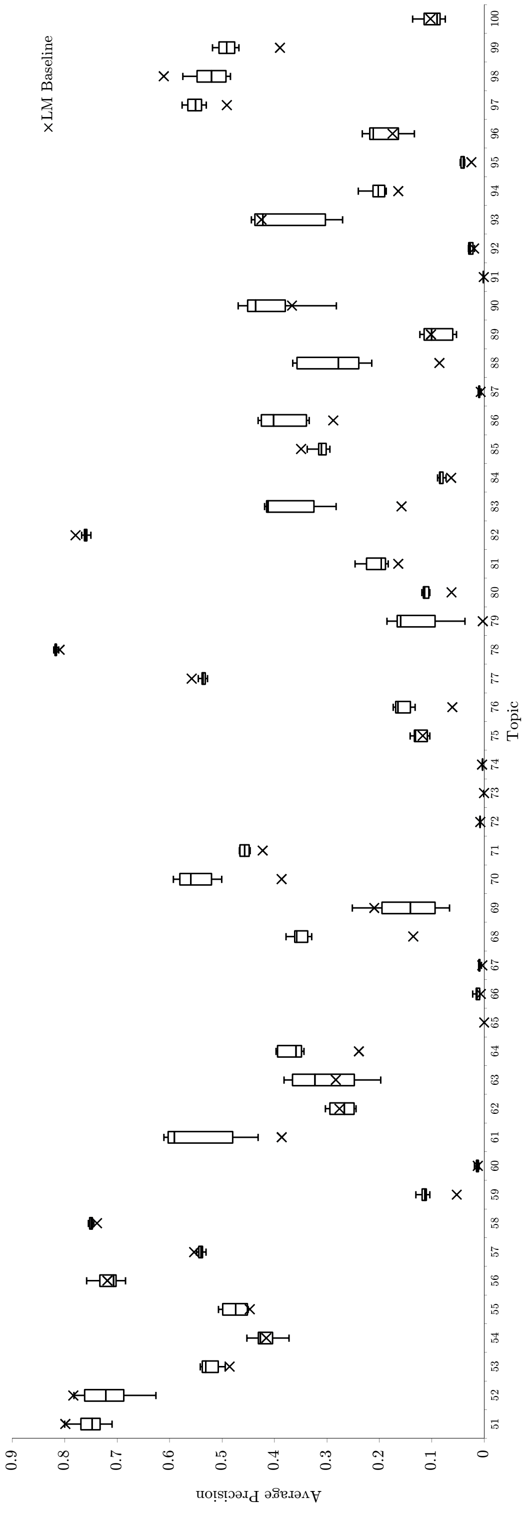


Figure 6.4: Box-plot of per-topic LM average precision (signified as the '×'), and dispersion of the TSQE approach retrieval average precision for AP-TREC1 topics at all β parameters swept from $[0, 1]$.

6.5.5 Discussion

In this section I discuss the results presented in the previous section in relation to the four research sub-questions outlined for this chapter.

RQ3.1: Can temporal semantic QE, combining non-temporal (i.e., term frequency) and temporal (in the form of temporal semantic similarity) evidence improve overall retrieval effectiveness?

I conducted extensive experiments with diverse test collections to determine whether TSQE reliably improves retrieval effectiveness in differing scenarios. As shown in Tables 6.3, 6.8 and 6.9, significant performance improvements can be achieved with the proposed TSQE PRF, which exploits both temporal and non-temporal evidence. There is a single static TSQE β setting mixing both temporal and non-temporal evidence (i.e., $0 < \beta < 1$) that is capable of significantly outperforming, on average, the performance of the LM+TQM for some collections/topic sets and measures, except Blogs06. In all cases, mixing temporal and non-temporal evidence is necessary. As such, neither non-temporal nor temporal evidence alone in PRF are able to outperform, on average, the baseline. Importantly, training/testing of the β and PRF parameters (i.e., Table 6.8) indicates that TSQE PRF can be reliably trained to improve MAP. While some improvement in precision can be attained, it is generally more sensitive and indeed often harmed. I discuss query drift, the likely reason for this effect, in the context of the next research question.

For the Blogs06 collection, significant performance improvement is not observed. I attribute poor PRF performance in Blogs06 to the fact that I was only able to mine the limited syndication feed summaries for term temporal dynamics. As such, the temporal dynamics did not adequately reflect the true temporal nature of index terms contained in the collection over time. This finding emphasises the importance of obtaining representative and accurate temporal dynamics for this application. Additionally, Peetz et al. (2014) highlight the relatively high initial retrieval performance, along with narrow and very well defined bursts observed for the Blogs06 collection/topics. These two factors mean that their proposed TQM approach (which exploits these bursts) provides an especially strong baseline, which TSQE is unable to beat.

RQ3.2: QE approaches can yield erratic per-query performance. How does the proposed QE approach perform for individual queries?

QE, and in particular approaches based on PRF are well known to improve some queries, yet severely harm others (Carpineto and Romano, 2012). This is due to query drift, where the expanded query includes neutral of poor terms. Unfortunately, query drift caused by introducing off-topic terms is a common problem, especially for PRF when the initial retrieval is poor, and thus a poor sample of relevant documents. While QE often leads to overall improvement, the erratic per-query performance fluctuations may be unacceptable for many applications since it can compromise user confidence in the system. It is therefore important to consider the sensitivity of the proposed approach in relation to individual query improvement or degradation to determine the robustness of the TSQE approach.

In Figure 6.4, I examined the per-topic effect of the TSQE β parameter on average precision in relation to the LM initial retrieval baseline. For the majority of topics, TSQE for either some or all β levels outperforms the LM baseline. In most cases, the TSQE approach either does relatively minimal harm, has no effect, or always has a positive effect on retrieval performance. There are relatively few topics where any β harms performance compared to the baseline. In any case, for most of these adversely affected topics, optimal β PRF performance is only marginally below LM performance (e.g., topics 51 and 52). β is therefore shown, in most cases, a relatively robust parameter in this instance. This behaviour of course relies on optimal PRF parameters – so care must be taken to select these effectively for the collection and task objectives, as discussed for RQ3.3 in the next section.

RQ3.3: What PRF conditions and QE approach parameters are necessary to maximise IR system effectiveness in diverse time-based scenarios?

There are no universal PRF parameters able to achieve optimum PRF performance for all collections – a similar finding to past PRF studies (Carpineto and Romano, 2012; Ogilvie et al., 2009). Different collections have varying temporal and non-temporal characteristics. Accordingly, the characteristics of the collection (e.g. document size and noisiness) and in most cases, the desired measure performance (i.e., MAP or precision) are important factors when determining the number of feedback documents and term selection scheme. The weight of the QE candidate terms in the expanded query is however a less critical factor.

Interestingly, despite overall performance gain in the MB collection, for some topics (e.g., MB002: "`2022 FIFA soccer`") seemingly relevant index terms have a very low temporal dynamic kurtosis – and thus, no discernable temporal dynamics (e.g., '`#fifa`', '`#qatar`' and '`world cup`'). This is likely due to the limited collection period, as they are unlikely to have a low kurtosis over a longer period. Accordingly, a longer-term collection may yield better PRF results with this approach.

RQ3.4: If optimal PRF and QE approach parameters can be predicted per-topic, what is the expected impact on retrieval effectiveness?

In many cases, the optimal PRF parameters are not sensitive, and so, can be set per collection. However, it is clear there is substantial performance improvement to be gained by setting the

TSQE β parameter on a per-topic basis. Oracle performance analysis reported in Table 6.9 suggests that individual topics have differing temporal characteristics, leading to a varying optimal β parameter setting per-topic. While some topics have a strong temporal dimension and so benefit from greater use of temporal evidence in PRF, others may instead favour non-temporal evidence in PRF. Query characteristics such as length, difficulty and ambiguity (Carpineto and Romano, 2012) which ultimately affect general retrieval performance are likely to also play a complex underlying role in retrieval performance both before and after PRF.

Although it is left to future work to predict the optimal β parameter per topic, I propose a number of features which may exhibit a relationship with the optimal topic β parameter. These include the distribution of temporal dynamic characteristics (e.g., kurtosis/peakedness) and the distribution of chronotype scores prior to running TSQE at extremes employing only temporal, or non-temporal evidence (i.e., $\beta = 0, 1$). By employing appropriate machine learning algorithms or heuristics, such as simulated annealing on performance estimates, may provide appropriate means of approximating optimal β values, given resource constraints.

6.6 Chapter Conclusions

Time-based collections such as blogs, tweets and news are composed of time-stamped documents covering events and phenomena are becoming increasingly popular. Conventional IR techniques, such as query expansion, typically view these collection as static over time. This belief may lead to sub-optimal retrieval performance when the composition of the collection changes as new documents are added over time, and the underlying statistical distributions (e.g., term and document frequency) become subject to temporal dynamics. In any case, IR approaches that ignore evolution of the collection do not exploit the rich temporal dimension available to characterise the meaning and structure of information during retrieval. In this chapter I posit that exploiting temporal dynamics in QE, based on PRF, can lead to improved retrieval effectiveness in time-based collections. PRF is a method of improving retrieval performance by expanding the original query with distinctive features found in the top-retrieved pseudo-relevant documents. Existing PRF approaches typically rely upon the measure of index term discriminability, or identify highly query-related terms to determine most distinctive terms for feedback via QE. I propose the Temporal Semantic Query Expansion (TSQE) PRF approach, which selects QE terms based on both temporal (i.e., temporal semantic similarity between terms) and non-temporal (i.e., term frequency in PRF documents) evidence. Temporal index term semantic similarity is measured by the correlation between document frequency temporal dynamics of the terms. Selected QE terms consist of terms which are both strongly temporally related to one another, and distinctive in the PRF documents. The proposed approach is practical in large-scale real-time scenarios as it relies upon temporal dynamics that can be readily obtained from time-sharded index statistics.

Experiments on four diverse time-based test collections demonstrate the retrieval improvement offered by TSQE for three of them. In particular, the proposed TSQE PRF model, mixing both temporal and non-temporal evidence, is able to significantly outperform the Language Model with Temporal Query Modelling baseline for many collections and retrieval performance measures. Different retrieval goals (i.e., recall or precision) in each collection require differing PRF and TSQE parameters to achieve optimal performance, demonstrating the varying temporal nature or queries and collections. Overall, I have found incorporating temporal dynamics into QE based on PRF is therefore valuable for improving retrieval performance in time-based collections.

