Information Retrieval on Time-Dependent Collections

Sérgio Nunes

Doctoral Program in Informatics Engineering

Faculdade de Engenharia
Universidade do Porto
2010

Contact Information

Sérgio Nunes
Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto

Rua Dr. Roberto Frias, s/n
4200-465 Porto
PORTUGAL

Email: sergio.nunes@fe.up.pt
URL: `http://www.fe.up.pt/~ssn`

Sérgio Nunes
"Information Retrieval on Time-Dependent Collections"
Copyright © 2010 by Sérgio Nunes. All rights reserved.

Ao meu irmão David e ao meu filho Miguel.

Abstract

Information Retrieval deals with the representation, storage and access to information items. The first works in this field date back to the late 40s of the 20th century. Technologies and innovations introduced in this discipline are at the core of information search systems, central pieces of today's information society. One of the first fundamental assumptions established was the existence of a static underlying collection of documents that is regularly replaced. Classic systems view the collection as fixed regardless of previous changes to the documents.

In this dissertation we relax this assumption and consider the document collection a time-dependent entity. The core statement of this work is that the dynamic facet of a document collection can be exploited to improve standard information retrieval tasks. More specifically, document retrieval performance is improved if the dynamic structure of the collection is taken into account to derive and integrate new signals.

This work uses the World Wide Web as the central document collection for all explorations. We start by proposing a taxonomy to catalog potential sources of temporal information on the web. These are organized in two main groups, document-based features and web-based features. Document-based features are structured according to the source of the data, more specifically by document's content, URL address and HTTP protocol. Correspondingly, web-based features can be based in a document's neighborhood, external archives and web logs.

We conduct an exploratory study to understand the use of time-dependent expressions in user issued web search queries. We find that temporal expressions are used in 1.5% of all queries. Moreover, most temporal expressions reference current or past dates, exhibiting a long tailed behavior. Future dates are rarely used. Also, these expressions are more frequently used in topics such as autos, sports, news and holidays.

In previous works, web documents' HTTP header fields were used to estimate a docu-

ment's most recent update. The main problem with this approach is that it is applicable only to a small part of web documents, those showing valid header information. We show that, by using a document's neighborhood, it is possible to estimate the date of documents where this information is either missing or unreliable. If the vicinity of each web document is considered, it is possible to increase the number of documents with a date estimate by up to 60%.

Classic link-based authority estimation algorithms are time-agnostic. We observe that the distribution of citations to a web document over time contains valuable information. We show that, by using temporal information, algorithms can respond to current events more quickly and produce a better overall query-independent authority estimation. We demonstrate that temporal information on the web can be used to derive stable time-dependent features, which can be successfully used in the context of web document ranking.

Keyword extraction from documents is a core step in any information retrieval system. Simple term frequency count is still one of the fundamental signals used to identify good quality terms for document indexing. We show that document history, given the rapid progression of content towards a stable version, is an important source of information for improving keyword extraction. The keyword extraction techniques proposed in this dissertation clearly outperform the classic term frequency count. Moreover, relevant terms, that are otherwise inaccessible, can be recovered using methods that skim through a document's history.

Resumo

A área da Recuperação de Informação tem como ponto central de investigação a representação, o armazenamento e o acesso a itens de informação. Os primeiros trabalhos nesta área remontam ao final dos anos 40 do século XX. As tecnologias e inovações introduzidas por esta disciplina estão no cerne dos sistemas de pesquisa de informação, peças fundamentais na actual sociedade da informação. Um dos mais antigos pressupostos nesta área considera a existência de uma colecção base de documentos fixa que é regularmente renovada. Os sistemas clássicos encaram esta colecção como estática, independentemente das mudanças que ocorrem nos documentos.

Nesta dissertação, relaxamos esta assumpção e consideramos que a colecção de documentos é uma entidade dependente do tempo. A afirmação central deste trabalho estabelece que a componente dinâmica de uma colecção de documentos pode ser explorada para melhorar as tarefas padrão, na área da recuperação de informação. Mais especificamente, o desempenho na recuperação de documentos é melhorado se a estrutura dinâmica da colecção for tida em conta para extrair e integrar novos sinais.