Part IV

Conclusions

Chapter 7

Conclusion and Future Work

In this chapter, I will discuss the findings of this thesis and their impact. I begin by outlining the thesis in terms of the high-level objectives and problems solved. I proceed to discuss the findings contained in each research chapter with respect to the three high-level research questions proposed in Chapter 1. As part of this, I detail the novel contributions made in each chapter. Finally, I make conclusions with respect to the thesis itself, and provide future directions for research in time-aware information retrieval (IR).

7.1 Introduction

In this thesis I have thoroughly investigated the involvement of time, and in particular, temporal dynamics in many aspects of IR. In Part I, I began by introducing and motivating the thesis statement. Following this, I provided an extensive background for IR and conceptualised the role of time in the IR process in relation to both users and IR systems, contributing a conceptual map of time-aware IR.

Since many users turn to IR systems to satisfy time-sensitive information needs, in Part II, I explored how an IR system must support the real-time temporal dynamics of changing user needs and expectations. In this part of the thesis I investigated how to improve the user's satisfaction consistently over time. In particular, I first proposed approaches to explicitly support the user in submitting adequate queries through time-sensitive query auto-completion. Following this, I examined approaches to implicitly support the user through temporal relevance ranking.

As well as supporting real-time temporal dynamics, in Part III, I considered how temporal dynamics captured in past information behaviour can be exploited to inform more effective IR systems. In this part of the thesis I investigated how evidence from past temporal dynamics can be integrated into IR models. With this in mind, I proposed query expansion based on the temporal semantics of index terms to improve retrieval effectiveness in time-based collections.

7.2 Discussion and Contributions

The main aim of this thesis was to study the involvement of temporal dynamics in information and information behaviour. In this section, I discuss the significance of the work presented in advancing the field of time-aware IR. I begin by outlining the important theoretical contribution made by this thesis – a conceptual map of time-aware IR, outlining the broad involvement of time in user- and system-oriented aspects of IR. Following this, for each high-level research question relating to a temporal dynamic evident in IR (i.e., RQ1 to 3, outlined in Chapter 1), I discuss the work, findings and subsequent novel contributions of this thesis.

7.2.1 Conceptual Map of Time-aware IR

Time has a very broad definition, and consequently, many implications on information and information behaviour. Two previous reviews (Alonso et al., 2011; Campos et al., 2014) provide extensive literature surveys of time-aware IR approaches. However, neither of these past works offer a principled model within which to organise and reason time-aware IR approaches, which either exploit or support the many implicit and explicit temporal clues present in IR. Since the influence of time in IR is multi-faceted, and the clues provided by information behaviour are broad, I argue such a model is necessary to understand diverse existing works and highlight future research opportunities. In Chapter 3, in addition to the three high-level research questions, I present a map of time-aware IR (i.e., Figure 3.1 on page 27), informed by established conceptual IR models.

I recognise the two-sided involvement of both the user and IR system during the process of IR. Consequently, I focus on the time-based aspects involved in six elements of the widely recognised conceptual IR model, namely: (1) user context, (2) cognitive factors, (3) information need formulation, (4) query formulation, (5) matching/relevance, and (6) the information collection. In particular, I focus on the following established time-aware IR areas: (i) temporal information need identification, (ii) temporal query modelling, (iii) temporal relevance modelling, and (iv) temporal information modelling. Alongside this proposed model I frame and map the extensive research across time-aware IR. The conceptual model serves to structure the work based on its objectives, and use of implicit and explicit temporal clues. Many opportunities in time-aware IR research are illuminated by this conceptual map – these are detailed as future work in Section 7.4.

Within the context of this map of time-aware IR, in the following sections I discuss the three high-level research questions of this thesis. For each research question I highlight the main findings and contributions provided by this thesis. The first two of these high-level research questions relate to *supporting* temporal dynamics in real-time information seeking. In the map of time-aware IR, these temporal dynamics arise from the implicit temporal clues attached to the “Query” and “Matching (Relevance Model)” aspects.

7.2.2 RQ1: Supporting Temporal Dynamics in Query Popularity

In this section I discuss the findings presented in Chapter 4 with respect to **RQ1** of this thesis: “Query popularity exhibits many patterns and trends over time as common information needs change. *Can these temporal dynamics be supported for consistently effective query auto-completion over time?*”

Query popularity varies dramatically over time as real-world events and phenomena influence information seeking. In Chapter 3, I explored several temporal dynamics present in query popularity activity, including periodic trends and one-off spiking patterns, etc., (Kulkarni et al., 2011). Indeed, these temporal dynamics may, or may not be predictable based on past observations.

One of the foremost the challenges for users during retrieval is formulating an adequate query to express their information need sufficiently to the IR system. A descriptive and error-free query maximises the chance of receiving relevant search results, and ultimately being satisfied. Query auto-completion (QAC) aims to assist users in supplying effective queries. However, consistently supporting users who are engaging in information seeking in real-time with effective QAC requires approaches which can anticipate changes in the queries submitted over time. I considered the trade-off between recent (i.e., most time-sensitive) and robust QAC. My objective was to develop effective QAC approaches which can provide both consistently popular and recently popular query completion suggestions alongside one another, when necessary.

I first outlined two existing common approaches to QAC, which are based on the most popular queries observed previously, and in recent periods of time (i.e., MLE-ALL and MLE-W). I proposed a new QAC approach based on tracking the last N queries with a given prefix (i.e., LNQ). I extended this approach to anticipate upcoming query popularity changes using query count prediction based on an incrementally-learnt short-range linear regression model (i.e., PNQ). Finally, I proposed an online parameter learning meta-approach to determine the optimal parameters of the aforementioned QAC approaches in a typical online QAC setting.

Using three real query logs in both English and Chinese language, I simulated a realistic real-time QAC system and experimented with each QAC approach for query prefix lengths of 2 to 5 characters over tens of millions of queries. The diverse temporal, language and demographic characteristics of each query log allowed me to compare approach performance in differing real-world scenarios. Noteworthy is the fact that the nature of temporal dynamics vary between scenarios, demonstrating while time is ubiquitous in all scenarios, it may hold differing implications depending on the user demographics and corresponding common usage patterns present. Hence, building an effective time-aware QAC system requires careful selection of approach, parameters and query popularity predictive model to suit the intrinsic temporal, language and demographic characteristics of the intended scenario. That is, I found no single parameter combination that is assured to be consistently effective in all situations.

Overall, supporting query popularity temporal dynamics is central to the problem of QAC. Accordingly, failing to take temporal dynamics into account leads to under-performing QAC. Results presented in Table 4.7 on page 85, show that incorporating the temporal dynamics of query popularity using relatively straight-forward online parameter learning techniques leads to the best observed QAC performance (i.e., +7.2% and +9.2% over the soft baseline for MSN and Sogou, for query prefixes of 2 characters). This represents reproducible state-of-the-art QAC performance in these cases – with a novel but practical time-aware QAC approach. Employing a more elaborate recent-trend linear regression model to rank completion suggestions based on short-range predicted query counts (i.e., PNQ) offers little improvement over an appropriately selected LNQ model. Hence, in this case the simplest but appropriately parameter tuned approach has been shown to outperform more elaborate approaches for time-aware QAC. The findings presented in this thesis therefore offer a foundation for future work in time-aware QAC.

It is worth noting that for the sake of practical online learning, I assumed query popularity could be predicted using a linear model. Of course this may not be the case, especially for certain types of queries that exhibit extreme or unusual burst patterns (e.g., the most major of news events) (Kulkarni et al., 2011), which in turn will not be adequately modelled by a linear model. As such, future work is likely to find more elaborate non-linear recent-trend regression models offer more gains for such cases.

Finally, one fundamental limitation is the relationship between the evaluation metric (i.e., MRR) used for evaluate this work, and real user satisfaction. To facilitate systematic evaluation, I rely on MRR as the metric to measure improvement in QAC approach performance. However, an increase in MRR may not necessarily correlate with a significant increase in user satisfaction when actually using the QAC system. Accordingly, further work is needed to better understand user models and satisfaction in QAC.

Contributions

Addressing **RQ1**, this thesis makes several novel contributions relating to time-aware query auto-completion to improve completion suggestion ranking over time:

1. Novel approaches to QAC based on past query distributions, and optionally incorporating short-range query popularity prediction. These approaches are designed to be computationally tractable and practical, given the constraints imposed by real-time QAC systems.

2. Extensive empirical analysis of QAC approach performance based on three real-world query logs (two English, and one Chinese) for varying prefix lengths.

3. Insight into the QAC approaches and parameters necessary to obtain optimal (i.e., *state-of-the-art*) time-aware QAC effectiveness in diverse scenarios (e.g., differing language and temporal characteristics).

4. Fully reproducible experimental findings due to the open query log datasets and experimental tools used[1].

7.2.3 RQ2: Supporting Temporal Dynamics in Query Intent

In this section I discuss the findings presented in Chapter 5 with respect to **RQ2** of this thesis: “Query intent exhibits many patterns and trends over time as collective influences and expectations change. *To what extent do query intents change over time, and, can these temporal dynamics be supported?*”

The user’s motivation – that is, intent – behind many queries exhibits several temporal dynamics as time influences *what* users a looking for, and *why*. In this thesis I have explored a number of temporal dynamics apparent in query intent. In Chapter 3, I explored periodic trends in query intent caused by repetitive and therefore predictable temporal events affecting what users are interested in over time, such as different types of party planning at different times of the year, e.g., Christmas, Summer or Halloween party planning. Further examples in Chapter 5 illustrate how unforeseeable and often short-lived events also have considerable temporal impact on query intent. These temporal dynamics need to be taken into account

[1] This has already allowed work presented in this thesis to be extended in Cai et al. (2014).

by relevance models looking to support users in their information seeking consistently over time.