Este trabalho utiliza a World Wide Web como colecção de documentos central em todas as explorações. Começamos por propor uma taxonomia para catalogar as potenciais fontes de informação temporal na web. Estas fontes são organizadas em dois grupos principais, as baseadas no documento e as baseadas na web. As características extraídas dos documentos são estruturadas de acordo com a fonte dos dados, mais especificamente a origem pode ser o conteúdo do documento, o endereço URL e o protocolo HTTP. Analogamente, as características extraídas da web podem ser extraídas da vizinhança de um documento, de arquivos externos e de registos de acesso web.

Realizamos um estudo exploratório para perceber como são usadas expressões temporais nas interrogações submetidas a sistemas de pesquisa na web. Verificamos que as expressões temporais são usadas em 1,5% das interrogações. Adicionalmente, vimos

que a maioria destas expressões referenciam datas actuais or datas passadas, exibindo uma distribuição de cauda longa. Datas futuras raramente são usadas. Constatamos, também, que estas expressões são mais frequentes nos seguintes tópicos: automotores, desportos, notícias e feriados/férias.

No passado, o cabeçalho HTTP de documentos web foi usado para estimar a data da última actualização ao documento. O principal problema desta abordagem reside no facto desta técnica só ser aplicável a uma parte reduzida dos documentos na web, aqueles que apresentam cabeçalhos válidos. Neste trabalho, demonstramos que, recorrendo à vizinhança de um documento, é possível estimar a data dos documentos em que esta informação é omissa ou não fiável. Se a vizinhança de cada documento web de uma colecção for considerada, é possível aumentar até 60% o número de documentos com uma estimativa de data.

Os algoritmos clássicos usados para estimar autoridade com base em ligações ignoram as propriedades temporais dos dados. Neste contexto, observamos que a distribuição das citações a um documento ao longo do tempo, contém informação valiosa. Mostramos que, usando informação temporal, os algoritmos reagem mais rapidamente a eventos recentes, o que resulta em estimativas de autoridade de maior qualidade. Demonstramos também que a informação temporal existente na web pode ser usada para extrair sinais estáveis dependentes do tempo, e que estes podem ser usados com sucesso no contexto da ordenação de documentos.

A extracção de palavras-chave de documentos é um passo fulcral em qualquer sistema de recuperação de informação. A frequência de ocorrência de termos é ainda um dos sinais fundamentais usados para identificar os termos de boa qualidade para a fase de indexação. Provamos que o histórico de um documento, tendo em conta a rápida estabilização do conteúdo, é uma importante fonte de informação para melhorar a extracção de palavras-chave. As técnicas de extracção de informação propostas nesta dissertação, são claramente mais eficazes do que a simples contagem de termos. Concluimos, também, que termos relevantes, de outra forma inacessíveis, podem ser extraídos usando técnicas que percorrem o historial dos documentos.

Résumé

La Recherche d'Information traite de la représentation, le stockage et l'accès aux éléments d'information. Les premiers travaux dans ce domaine remontent à la fin des années 40 du 20ème siècle. Les technologies et les progrès amenés par cette discipline sont cruciaux pour les systèmes de recherche d'informations et des éléments centraux de la société d'information d'aujourd'hui. Une des premières hypothèses fondamentales établies a été l'existence d'une collection sous-jacente de documents fixes qui est régulièrement mis à jour. Les systèmes traditionnels analysent cette collection comme fixe, indépendamment des changements opérés dans ces documents.

Dans cette thèse, nous laisserons de côté cette hypothèse et nous considèrerons que cette collection de documents une chose intrinsèquement liée au temps. Laffirmation centrale de ce travail est que la composante dynamique d'une collection de documents peut être exploitée pour améliorer le niveau des tâches de recherche d'informations. Plus précisément, la récupération des documents est améliorée si la structure dynamique de la collection est prise en compte dans lextraction et lintégration de nouvelles données.

Ce travail utilise le World Wide Web pour toutes ces recherches et pour la collection de documents centraux. Nous commencerons par proposer une taxonomie pour cataloguer les sources potentielles d'informations temporelles sur le web. Elles sont organisées en deux groupes principaux, les caractéristiques du document et les fonctionnalités web. Les caractéristiques des documents sont structurées en fonction de la source des données, plus particulièrement par le contenu du document, l'adresse URL et le protocole HTTP. En conséquence, les caractéristiques sur le web peuvent être fondées sur les similarités d'un document, des archives et de journaux web.

Nous mènerons une étude exploratoire pour comprendre comment sont utilisées les expressions temporelles à travers des interrogations soumises aux systèmes de recherche

du web. Nous constaterons que les expressions temporelles sont utilisées dans 1,5% de toutes les requêtes. En outre, la plupart de ces expressions font référence à des dates actuelles ou passées, sur une très longue période. Les dates futures sont rarement utilisées. Nous avons également constaté que ces expressions sont plus fréquentes sur des sujets tels que : lautomobile, le sport, les actualités et les jours fériés/vacances.

Dans le passé, len-tête HTTP des documents web a été utilisé pour estimer la date de la dernière mise à jour d'un document. Le principal problème avec cette approche est qu'elle est applicable seulement à une petite partie des documents du web : les documents ayant un en-tête valide. Nous démontrerons que, en utilisant lanalogie d'un document, il est possible d'estimer la date des documents lorsque cette information est absente ou peu fiable. Si lanalogie de chaque document web était prise en compte, il serait possible d'augmenter le nombre de documents avec une estimation de date de près de 60%.

Les algorithmes classiques utilisés pour évaluer lautorité des liens ne sont pas sujets au temps. Nous observerons que la distribution des citations d'un document web au fil du temps contient de précieuses informations. Nous démontrerons que, en utilisant des informations temporelles, les algorithmes peuvent réagir à l'actualité plus rapidement et produire une meilleure estimation globale de lautorité. Nous démontrerons encore que linformation temporelle sur le web peut être utilisée pour obtenir des données stables dépendantes du temps, et que ces dernières peuvent être utilisées avec succès dans le cadre du classement des documents du web.

L'extraction de mots-clés dun document est une étape de base dans le processus de base de la récupération d'information. Un décomptage simple de la fréquence des termes est même une des techniques fondamentales utilisées pour identifier les termes de bonne qualité pour l'indexation de documents. Nous démontrerons que l'historique des documents, lorsque lon prend en compte la stabilisation rapide du contenu, est une source importante d'informations pour l'amélioration de l'extraction des mots clés. Les techniques d'extraction dinformations proposées dans cette thèse sont nettement supérieures à la fréquence de décompte classique pour chaque terme. En outre, les termes pertinents qui seraient inaccessibles autrement, peuvent être récupérés en utilisant des techniques qui se trouvent dans lhistorique du document en question.

Acknowledgements

First and foremost, I thank my advisor, Cristina Ribeiro, for the consistent and dedicated support and encouragement throughout these years. I am forever in debt for her mentoring in many aspects of my academic and personal life. Second, I express my gratitude to my co-advisor, Gabriel David, for his help and wise counseling in many occasions. Both have significantly contributed to make this journey of learning and discovering a true pleasure.

I had the unique opportunity to fully dedicate myself to research in the course of my Ph.D. I thank DEEC and DEI for making this possible, in particular to José Silva Matos and to Raul Moreira Vidal, for authorizing my exemption from faculty service.

This work was financially supported by a scholarship from the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006. I thank FCT and FSE for this important contribution.

I thank Mark Sanderson and Mário J. Silva, both serving on my thesis committee, for their valuable contributions on the initial stages of my Ph.D. I also thank SAPO and PT Comunicações, in particular to Benjamin Júnior for his pivotal role, for giving me the opportunity to access data and services from SAPO's infrastructure. Many of the experiments were made possible by this rewarding cooperation.

I thank Carla, Manuel, my mother Isabel and Mário Júnior for their help in reviewing the final versions of this text.

My friends have always been a source of strength and balance in my life. I am fortunate to have such loyal and inspiring friendships. They were, most times without realizing, a source of motivation and energy.