The primary objective of an IR system is to identify information relevant to the user's information need. However, short and non-specific queries provided by users are a common problem in web search. An estimated 16% queries are ambiguous (i.e., they have one expected interpretation) (Song et al., 2007), and 25% are multi-faceted (i.e., they have multiple complementary interpretations, of which one or a few may be most desired) (Jansen et al., 2007). In both cases, these types of query lead to uncertainty about the user's information need, and hence, what information is relevant. In response to this problem, intent-aware search result diversification approaches interleave search results covering the most popular interpretations (that is, query intents) to satisfy as many users as possible. For intent-aware diversification to be successful, the search engine must have knowledge of: (i) possible query intents, and (ii) the popularity of each of these intents. Temporal dynamics in query intent can be viewed as both the emergence and evolution of query intents, and changing popularity of those intents over time. Hence, I studied the temporal dynamics of query intent in the context of both ambiguous and multi-faceted queries as a first step of framing temporal relevance as time-sensitive search result diversification. Since uncertain queries account for a considerable proportion of daily query volume, supporting their temporal dynamics has potentially high impact for improving user satisfaction over time.

Firstly, in Section 5.4 on page 107, I showed temporal dynamics play a central role in the intent popularity of many ambiguous queries. The distribution of ambiguous topics with temporal intent ranking changes shown in Figures 5.6, 5.7, 5.8 and 5.9 on pages 113-115 demonstrate that optimal intent-aware result ranking is far from stationary for the overwhelming majority of ambiguous queries. Indeed, ranking changes occur frequently over hours, days and months in many cases. Interestingly, I observed differing types of ambiguous queries exhibit varying temporal dynamics over several periods of time. For example, optimal intent-aware ranking for person, place and film entities changes hourly throughout the day, in contrast to short ambiguous acronyms which change daily instead. Neither consistent periodicity nor occasional event-driven ranking changes are exclusive. In fact, many distinct temporal dynamics interact to produce complex temporal ranking effects. Modelling these compound temporal effects is a considerable challenge. Established time series modelling approaches (Radinsky et al., 2013b) will likely need to be adapted to provide truly time-sensitive ranking, which is proactive to the changing influences and expectations of web search users – which is often not predictable solely from past popularity evidence. Overall, the findings of this work demonstrate temporal dynamics have considerable impact on a large proportion ambiguous queries, over several time periods. Accordingly, these findings hold several implications for developing temporal relevance approaches to support real-time information seeking behaviour, including: (1) the need to support varying temporal dynamics for intrinsically different types of ambiguous queries over hours, days and months, (2) past query intent popularity evidence is important, but cannot be considered in isolation, and (3) external sources may offer insight for anticipating upcoming ranking changes caused by transient and unpredictable event-driven influences. An important point to bear in mind is that these findings are based on surrogate temporal dynamics data from Wikipedia. Based on the findings of several other works I argued these temporal dynamics are representative of what is found in a query log, however, future work with long-term proprietary query logs is necessary to further verify the generalisability of findings made by this work.

Moreover, in Section 5.5 on page 116, I found temporal dynamics play a central role in the query intent popularity of event-driven multi-faceted queries – which are account for a considerable proportion of daily trending web search queries (Kairam et al., 2013). Without knowledge of the emerging query intents, and their temporal popularity, applying effective intent-aware query ranking is problematic. Reducing the reliance on past query logs to mine query intents, I found related content structure and dynamics can instead be used to derive multi-faceted query intents for event-driven topics (e.g., long-running news events). In particular, results presented in Table 5.2 on page 123 show the majority of event-driven query intents can be represented by hierarchical Wikipedia article structure, namely user-curated sections and subsections. Moreover, results shown in Table 5.3 on page 123 demonstrate the popularity of each of these intents over time is equally reflected by the editing activity of informational content in the respective section(s). Consequently, Wikipedia article structure offers a means to understand (i) the query intents currently present and emerging, and (ii) the temporal popularity of each intent. Overall, these findings will underpin new methods to support intent-aware search result diversification in scenarios when there is insufficient, or now outdated past query log evidence to adequately rank query intents in real-time.

Although I found representing event-driven query intents with Wikipedia is not perfect in terms of full intent coverage and precisely measuring intent popularity, future work combining external signals of query intents (i.e., from Wikipedia articles, or social sharing, etc.) with internal query log interaction signals (e.g., available result clicks) is likely to lead to more effective intent-aware search result diversification for real-time event-driven queries.

Contributions

Addressing **RQ2**, this thesis makes several novel contributions relating to framing temporal relevance as time-sensitive search result diversification to improve search ranking satisfaction over time:

1. An in-depth study into the impact of temporal dynamics ambiguous queries in terms of periodic, and event-driven occasional trends on query intent over varying periods of time.

2. I characterised the impact of temporal dynamics on query intent for short- and long-running (e.g., days to months long) event-driven multi-faceted queries.

3. I demonstrated that Wikipedia article structure strongly reflects query intents of multi-faceted event driven queries, allowing it to be used for real-time query intent modelling. Further, I showed the stream of collaborative changes to this structured content is indicative of user interest in the particular aspect, and thus highly correlated with the respective query intent popularity.

4. A robust methodology for quantifying temporal ranking changes in order to study the effect of temporal dynamics on past result ranking. Although I studied web search rankings, the proposed method is generalisable and so can be applied to other rankings, such as completion suggestions in query auto-completion.

5. Finally, a secondary contribution of this work is a set of guidelines for using Wikipedia as a extensive and diverse source of evidence for both temporal dynamics and temporal structured knowledge in time-aware IR research. This includes an overview of findings from several studies which support using temporal dynamics sourced from Wikipedia as a valid surrogate to those found in suitably large-scale long-term query logs, which are unavailable for open research.

The remaining two high-level research questions this thesis relate to *exploiting* past temporal dynamics captured in information collections. In the map of time-aware IR proposed in Chapter 3, these temporal dynamics arise from the implicit temporal clues attached to the "Collection" aspect.

7.2.4 RQ3: Exploiting Temporal Dynamics in Term Popularity to Improve IR System Effectiveness

In this section I discuss the findings presented in Chapter 6 with respect to **RQ3** of this thesis: "Term popularity exhibits many patterns and trends over time in a time-based document collection. *Can these temporal dynamics be exploited to improve IR system effectiveness over time?*"

Information collections capture various temporal dynamics in past information behaviour as content authors discuss topical events and phenomena over time. In Chapter 3, I explored several temporal dynamics present in time-based collections, including term popularity (e.g., term and document frequency), specificity and relationships.

Although time-based collections – such as ever-evolving web, news and social media – are becoming increasingly common, conventional IR techniques typically view the collection as stationary over time. This assumption may lead to sub-optimal retrieval performance when the composition of the collection changes as new documents covering both old and new topics are added over time, and the underlying statistical distributions (e.g., term and document frequency) used in retrieval models become subject to temporal dynamics. In any case, IR approaches which ignore collection evolution do not exploit the rich temporal dimension available to characterise the meaning and structure of information beyond static distributions during retrieval.

Past work showed the temporal dynamics of term popularity in a collection can be used to estimate semantic relationships (i.e., temporal semantic similarity) between index terms over time (Radinsky et al., 2011). However, this finding had not been previously operationalised to improve IR system effectiveness in time-based collections. Consequently, I set out to exploit these temporal dynamics to improve IR system effectiveness through query expansion (QE), based on pseudo-relevance feedback (PRF).

PRF is a method of improving retrieval performance by expanding the original query with distinctive features found in the top-retrieved pseudo-relevant documents. Existing PRF approaches typically rely upon the measures of index term discriminability to identify terms highly related to the query topic, in order to determine most distinctive terms for feedback via QE. I proposed the Temporal Semantic Query Expansion (TSQE) PRF approach, which selects QE terms based on both temporal (i.e., temporal semantic similarity between terms) and non-temporal (i.e., term frequency in PRF documents) evidence. Temporal index term semantic similarity is measured by the Pearson correlation between document frequency temporal dynamics of the terms. Terms selected for QE consist of terms which are strongly temporally interdependent on one another (i.e., there is a high degree of temporal semantic similarity amongst them), and also at the same time, are distinctive in PRF documents. I name these the topic's "chronotype" terms, and hence posit they are the best terms to exploit for QE in a time-based collection. Note that while this approach is computationally demanding, I proposed various methods to make it more tractable. That said, future work will need to explore further techniques and trade-offs for reducing the modelling complexity necessary (in particular, the size of the dense Temporal Semantic Network graph) to make the approach more practicable in an online system, yet still maximising retrieval performance. This is likely to include better candidate QE term selection, for example, focusing on more definitive entities rather than simply all terms found feedback documents, regardless of their importance.

I conducted extensive retrieval experiments on four diverse time-based test collections (i.e., two news collections, Twitter microblogs and web blogs) to demonstrate the retrieval improvement provided by the TSQE approach in differing scenarios. Based on results presented in Table 6.8 on page 155, I found the proposed TSQE approach, mixing both temporal and non-temporal evidence, is able to significantly outperform a Language Model with Temporal Query Modelling PRF baseline for many collections and measures, particularly for the MAP retrieval performance metric. Accordingly, this increased MAP shows TSQE is able to higher rank relevant results the majority of cases. In some cases TSQE did not outperform the baseline, or only provided statistically insignificant improvement, however there are likely inconsequential factors causing this effect. In particular, the Blogs06 collection proved problematic, likely due to its unreliable composition and associated temporal evidence issues. Retrieval precision metrics were less reliably improved by TSQE; precision is often adversely affected by PRF models, which often lead to query topic drift if the initial retrieval is poor (Carpineto and Romano, 2012). Additionally, the lack of statistical significance in some cases may be down the relatively small test topic sets available for some collections following train/test splits – for example only 25 queries for the TREC Microblog collection[1]. Importantly, analysis presented in Table 6.3 on page 150 showed different retrieval goals (i.e., recall or precision) in diverse time-based collections require varying PRF and TSQE parameters to achieve optimal performance, demonstrating the varying temporal nature of both queries and collections. However, through train-test validation (i.e., Table 6.8) I demonstrated that these parameters could be effectively learnt for each collection and topic set. As a result, overall I have shown incorporating temporal dynamics into QE (based on PRF) mixing both temporal (i.e., temporal dynamics) and non-temporal evidence is beneficial for improving IR system retrieval effectiveness for time-based collections.

Contributions

Addressing **RQ3**, this thesis makes several novel contributions for improving retrieval performance in time-based collections, based on exploiting temporal dynamics in term popularity for QE through PRF:

1. A novel network-based structure, named a Temporal Semantic Network (TSN) to capture temporal and non-temporal evidence contained in PRF documents and index term temporal dynamics.

2. A novel approach for temporal semantic query expansion flexibly combining temporal and non-temporal evidence, based on network analysis of the TSN to identify a query

[1] An additional set of 50 query topics for the TREC Microblog 2011 collection has become available since this work. These further topics are expected to lead to statistically significant results in future work.

topic's chronotype terms.

3. Comprehensive experimental validation and analysis of the proposed temporal semantic QE approach, and related PRF parameters for four diverse time-based collections.

4. An in-depth understanding of the training and parameters needed for TSQE to beat state-of-the-art temporal PRF baseline performance for different retrieval metrics (i.e., MAP versus precision) in diverse time-based collections.

7.3 Conclusion

Time is a deep-rooted and systemic part of individual and social human behaviour. This leads to time-based events and phenomena influencing information interaction, interpretation and expectations in many ways. Accordingly, elements of time are woven throughout information itself, and information behaviours such as creation, seeking and utilisation.

Many patterns and trends – namely *temporal dynamics* – are evident in many aspects of IR, such as what users seek, and the meaning of information over time. A temporal dynamic can refer be a periodic regularity, or, a one-off or irregular past, present (i.e. recent/ongoing) or future – driven by predictable and unpredictable events and phenomena. Despite the extensive involvement of temporal dynamics in IR, the majority of conventional IR approaches do not consider the impact of change over time on their approach or effectiveness. Instead, they view the world and information as stationary over time. As a result, they have neither sought to support nor exploit the inherently temporal dimensions of information and interaction.

This thesis posited that temporal dynamics embody tacit meaning and structure of information and information seeking. I observed temporal dynamics in many of the fundamental elements of IR, including queries and query intent in information seeking, as well as index term frequency, specificity and relationships in information collections. From several perspectives, in this thesis I have demonstrated that temporal dynamics offer rich insight into the tacit temporal structure of many aspects of information behaviour relevant to IR, which can in turn be used to improve IR system satisfaction consistently over time. Overall, since all these elements are main stays of IR approaches, temporal dynamics should be viewed as a fundamental cornerstone of IR.

I conjectured that temporal dynamics are a 'two-way street' because they present both challenges and opportunities for existing IR approaches. Real-time temporal dynamics of user expectations must be supported while users are engaging with IR systems during their daily life. If this is achieved effectively, then users will continue to use IR systems during their real-time tasks as they are consistently satisfied over time. Indeed, the role of temporal dynamics in IR is set to get increasingly important as IR systems become even more central to our real-time dependence on information systems to support our everyday activities. At the same time, past temporal dynamics provides a dimension to structure the meaning of information, beyond the conventional characterisation of information afforded by stationary statistical distributions which assume information and information behaviour remain static over time. In long-term time-based information collections, distinguishing the temporal dimension is likely to become increasingly important in characterising the meaning of information. Consequently, supporting and exploiting temporal dynamics can be viewed complementarily as '*two sides of the same coin*'. That is, supporting real-time temporal dynamics requires exploiting past temporal dynamics, and vice-versa for consistent effectiveness over time.

Overall, in this thesis I have presented and evaluated several techniques which (i) support the real-time temporal dynamics of information seeking to maintain consistent user satisfaction, and (ii) exploit past temporal dynamics as a tacit dimension to inform more effective IR systems. Connected strands of time are diverse in IR, but temporal dynamics are a common thread in many streams of individual and social information and information behaviour. Bringing these strands together with a unified understanding of why users develop information needs, what motivates their need and thus their expectations, and the temporal signals defining the meaning of relevant information over time will lead to considerable future impact by providing more consistently effective IR systems. While a unified model of time in IR is some way off, together the novel contributions both reinforce and establish several novel avenues for incorporating temporal dynamics in time-aware IR, and therefore provide a strong foundation to support information behaviour in present and future IR systems.

7.4 Future Directions

This thesis has explored many challenges are opportunities of temporal dynamics in IR. Several avenues have emerged for future work. In the conclusions provided at the end of each research contribution chapter (i.e., Chapters 4 to 6, inclusive), I included future work specific to the approaches and techniques explored in the respective chapter. In the following sections I detail more general opportunities relating to temporal dynamics in time-aware IR.

Understanding the Impact of Time-Evolving Collections on Retrieval Models

In this thesis I have shown that index term frequency, specificity and relationships all exhibit temporal dynamics in time-based collections which evolve over time. Despite the majority of collections evolving over time, there has been little work to study the impact of changing collections on retrieval model performance. The standard systems-based evaluation methodology (i.e., much of TREC, with the exception of TREC Microblogging), assumes a static collection. Consequently, many retrieval models and techniques (e.g., query expansion) have been developed and tested using static collection snapshots. Given that the fundamental statistical distributions upon which many of these approaches are based change over time, their transferability to real scenarios where temporal change is ever present needs more investigation. Indeed, retrieval effectiveness is likely subject to temporal dynamics. Further work should examine how retrieval performance varies as the collections grows, and new and evolving topics impact retrieval model effectiveness for previously satisfactory results.

Enhanced Time Series Modelling for Event-aware IR

Time-aware IR is often reliant on predicting the occurrence and impact of future events and phenomena. Accordingly, time series models have been used extensively to predict future time-based activity, based on modelling short- and long-term trends, periodicity and surprises in previously observed activity (Radinsky et al., 2013b). In particular, in IR times series models have been employed to predict future query popularity (Shokouhi, 2013), result clicks (Radinsky et al., 2013b) and index term weighting (Efron, 2010).

Conventional time series models such as auto-regressive moving averages (e.g., ARMA and ARIMA techniques) and exponential smoothing are based purely on past time series observations. Hence, they do not take external factors outwith the previously observed time series into consideration. However, structured and unstructured external factors, such as topics of discussion in a social network, and their authors influence, could prove very predictive of a given time series such as the popularity of a related query. For example, discussion about a particular actor in an upcoming highly anticipated television show may be highly predictive of upcoming query popularity for that actor, and their related entities.

Recent advances in machine learning techniques, in particular deep artificial neural networks, facilitate representing and learning vast and deep feature spaces for time series modelling (Boulanger-Lewandowski, 2014). Indeed, these new large-scale knowledge representation techniques may open up many opportunities for powerful time series modelling based on rich sets of previously observed time series trends and patterns, in addition to external factor features (e.g., high dimensional term occurrence during information cascades in a social network). The potential prediction accuracy afforded by these techniques is likely to have considerable impact on time-aware IR techniques which can better anticipate upcoming change. With more reliable knowledge of the future, there will be an opportunity to pursue *event-aware* IR which pro-actively responds to upcoming changes expected to affect user expectations. This is in contrast to current time-aware IR approaches which are typically more reactive as they evolve based on changes seen in previous observations. Indeed, event-aware

IR approaches may be able to find theoretical foundations in human perception of events over time, as discussed in user-oriented factors in Chapter 3.

Further Development of Temporal Relevance

As part of RQ2, in Chapter 5, I outlined the problem of temporal relevance as the variability of information item usefulness for satisfying a particular information need, given external temporal factors such as the current time or recent or upcoming events which may dictate the motivation and thus expectations of the user. Importantly, I framed this problem as time-sensitive search result diversification within an existing intent-aware ranking framework. I found that supporting this view of temporal relevance would have potential for considerable impact in improving time-aware IR. However, since modelling temporal relevance in an intent-aware framework is a complex problem consisting of many modelling and evaluation challenges, the findings I presented were only the first steps of solving the larger problem of temporal relevance. Future work should look to the enhanced query intent popularity modelling, most likely using the time series modelling techniques I outlined in the previous section as a means to model the complex interactions between query intent popularities over time. Furthermore, it is difficult to evaluate temporal relevance reliably as it is inherently an online problem, for which an appropriately long-term temporal test collection does not currently exist. Accordingly, temporal relevance approaches need to be validated and evaluated in a real scenario, most likely through some form of online user testing (e.g., A/B testing).

7.5 Chapter Overview

Time is undoubtedly a cornerstone of IR, from both systems- and user-oriented perspectives. Elements of time are woven throughout information and information interaction. Temporal dynamics are an implicit temporal clue evident in many streams of information behaviour. Indeed, this thesis concludes that *temporal dynamics play a valuable role in many aspects of IR*. Consequently, temporal dynamics should be both supported and exploited in time-aware IR. The main theoretical and practical contributions of the thesis were presented and discussed with respect to the four high-level research questions of this thesis (i.e., RQ1-3). Finally, avenues for future work were proposed.