Finally, I express my gratitude to three of the most important people in my life, Carla, my mother Isabel and Manuel. I know that I can always count on you for love, support, patience and honesty. I aspire to give you what you have always given me.

Since I started my Ph.D. I have lost and gained more than I could ever imagine possible. This work is entirely dedicated to my brother David and my son Miguel.

Brief Contents

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of information retrieval has extensively addressed the problems related to the storage, access and retrieval of information. The classic approach in this area assumes a static underlying collection of documents. We relax this basic presumption and address the challenges and opportunities introduced by incorporating a time-aware view of the documents. In the context of dynamic collections, we address several problems related to information retrieval, ranging from information extraction for feature selection to content and link analysis for ranking. This is a topic that has received little attention from the research community when compared with other themes in information retrieval.

This chapter presents a general introduction to the dissertation. It starts with a brief review of the tasks in information access and retrieval, with a strong emphasis on the dynamic nature of information. The temporal dimension of information is used to introduce the background of the problems addressed here. These problems are presented as a thesis statement in this chapter. Finally, the chapter concludes with a summary of the structure of this document.

1.1 Information Retrieval

Information access has become an activity of major importance in today's society. The advent of a *knowledge-based economy* inline with the development of the *information society* has promoted the appearance of a myriad of information-related tasks and jobs.

These activities are not limited to the business realm, in reality they have become ubiquitous in today's world. Archiving, retrieving or searching through information are everyday activities either in professional or in personal contexts.

The field of Informatics has studied many of the problems related to information access and retrieval within the discipline of *Information Retrieval*. The first references to the use of computers for information management can be traced back to the seminal essay "As We May Think" published in 1945 [23]. In this text, Vannevar Bush envisioned a future where computers would play a fundamental role in all tasks related to information access. Since these early years, the field of information retrieval has grown to be an active and prominent area in computer and information science. A contemporary definition of *Information Retrieval* can be found in Manning et al. [61]:

> Information retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.

In this sense, information retrieval, or simply *information search*, deals with a wide range of technologies and concepts. The main topics in information retrieval can be outlined as "the representation, storage, organization of, and access to information items" [8]. Information retrieval includes not only low-level computer-science concepts such as data storage and organization in physical disks, but also high-level concepts involving the fields of human-computer interaction, knowledge representation and information science. Although search is the most visible problem addressed in this discipline, there are several other more specific tasks that are also studied like text summarization, information filtering or automatic question answering.

The first developments in IR, starting in the late 1940s [61], were focused on using computer-based mechanisms for performing classic information-related tasks typically executed by humans. An example is the automatic construction of indices for fast keyword-based search within a single document or a collection of texts. The first studies over concepts such as *term frequency* and *document frequency* date back to those years. This led to the development of the *tf-idf* weighting scheme, one of the most important techniques for traditional keyword-based retrieval, first defined by Spärck Jones in 1972 [48]. In a nutshell, the *tf-idf* measure defines a weight for each term in each document of the collection. This score depends on the frequency of the term within the document and on the inverse frequency of documents containing the term in the collection. This approach is still at the core of many information retrieval systems.

The second half of the 20th century witnessed the development of commercial solutions

in the information retrieval sector [65]. This evolution was led by the growth observed in the science and technology literature. The main publishers developed information systems that allowed an easy access to past and current abstracts. These solutions were highly specialized services, usually designed to be primarily used by librarians and documentation experts (e.g. MEDLINE in the medical area, LEXIS/NEXIS in the legal area). Before the internet, access to these products was made through private networks and custom software applications. The invention of the World Wide Web is a landmark event in the field of information retrieval. First, the development of the internet established a common platform for computer-based communication over diverse networks. Second, the web introduced a set of standards for publishing and retrieving information in a hyperlinked fashion. The web abolished the existing barriers between heterogeneous document publishing systems and is now considered the *de facto* medium for online information publishing. With the growth of the web as a vast information repository, the need for search in this space was evident from early on. As a result, the web had a leading role in bringing information retrieval concepts and techniques to the masses. Information search tasks that were once in the realm of professional documentation services are now executed on a daily basis by a large part of the population.