Part V

References and Appendix

References

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., and Steggles, P. (1999). Towards a better understanding of context and context-awareness. In *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, HUC '99, pages 304–307, London, UK. Springer-Verlag. 3, 29

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using twitter data. In *Information Communications Workshop, 2011 IEEE Conference on Computer Communications*, pages 702–707. 53

Adar, E., Teevan, J., Dumais, S. T., and Elsas, J. L. (2009). The web changes everything: Understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 282–291, New York, NY, USA. ACM. 46

Adar, E., Weld, D. S., Bershad, B. N., and Gribble, S. S. (2007). Why we search: Visualizing and predicting user behavior. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 161–170, New York, NY, USA. ACM. 4, 6, 30, 37, 59, 61, 74, 87, 94, 117

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM. 92

Alfonseca, E., Ciaramita, M., and Hall, K. (2009). Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1046–1055, Stroudsburg, PA, USA. Association for Computational Linguistics. 7, 38, 130

Allan, J. (2002). *Introduction to topic detection and tracking*, pages 1–16. Kluwer Academic Publishers, Norwell, MA, USA. 46, 131

Allan, J., Papka, R., and Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 37–45, New York, NY, USA. ACM. 24, 46

Alonso, O., Gertz, M., and Baeza-Yates, R. (2009). Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 97–106, New York, NY, USA. ACM. 52

Alonso, O., Stroetgen, J., Baeza-Yates, R., and Gertz, M. (2011). Temporal information retrieval: Challenges and opportunities. In *Temporal Web Analytics Workshop*, WSDM '11, pages 59–66, Hyderabad, India. ACM. 4, 7, 25, 26, 163

Arampatzis, A. (2001). *Adaptive and Temporally-dependent Document Filtering*. PhD thesis, University of Nijmegen. 52

Arguello, J., Diaz, F., Callan, J., and Crespo, J.-F. (2009). Sources of evidence for vertical selection. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 315–322, New York, NY, USA. ACM. 98

Aslam, J., Ekstrand-Abueg, M., Pavlu, V., Diaz, F., and Sakai, T. (2013). Trec 2013 temporal summarization. In *Proceedings of the 22nd Text Retrieval Conference (TREC), November*. 52

Baeza-Yates, R., Gionis, A., Junqueira, F., Murdock, V., Plachouras, V., and Silvestri, F. (2007). The impact of caching on search engines. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 183–190, New York, NY, USA. ACM. 33, 52

Bar-Yossef, Z. and Kraus, N. (2011). Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 107–116, New York, NY, USA. ACM. 58, 59, 60, 63, 64, 65, 75, 76, 77, 79, 87

Bast, H. and Celikik, M. (2011). Fast construction of the hyb index. *ACM Trans. Inf. Syst.*, 29(3):16:1–16:33. 63

Bates, M. J. (2010). Information behavior. In Bates, M. J. and Maack, M. N., editors, *Encyclopedia of Library and Information Sciences*, volume 3, pages 2381–2391, New York. CRC Press. 4

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O., and Grossman, D. (2007). Temporal analysis of a very large topically categorized web query log. *J. Am. Soc. Inf. Sci. Technol.*, 58(2):166–178. 37, 75

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., and Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 321–328, New York, NY, USA. ACM. 37

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, (5):133–143. 3, 34

Berberich, K. and Bedathur, S. (2013). Temporal diversification of search results. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA2013)*, Dublin, Ireland. 41

Berberich, K., Bedathur, S., Alonso, O., and Weikum, G. (2010). A language modeling approach for temporal information needs. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, ECIR'2010, pages 13–25, Berlin, Heidelberg. Springer-Verlag. 34, 51

Bhatia, S., Majumdar, D., and Mitra, P. (2011). Query suggestions in the absence of query logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 795–804, New York, NY, USA. ACM. 64

Blanco, R. and Lioma, C. (2012). Graph-based term weighting for information retrieval. *Inf. Retr.*, 15(1):54–92. 133

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM. 45

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022. 45

Borlund, P. (2003). The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Inf. Res.*, 8(3). 13, 20

Boulanger-Lewandowski, N. (2014). *Modeling High-Dimensional Audio Sequences with Recurrent Neural Networks*. PhD thesis, Université de Montréal. 174

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10. 6

Brucato, M., Derczynski, L., Llorens, H., Bontcheva, K., and Jensen, C. S. (2013). Recognising and interpreting named temporal expressions. In *Recent Advances in Natural Language Processing, RANLP 2013*, pages 113–121. 25, 38

Cai, F., Liang, S., and de Rijke, M. (2014). Time-sensitive personalized query autocompletion. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1599–1608, New York, NY, USA. ACM. 166

Campbell, I. (2000). *The Ostensive Model of Developing Information-Needs*. PhD thesis. 29

Campos, R., Dias, G., Jorge, A. M., and Jatowt, A. (2014). Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15:1–15:41. 7, 26, 34, 163

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM. 98

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1. 18, 129, 132, 146, 157, 158, 159, 171

Catledge, L. D. and Pitkow, J. E. (1995). Characterizing browsing strategies in the world-wide web. In *Proceedings of the Third International World-Wide Web conference on Technology, tools and applications*, pages 1065–1073, New York, NY, USA. Elsevier North-Holland, Inc. 73

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623 – 1640. 52

Chang, Y., Yamada, M., Ortega, A., and Liu, Y. (2014). Ups and downs in buzzes: Life cycle modeling for temporal pattern discovery. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 749–754. 62

Chau, M., Lu, Y., Fang, X., and Yang, C. C. (2009). Characteristics of character usage in chinese web searching. *Information Processing & Management*, 45(1):115–130. 75

Chaudhuri, S. and Kaushik, R. (2009). Extending autocompletion to tolerate errors. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, SIGMOD '09, pages 707–718, New York, NY, USA. ACM. 63

Chien, S. and Immorlica, N. (2005). Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 2–11, New York, NY, USA. ACM. 38, 130, 137

Cho, J. and Garcia-Molina, H. (2000). The evolution of the web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 52

Choi, J. and Croft, W. B. (2012). Temporal models for microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2491–2494, New York, NY, USA. ACM. 45, 131

Collins-Thompson, K. and Callan, J. (2005). Query expansion using random walk models. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 704–711, New York, NY, USA. ACM. 133

Craswell, N., Jones, R., Dupret, G., and Viegas, E. (2009). Wscd '09: Proceedings of the 2009 workshop on web search click data. New York, NY, USA. ACM. 73

Croft, W., Metzler, D., and Strohmann, T. (2010). *Search Engines: Information Retrieval in Practice*. Pearson, London, England. 15, 16, 17, 19, 28, 77

Da, J. (2004). A corpus-based study of character and bigram frequencies in chinese e-texts and its implications for chinese language instruction. In *4th International Conference on New Technologies in Teaching and Learning Chinese*, pages 501–511, Beijing, China. The Tsinghua University Press. 75

Dai, N. and Davison, B. D. (2010). Freshness matters: In flowers, food, and web authority. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 114–121, New York, NY, USA. ACM. 45

Dakka, W., Gravano, L., and Ipeirotis, P. (2012). Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):220–235. 32, 40, 41, 95

de Jong, F., Rode, H., and Hiemstra, D. (2005). Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168, Amsterdam, The Netherlands. Royal Netherlands Academy of Arts and Sciences. 43

Deng, T., Zhao, L., Feng, L., and Xue, W. (2011). Information re-finding by context: A brain memory inspired approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1553–1558, New York, NY, USA. ACM. 32

Derrick, T. R., Bates, B. T., and Dufek, J. S. (1994). Evaluation of time-series data sets using the Pearson product-moment correlation coefficient. *Medicine & Science in Sports & Exercise*, 26(7):919–928. 137

Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., Buchner, K., Liao, C., and Diaz, F. (2010a). Towards recency ranking in web search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 11–20, New York, NY, USA. ACM. 117

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. (2010b). Time is of the essence: Improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 331–340, New York, NY, USA. ACM. 30, 32, 117

Druk, A. (2014). Wikipedia traffic patterns vs. google trends. http://www.wikipediatrends.com/blog/wikipedia-traffic-patterns-vs-google-trends/. 106

Drutsa, A., Gusev, G., and Serdyukov, P. (2015). Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 27–36, New York, NY, USA. ACM. 35

Duan, H. and Hsu, B.-J. P. (2011). Online spelling correction for query completion. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 117–126, New York, NY, USA. ACM. 59, 63

Dumais, S. T. (2004). Latent Semantic Analysis. *Annual Review of Information Science and Technology*, 38:189–230. 48

Efron, M. (2010). Linear time series models for term weighting in information retrieval. *Journal of the American Society for information Science and Technology*, 61(7):1299–1312. 7, 51, 131, 174

Efron, M. (2013). Query representation for cross-temporal information retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 383–392, New York, NY, USA. ACM. 48

Efron, M. and Golovchinsky, G. (2011). Estimation methods for ranking recent information. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 495–504, New York, NY, USA. ACM. 32

Efron, M., Lin, J., He, J., and de Vries, A. (2014). Temporal feedback for tweet search with non-parametric density estimation. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 33–42, New York, NY, USA. ACM. 41

Ellis, P. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press. 79

Elsas, J. L. and Dumais, S. T. (2010). Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 1–10, New York, NY, USA. ACM. 46

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479. 133

Eysenck, M. W. and Keane, M. (2005). *Cognitive Psychology: A Student's Handbook*. Psychology Press. 30

Fafalios, P., Kitsos, I., and Tzitzikas, Y. (2012). Scalable, flexible and generic instant overview search. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 333–336, New York, NY, USA. ACM. 63

Fetterly, D., Manasse, M., Najork, M., and Wiener, J. (2003). A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 669–678, New York, NY, USA. ACM. 46

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Niu, F., Zhang, C., Re, C., and Soboroff, I. (2012). Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In *Proceedings of the Text REtrieval Conference (TREC)*. 52

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164 – 181. 7, 55

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611. 48

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, In Press. 71

Georgescu, M., Kanhabua, N., Krause, D., Nejdl, W., and Siersdorfer, S. (2013a). Extracting event-related information from article updates in wikipedia. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 254–266, Berlin, Heidelberg. Springer-Verlag. 100

Georgescu, M., Pham, D. D., Kanhabua, N., Zerr, S., Siersdorfer, S., and Nejdl, W. (2013b). Temporal summarization of event-related updates in wikipedia. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 281–284, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 52, 100

Gibson, J. (1975). Events are perceivable but time is not. In Fraser, J. and Lawrence, N., editors, *The Study of Time II*, pages 295–301. Springer Berlin Heidelberg. 3, 31

Golbandi, N. G., Katzir, L. K., Koren, Y. K., and Lempel, R. (2013). Expediting search trend detection via prediction of query counts. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 295–304, New York, NY, USA. ACM. 38, 64, 69, 70

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA. ACM. 45

Halavais, A. and Lackaff, D. (2008). An analysis of topical coverage of wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440. 101

Harper, D. J. and van Rijsbergen, C. J. (1978). An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216. 133

He, Q., Chang, K., and Lim, E.-P. (2007). Analyzing feature trajectories for event detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 207–214, New York, NY, USA. ACM. 46

Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition. 14

Hiemstra, D. and Vries, A. P. D. (2000). Relating the new language models of information retrieval to the traditional retrieval models. Technical report, CTIT. 17

Hjørland, B. (2010). The foundation of the concept of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 61(2):217–237. 39

Hu, J., Wang, G., Lochovsky, F., Sun, J.-t., and Chen, Z. (2009). Understanding user's query intent with wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 471–480, New York, NY, USA. ACM. 94, 117, 120

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 3, 28, 29, 30, 31, 33

Jansen, B. J., Booth, D. L., and Spink, A. (2007). Determining the user intent of web search engine queries. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1149–1150, New York, NY, USA. ACM. 93, 167

Jansen, B. J. and Spink, A. (2006). How are we searching the world wide web? a comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248 – 263. 92

Jatowt, A. and Ishizuka, M. (2004). Temporal web page summarization. In Zhou, X., Su, S., Papazoglou, M., Orlowska, M., and Jeffery, K., editors, *Web Information Systems, WISE 2004*, volume 3306 of *Lecture Notes in Computer Science*, pages 303–312. Springer Berlin Heidelberg. 52

Jones, R. and Diaz, F. (2007). Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14–es. 40, 136, 202

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux. 32

Kairam, S. R., Morris, M. R., Teevan, J., Liebling, D. J., and Dumais, S. T. (2013). Towards supporting search over trending events with social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. 6, 30, 59, 61, 98, 117, 126, 168

Kanhabua, N. (2009). Exploiting temporal information in retrieval of archived documents. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 848–848. ACM. 52

Kanhabua, N. and Nørvåg, K. (2008). Improving temporal language models for determining time of non-timestamped documents. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '08, pages 358–370, Berlin, Heidelberg. Springer-Verlag. 43

Kanhabua, N. and Nørvåg, K. (2010a). Determining time of queries for re-ranking search results. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'10, pages 261–272, Berlin, Heidelberg. Springer-Verlag. 39

Kanhabua, N. and Nørvåg, K. (2010b). Exploiting time-based synonyms in searching document archives. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 79–88, New York, NY, USA. ACM. 48

Kanhabua, N. and Nørvåg, K. (2010c). QUEST: query expansion using synonyms over time. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, pages 595–598. 48

Kastrinakis, D. and Tzitzikas, Y. (2010). Advancing search query autocompletion services with more and better suggestions. In *Proceedings of the 10th international conference on Web engineering*, ICWE'10, pages 35–49, Berlin, Heidelberg. Springer-Verlag. 63

Keegan, B., Gergle, D., and Contractor, N. (2013). Hot off the wiki: Structures and dynamics of wikipedia's coverage of breaking news events. *American Behavioral Scientist*. 100, 102

Keikha, M., Gerani, S., and Crestani, F. (2011). Temper: A temporal relevance feedback method. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 436–447, Berlin, Heidelberg. Springer-Verlag. 42

Khodaei, A., Shahabi, C., and Khodaei, A. (2012). Temporal-textual retrieval: Time and keyword search in web documents. *International Journal of Next-Generation Computing (IJNGC)*, 3(3). 51

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 91–101, New York, NY, USA. ACM. 45

Kleinberg, J. (2006). Temporal dynamics of on-line information streams. In Garofalakis M, Gehrke J, R. R., editor, *Data Stream Management: Processing High-Speed Data Streams*. Springer. 45

Knuth, D. E. (1997). *The art of computer programming, volume 1 (3rd ed.): fundamental algorithms*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA. 68

Koren, Y. (2009). Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 447–456, New York, NY, USA. ACM. 53

Kramar, T. and Bielikova, M. (2014). Context of seasonality in web search. In *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 644–649. Springer International Publishing. 36

Kulkarni, A., Teevan, J., Svore, K. M., and Dumais, S. T. (2011). Understanding temporal query dynamics. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 167–176, New York, NY, USA. ACM. 6, 36, 37, 41, 59, 61, 68, 75, 94, 95, 98, 117, 164, 165

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2004). Structure and evolution of blogspace. *Commun. ACM*, 47:35–39. 45

Lathia, N., Hailes, S., Capra, L., and Amatriain, X. (2010). Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 210–217, New York, NY, USA. ACM. 53

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM. 17, 18, 129, 133

Lee, K. S., Croft, W. B., and Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 235–242, New York, NY, USA. ACM. 133

Lehmann, J., Lalmas, M., Yom-Tov, E., and Dupret, G. (2012). Models of user engagement. In *Proceedings of 20th User Modeling, Adaptation, and Personalization International Conference (UMAP)*, pages 164–175. 35

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA. ACM. 45

Li, X. and Croft, W. B. (2003). Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 469–475, New York, NY, USA. ACM. 32

Liebscher, R. and Belew, R. K. (2003). Lexical dynamics and conceptual change : Analyses and implications for information retrieval. *Cognitive Science*, 1:46–57. 7, 51, 130

Lin, J. and Efron, M. (2013). Temporal relevance profiles for tweet search. In *Proceedings of the 2013 SIGIR Workhop on Time-Aware Information Access*, TAIA '13. 43

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165. 50, 136

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *European Physical Journal B*, 11:193–197. 139

Masanès, J. (2007). *Web Archiving*. Springer. 53

Massoudi, K., Tsagkias, M., de Rijke, M., and Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 362–367, Berlin, Heidelberg. Springer-Verlag. 41

Maxwell, D. and Azzopardi, L. (2014). Stuck in traffic: how temporal delays affect search behaviour. In *Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26-29, 2014*, pages 155–164. 33

McParlane, P. J. and Jose, J. M. (2013). Exploiting time in automatic image tagging. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 520–531, Berlin, Heidelberg. Springer-Verlag. 52

McParlane, P. J., Whiting, S., and Jose, J. M. (2013). Improving automatic image tagging using temporal tag co-occurrence. In *Advances in Multimedia Modeling (2)*, pages 251–262. 52

Metzler, D., Cai, C., and Hovy, E. (2012). Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 646–655, Stroudsburg, PA, USA. Association for Computational Linguistics. 46

Metzler, D., Jones, R., Peng, F., and Zhang, R. (2009). Improving search relevance for implicitly temporal queries. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 700–701, New York, NY, USA. ACM. 98

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182. xi, 47, 49, 130, 202

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. 133

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations Workshop*. 48

Mitra, B. and Craswell, N. (2015). Query auto-completion for rare prefixes. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 1755–1758, New York, NY, USA. ACM. 76

Mizzaro, S. (1997). Relevance: the whole history. *J. Am. Soc. Inf. Sci.*, 48(9):810–832. 95

Mooers, C. N. (1952). Information retrieval viewed as temporal signaling. In *Proceedings of the International Congress of Mathematicians*. 7

Moshfeghi, Y. and Jose, J. M. (2013). On cognition, emotion, and interaction aspects of search tasks with different search intentions. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 931–942, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 58

Moshfeghi, Y., Pinto, L. R., Pollick, F. E., and Jose, J. M. (2013). Understanding relevance: an fmri study. In *Advances in Information Retrieval*, number 7814 in Lecture notes in computer science, pages 14–25. Springer. 32

Mourão, F., Rocha, L., Araújo, R., Couto, T., Gonçalves, M., and Meira, Jr., W. (2008). Understanding temporal aspects in document classification. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 159–170, New York, NY, USA. ACM. 53

Nguyen, T. N. and Kanhabua, N. (2014). Leveraging dynamic query subtopics for time-aware search result diversification. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 222–234. 94, 99

Nunes, S., Ribeiro, C., and David, G. (2007). Using neighbors to date web documents. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, WIDM '07, pages 129–136, New York, NY, USA. ACM. 43, 44

Nunes, S., Ribeiro, C., and David, G. (2008). Use of temporal expressions in web search. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 580–584, Berlin, Heidelberg. Springer-Verlag. 38

Odijk, D., Santucci, G., de Rijke, M., Angelini, M., and Granato, G. (2012). Time-aware exploratory search: Exploring word meaning through time. In *SIGIR 2012 Workshop on Time-aware Information Access*, Portland, OR, USA. 47, 49

Ogilvie, P., Voorhees, E., and Callan, J. (2009). On the number of terms used in automatic query expansion. *Inf. Retr.*, 12:666–679. 146, 158

Onnela, J. P., Chakraborti, A., Kaski, K., Kertész, J., and Kanto, A. (2003). Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E*, 68(5):56110. 139

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., and Ounis, I. (2012). Bieber no more: First Story Detection using Twitter and Wikipedia. *SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)*. 100, 102

Ounis, I., Macdonald, C., Lin, J., and Soboroff, I. (2011). Overview of the trec 2011 microblog track. *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*. 32

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. 44, 141, 143

Pass, G., Chowdhury, A., and Torgeson, C. (2006). A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, InfoScale '06, New York, NY, USA. ACM. 73

Peetz, M.-H. and de Rijke, M. (2013). Cognitive temporal document priors. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 318–330, Berlin, Heidelberg. Springer-Verlag. 31, 32

Peetz, M.-H., Meij, E., and Rijke, M. (2014). Using temporal bursts for query modeling. *Inf. Retr.*, 17(1):74–108. 41, 131, 134, 147, 157

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM. 32

Pöppel, E. (1978). Time perception. In Held, R., Leibowitz, H., and Teuber, H.-L., editors, *Perception*, volume 8 of *Handbook of Sensory Physiology*, pages 713–729. Springer Berlin Heidelberg. 3, 31

Potthast, M., Stein, B., and Gerling, R. (2008). Automatic vandalism detection in wikipedia. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 663–668, Berlin, Heidelberg. Springer-Verlag. 102

Preum, S. M., Stankovic, J., and Qi, Y. (2015). Maper: A multi-scale adaptive personalized model for temporal human behavior prediction. In *Proceedings of the Twenty Fourth International Conference on Information and Knowledge Management*, CIKM '15, pages 469–475, New York, NY, USA. ACM. 34, 35

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA. ACM. 48, 49, 130, 134, 137, 170

Radinsky, K., Diaz, F., Dumais, S., Shokouhi, M., Dong, A., and Chang, Y. (2013a). Temporal web dynamics and its application to information retrieval. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 781–782, New York, NY, USA. ACM. 42, 98

Radinsky, K., Svore, K. M., Dumais, S. T., Shokouhi, M., Teevan, J., Bocharov, A., and Horvitz, E. (2013b). Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM Trans. Inf. Syst.*, 31(3):16:1–16:37. 7, 26, 34, 43, 55, 94, 95, 98, 125, 167, 174

Radlinski, F., Kurup, M., and Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 43–52, New York, NY, USA. ACM. 95

Richardson, M. (2008). Learning about the world through long-term query logs. *ACM Trans. Web*, 2(4):21:1–21:27. 30

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1996). Okapi at trec-3. pages 109–126. 17

Robertson, S. E. (1997). *The Probability Ranking Principle in IR*, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 16

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall. 129, 133

Rodriguez, M. A. and Bollen, J. (2006). Simulating network influence algorithms using particle-swarms: Pagerank and pagerank-priors. Technical report, Los Alamos National Laboratory. LA-UR-06-2139. 142

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 513–522, New York, NY, USA. ACM. 53

Safran, N. (2013). Excluding one word queries, wikipedia appears on page one about 10 percent more often on google vs. bing. http://www.conductor.com/blog/2012/05/googles-love-affair-with-wikipedia-far-more-serious-than-bings-study. 106

Sakai, T., Dou, Z., and Clarke, C. L. (2013). The impact of intent selection on diversified search evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 921–924, New York, NY, USA. ACM. 111

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM. 45, 102

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620. 15

Sanderson, M. (2008). Ambiguous queries: Test collections need more sense. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 499–506, New York, NY, USA. ACM. 3, 92

Santos, R. L., Macdonald, C., and Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 595–604, New York, NY, USA. ACM. 93, 94, 95

Saracevic, T. (2007). 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. 39

Savolainen, R. (2006). Time as a context of information seeking. *Library & Information Science Research*, 28(1):110 – 127. 33

Sengstock, C. and Gertz, M. (2011). Conquer: A system for efficient context-aware query suggestions. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 265–268, New York, NY, USA. ACM. 64

Shokouhi, M. (2011). Detecting seasonal queries by time-series analysis. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1171–1172, New York, NY, USA. ACM. 38, 61, 62, 63, 68, 75

Shokouhi, M. (2013). Learning to personalize query auto-completion. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 103–112, New York, NY, USA. ACM. 63, 76, 174

Shokouhi, M. and Radinsky, K. (2012). Time-sensitive query auto-completion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 601–610, New York, NY, USA. ACM. 7, 38, 58, 63, 64, 65, 66, 68, 69, 75, 76, 77, 78

Siblini, R. and Kosseim, L. (2013). Using a weighted semantic network for lexical semantic relatedness. In *Recent Advances in Natural Language Processing*, pages 610–618. ACL. 133

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. 92

Smucker, M. D. and Clarke, C. L. A. (2012). Modeling user variance in time-biased gain. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '12, pages 3:1–3:10, New York, NY, USA. ACM. 35

Song, R., Luo, Z., Wen, J.-R., Yu, Y., and Hon, H.-W. (2007). Identifying ambiguous queries in web search. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1169–1170, New York, NY, USA. ACM. 93, 167

Sparck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21. 16, 50

Steiner, T., van Hooland, S., and Summers, E. (2013). Mj no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 791–794, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 100

Strizhevskaya, A., Baytin, A., Galinskaya, I., and Serdyukov, P. (2012). Actualization of query suggestions using query logs. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 611–612, New York, NY, USA. ACM. 38, 64, 69

Strötgen, J., Alonso, O., and Gertz, M. (2012). Identification of top relevant temporal expressions in documents. In *Proceedings of the 2nd Temporal Web Analytics Workshop*, TempWeb '12, pages 33–40, New York, NY, USA. ACM. 38, 51

Sun, M., Lebanon, G., and Collins-Thompson, K. (2010). Visualizing differences in web search algorithms using the expected weighted hoeffding distance. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 931–940, New York, NY, USA. ACM. 98

Teevan, J., Adar, E., Jones, R., and Potts, M. A. S. (2007). Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 151–158, New York, NY, USA. ACM. 31, 74

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 the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '13, pages 1:1–1:10, New York, NY, USA. ACM. 33

Teevan, J., Ramage, D., and Morris, M. R. (2011). #twittersearch: A comparison of microblog search and web search. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 35–44, New York, NY, USA. ACM. 6, 129, 144

Tsai, F. S., Etoh, M., Xie, X., Lee, W.-C., and Yang, Q. (2010). Introduction to mobile information retrieval. *IEEE Intelligent Systems*, 25(1):11–15. 6

Uzzaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics. 25

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–199. 133

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth. 15, 133

Vis, F. (2009). Wikinews reporting of hurricane katrina. In *Citizen Journalism: Global Perspectives*, Global Crises and the Media. Peter Lang. 100

Vlachos, M., Meek, C., Vagena, Z., and Gunopulos, D. (2004). Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 131–142, New York, NY, USA. ACM. 38

Voorhees, E. M. and Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press. 19, 21

Wang, P., Berry, M. W., and Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):743–758. 37

Wang, X., Zhai, C., Hu, X., and Sproat, R. (2007). Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 784–793, New York, NY, USA. ACM. 45

Wang, Y. and Lin, J. (2014). The impact of future term statistics in real-time tweet search. In de Rijke, M., Kenter, T., de Vries, A., Zhai, C., de Jong, F., Radinsky, K., and Hofmann, K., editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 567–572. Springer International Publishing. 145

Wattenberg, M., Viégas, F. B., and Hollenbach, K. (2007). Visualizing activity on wikipedia with chromograms. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part II*, INTERACT'07, pages 272–287, Berlin, Heidelberg. Springer-Verlag. 52

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38. 108

White, R. W., Jose, J. M., and Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Inf. Process. Manage.*, 42(1):166–190. 29

White, R. W. and Roth, R. A. (2009). *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers. 14, 29

White, S. and Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 266–275, New York, NY, USA. ACM. 141, 142

Whiting, S., Alonso, O., and Jose, J. M. (2015). Temporal dynamics of ambiguous queries. In *SIGIR 2015 Workshop on Time-aware Information Access (TAIA2015)*, Santiago, Chile. 92

Whiting, S., Jose, J., and Alonso, O. (2014). Wikipedia as a time machine. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 857–862, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 92

Whiting, S. and Jose, J. M. (2014). Recent and robust query auto-completion. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 971–982, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 58

Whiting, S., Klampanos, I. A., and Jose, J. M. (2012a). Temporal pseudo-relevance feedback in microblog retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 522–526, Berlin, Heidelberg. Springer-Verlag. 128

Whiting, S., McMinn, J., and Jose, J. M. (2013a). Exploring real-time temporal query auto-completion. In *Dutch-Belgian IR Workshop (DIR)*, pages 12–15. 58, 198

Whiting, S., Moshfeghi, Y., and Jose, J. M. (2011). Exploring term temporality for pseudo-relevance feedback. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1245–1246, New York, NY, USA. ACM. 128, 201

Whiting, S., Zhou, K., Jose, J., Alonso, O., and Leelanupab, T. (2012b). Crowdtiles: Presenting crowd-based information for event-driven information needs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2698–2700, New York, NY, USA. ACM. 52

Whiting, S., Zhou, K., Jose, J., and Lalmas, M. (2013b). Temporal variance of intents in multi-faceted event-driven information needs. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 989–992, New York, NY, USA. ACM. 92

Wijaya, D. T. and Yeniterzi, R. (2011). Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, DETECT '11, pages 35–40, New York, NY, USA. ACM. 49

Yin, Z., Shokouhi, M., and Craswell, N. (2009). Query expansion using external evidence. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 362–374, Berlin, Heidelberg. Springer-Verlag. 133

Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410, New York, NY, USA. ACM. 129, 133

Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 10–17, New York, NY, USA. ACM. 98

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 116–, New York, NY, USA. ACM. 71

Zhang, Y., Jansen, B. J., and Spink, A. (2009). Time series analysis of a web search engine transaction log. *Inf. Process. Manage.*, 45(2):230–245. 37

Zhao, Q., Liu, T.-Y., Bhowmick, S. S., and Ma, W.-Y. (2006). Event detection from evolution of click-through data. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 484–493, New York, NY, USA. ACM. 37

Zhou, K., Whiting, S., Jose, J. M., and Lalmas, M. (2013). The impact of temporal intent variability on diversity evaluation. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 820–823, Berlin, Heidelberg. Springer-Verlag. 94, 96, 99

Zhuang, Y. (2014). Building a complete tweet index. https://blog.twitter.com/2014/building-a-complete-tweet-index. 136

Zipf, G. (1949). Human behaviour and the principle of least-effort. Addison-Wesley, Cambridge, MA. 8, 140

Zliobaite, I., Bifet, A., Pfahringer, B., and Holmes, G. (2014). Active learning with drifting streaming data. *IEEE Trans. Neural Netw. Learning Syst.*, 25(1):27–39. 71

Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J. M., and Azzopardi, L. (2013). Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Inf. Retr.*, 16(2):267–305. 20

Appendix A

Preliminary Recent Query Auto-completion Findings

In early published work (Whiting et al., 2013a), I explored value of recent query auto-completion (QAC). This work studied QAC performance using only recent query log evidence in the common maximum likelihood QAC approach (i.e., MLE-W) versus using all past query log evidence in the maximum likelihood approach (MLE-ALL). MLE-W was found to significantly improve QAC effectiveness so it was employed as the hard baseline for experiments in Chapter 4.

I report the findings of this preliminary work in Table A.1 for reference. Note that the raw query logs (i.e., AOL and MSN) and approaches (i.e., MLE-ALL and MLE-W) used to generate the results presented in Table A.1 are the same as those presented in Chapter 4 on page 79. However, the QAC performance as measured by the mean reciprocal rank (MRR) metric reported in these two tables differs, due additional query log data cleaning in the later work presented in Chapter 4. This improved cleaning filtered web addresses and repetitive spam from the query log, which temporal QAC approaches could overfit and unfairly exploit. Accordingly, this pre-processing makes the final results reported in Chapter 4 more robust. Also important to note is that the scale-invariant pattern and ranking of the effect of MLE-W and different window sizes (in days) on QAC performance remains the same regardless of the query log filtering, thus emphasising the effect of time in QAC.

Table A.1: MRR observed for QAC when using all prior query log evidence, and the past 2, 4, 7 or 14 days of query-log evidence. Prefix (ρ) lengths of 2 to 5 characters are reported for the AOL and MSN query logs. The best performing sliding window setting is highlighted for each prefix length (although in some cases this is still outperformed by the baseline, at least for the reported window periods).

ρ	**MLE-ALL**	**MLE-W2**	**MLE-ALL**	**MLE-W4**	**MLE-ALL**	**MLE-W7**	**MLE-ALL**	**MLE-W14**
AOL								
2	0.090	**0.091 (1.11%)**	0.090	0.091 (1.11%)	0.090	0.091 (1.11%)	0.090	0.091 (1.11%)
3	0.143	**0.147 (2.80%)**	0.143	0.146 (2.10%)	0.143	0.145 (1.40%)	0.143	0.145 (1.40%)
4	0.185	0.189 (2.16%)	0.184	**0.189 (2.72%)**	0.184	0.188 (2.17%)	0.184	0.187 (1.63%)
5	0.217	0.215 (-0.92%)	0.216	0.217 (0.46%)	0.217	0.218 (0.46%)	0.217	**0.219 (0.92%)**
MSN								
2	0.112	**0.117 (4.46%)**	0.111	0.115 (3.60%)	0.111	0.113 (1.80%)	0.110	0.111 (0.91%)
3	0.164	0.163 (-0.61%)	0.164	**0.165 (0.61%)**	0.164	**0.165 (0.61%)**	0.164	0.164 (0.00%)
4	0.197	0.188 (-4.57%)	0.197	0.193 (-2.03%)	0.197	0.196 (-0.51%)	0.197	**0.197 (0.00%)**
5	0.215	0.197 (-8.37%)	0.216	0.205 (-5.09%)	0.216	0.211 (-2.31%)	0.218	**0.216 (-0.92%)**

Appendix B

Multi-faceted Test Events/Queries

The following twenty queries were used as test queries in Section 5.5 on page 116.

Short-term Events:

(1) Eyjafjallajokull, (2) G-20 Cannes summit, (3) 54th Grammy Awards, (4) Kony 2012, (5) Tucson Shooting, (6) Costa Concordia disaster, (7) Hurricane Irene, (8) 2011 England riots, (9) 2011 Lokomotiv Yaroslavl air disaster, and (10) 2011 Van earthquake.

Long-term Events:

(1) 2011 military intervention in Libya, (2) Arab Spring, (3) Credit crunch, (4) European sovereign-debt crisis, (5) 2011 Norway attacks, (6) Occupy movement, (7) News International phone hacking scandal, (8) Fukushima Daiichi nuclear disaster, (9) Wedding of Prince William and Catherine Middleton, and (10) 2011 Thailand floods.

Appendix C

Preliminary Temporal Query Expansion Findings

In early published work (Whiting et al., 2011), I explored the possibility of exploiting the temporal dynamics of terms to enhance pseudo-relevance feedback performance. I report the findings of this preliminary work here for completeness, since the Temporal Semantic Query Expansion (TSQE) approach I present and exhaustively evaluate in Chapter 6 supersedes this preliminary naive approach by addressing several limitations.

C.1 Approach

I proposed two straight-forward but novel pseudo-relevance feedback (PRF) approaches which relied on the temporal dynamics of document and query terms to identify valuable terms for query expansion (QE). Conventional PRF approaches rank possible expansion terms by their term frequency (TF) in initially retrieved documents, i.e., PRF(Text). The first experimental approach ignored TF and instead ranked terms solely by their temporal dynamic correlation with the most temporally significant term extracted from the query – namely PRF(Temporal). This temporal correlation (TC) is computed as the Pearson's correlation co-efficient between each term's temporal dynamics, the source of which is further detailed below. The second approach, PRF(Text+Temporal), combines both sources of evidence by computing $TC \times TF$ – thereby taking temporal and traditional TF evidence into account simultaneously.

In this early work I used externally sourced temporal dynamics, rather than the temporal dynamics extracted from the collection itself. Accordingly, the temporal dynamics of a term over a prolonged period was obtained from the long-term Google Books (GB) n-gram dataset[1]

[1] This extensive dataset provides annual term popularity statistics through the extraction of 1 (single word) to 5-gram (five word phrases) from the digitised Google Books corpus, containing over 5m published works (roughly 4% of books ever published), and over 8 billion English words and phrases.

Michel et al. (2010). While the resolution of these temporal dynamics is coarse (i.e., annual popularity), books are likely to have sufficient coverage of significant temporal words and phrases for this application.

To use the GB dataset, I first had to extract descriptive n-grams from the query (in this case, the TREC topic title) and the PRF documents. The Yahoo Term Extraction Service[1] was employed for this purpose since it identified meaningful keywords, entities and noun phrases for further processing.

For this work, I considered the extracted query term with the strongest temporal pattern – measured as the term with the greatest temporal dynamic kurtosis[2] – as the primary clue of a topic's temporal nature. Notably, analysis showed frequent terms in the collection (that is, *stopwords*) have the lowest kurtosis since they are used frequently over time without temporal discrimination – so this step also serves to discard them from further processing. Consequently, the highest kurtosis topic related n-gram is selected to temporally correlate with candidate PRF terms extracted from pseudo-relevant documents. As in Chapter 6, Pearson's correlation co-efficient was used to measure temporal similarity between document and query term temporal dynamics.

C.2 Experiment and Results

Documents were indexed using Indri 5.0[3] with standard stop-word removal and Krovetz stemming applied. The top 1000 documents were retrieved for each topic using a unigram language model (LM) with Indri default Dirichlet smoothing ($\mu = 2500$). This run formed the baseline (identified as LM in Table C.1) against which I compared the three PRF approaches. For comparison, I consider performance of PRF(Text) as a much stronger PRF baseline. In each PRF approach, the top n temporally correlated or frequent terms (with $n = 10$, 20) in the top 10 retrieved documents were linearly-combined and used to expand the query[4]. To reduce query drift, the expanded query includes the original query terms weighted at 0.7, with possibly non-relevant PRF expansion terms weighted at 0.3.

The GB data was indexed, making n-grams case-insensitive. The yearly resolution temporal dynamic for each term from 1950-1994 was obtained from the GB index (aligning with test collection periods). Due to computational time and space limitations, only 1- and 2-grams were used in this study.

[1] `http://developer.yahoo.com/search/content/V1/termExtraction.html`

[2] A statistical measure of the 'peakedness' of a time series (Jones and Diaz, 2007)

[3] `http://www.lemurproject.org/indri`

[4] Indri `#combine` query operators are used, ultimately treating n-grams as multiple unigrams in the unigram language model.

Experiment Settings. Preliminary evaluation of the PRF technique is performed using the Associated Press (AP88-89) and Wall Street Journal (WSJ87-92) news wire test collections, along with TREC-1 ad hoc topics 51-100. I used a news collection at this stage as it is was most likely to contain topics with some temporal element. Mean average precision (MAP) for the top 1000 documents is used as the main effectiveness evaluation measure. Precision at 30 (P@30) and Recall have been included to observe whether PRF increases recall at the cost of precision. Furthermore, the improvement metric (IMP) is included as the number of individual topics with improved MAP over the LM baseline, in order to analyse per-topic variation.

C.2.1 Results

The results reported in Table C.1 indicate that the temporal dynamics of term popularity can be employed to enhance retrieval effectiveness. More specifically, the temporal-based PRF(Temporal) approach displays better performances than LM. We also observe that combining textual and temporal aspects of terms in PRF documents – i.e., PRF(Text+Temporal) – leads to the best performance (shown to be statistically significant), increasing over that of PRF(Text) alone. Thus, whilst effectiveness gains are marginal on average in this preliminary study, there is a case for using the long-term temporal correlation of terms in PRF term selection, particularly when considering individual topic improvements (i.e., IMP). Improvement gain across approaches is due to a significant variation in the selected terms as reflected by a per-topic average Spearman's ρ of 0.05 and 0.06 for AP88-89 and WSJ87-92 respectively.

Overall, the results of this initial study were encouraging since they show improvement when including temporal profile correlation in PRF term selection. Of course there were several limitations in how the temporal aspect was identified, extracted and exploited, so further work addressing these issues led to a more elaborate model (cf. Temporal Semantic Query Expansion) which was presented in Chapter 6 of this thesis.

Table C.1: Reporting MAP, P@30, Recall and IMP (based on MAP) values for LM (i.e., baseline), PRF(Temporal), PRF(Text), and PRF(Text+Temporal) using the AP88-89 and WSJ87-92 test collections and TREC Topics (51-100). Best performing approach for each metric is highlighted. Paired t-test statistical significance is denoted as * being $p < 0.05$ and ** being $p < 0.01$.

AP88-89	LM	PRF(Temporal)		PRF(Text)		PRF(Text+Temporal)	
		Top 10 Terms	Top 20 Terms	Top 10 Terms	Top 20 Terms	Top 10 Terms	Top 20 Terms
MAP	0.274	0.286 (+4.4%)*	0.288 (+5.2%)**	0.303 (+10.4%)**	0.304 (+11.0%)**	__0.305 (+11.2%)**__	0.304 (+11.0%)**
P@30	0.368	0.374 (+1.4%)	0.379 (+2.9%)*	__0.4 (+8.6%)__*	0.395 (+7.4%)**	0.393 (+6.8%)**	0.395 (+7.2%)**
Recall	0.640	0.663 (+3.6%)	0.662 (+3.5%)**	0.696 (+8.8%)**	0.694 (+8.5%)**	__0.699 (+9.2%)**__	0.694 (+8.5%)**
IMP		20	23	22	26	26	__28__

WSJ87-92	LM	PRF(Temporal)		PRF(Text)		PRF(Text+Temporal)	
		Top 10 Terms	Top 20 Terms	Top 10 Terms	Top 20 Terms	Top 10 Terms	Top 20 Terms
MAP	0.256	0.27 (+5.4%)**	0.27 (+5.1%)**	0.284 (+11.0%)**	0.283 (+10.6%)**	__0.286 (+11.7%)**__	0.284 (+11.1%)**
P@30	0.373	0.384 (+3.1%)	0.387 (+3.9%)*	0.393 (+5.5%)*	0.395 (+5.9%)*	__0.399 (+6.9%)*__	__0.399 (+6.9%)*__
Recall	0.596	0.614 (+3.1%)**	0.617 (+3.6%)**	0.650 (+9.2%)**	0.652 (+9.4%)**	__0.652 (+9.6%)**__	0.649 (+8.9%)**
IMP		24	26	31	33	34	__35__