Generic web search engines like Google are typically found among the most visited sites on the web. This fact demonstrates the significance of search as a principal tool in information access. Moreover, it confirms the success of the theories and techniques developed in the area of information retrieval to address complex problems. Given the universality of web search, this is a good real-world example to illustrate how an information retrieval system works. Figure 1.1 outlines the main components and processes of a web information retrieval system. Web search systems have a distinctive component, identified as *web crawling*, that handles the collection of web documents. Given the heterogeneous and dynamic nature of the web, this module faces important technical challenges — e.g. find new documents or changes to old documents, detect duplicates, deal with document encodings, cope with growth, etc. The web crawler keeps an up-to-date internal copy of the documents gathered from the web. These documents are then parsed by the indexer module to prepare a data structure, the *inverted index*, designed for fast retrieval. The most important piece of this architecture is the *scoring & ranking module* that takes a user query and returns a list of documents ordered by relevance. The final weight of a document for a given query is given by a formula that combines a large number of independent *signals*. These signals can be *query-dependent* or *query-independent*. The former depend on the query, i.e. the docu-

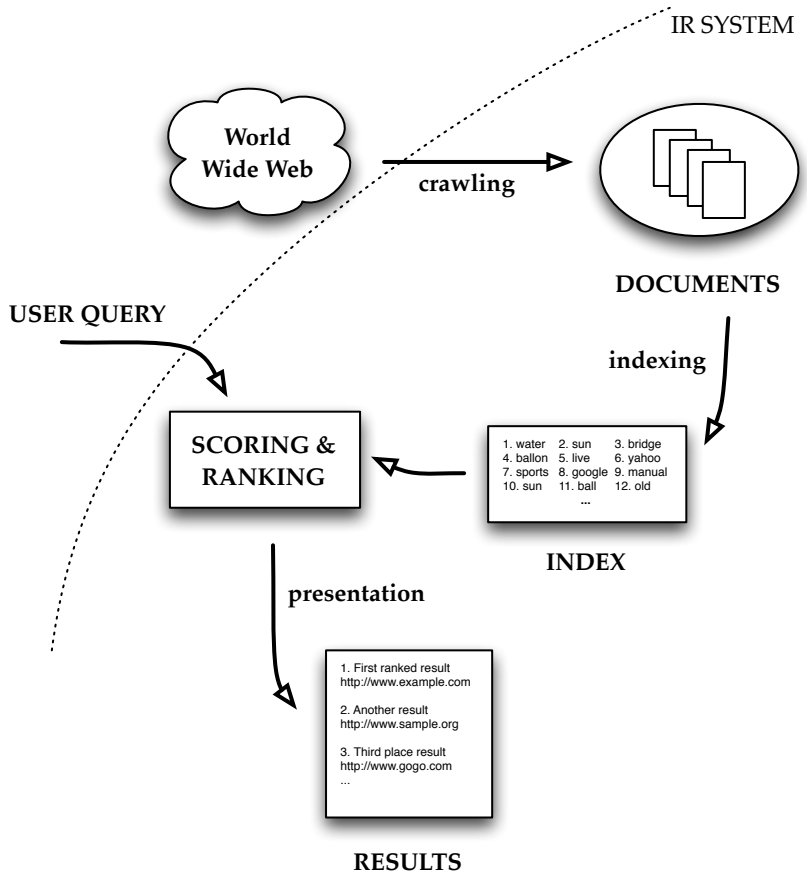

Figure 1.1: Outline of a web information retrieval system.

ment's score varies with each query. An example of a query-dependent signal is the *tf-idf* weighting scheme mentioned previously. On the other hand, the value of query-independent signals is fixed for each document; an example being the number of links pointing to the document. This value is determined regardless of the query being considered. The number of possible signals is vast and diverse. It is commonly mentioned that Google uses more than 200 independent signals in its non-disclosed ranking formula.

Web search engines are a well-known example of a real-word application of information retrieval that builds upon the many years of research in this field. Many more examples of the pervasiveness of information retrieval can be found in our everyday activities, for instance: searching for contacts in a mobile phone, using voice-activated services over phone, or simply searching for a file in a personal computer. Despite

its success and maturity, the field of information retrieval is still a very active area of research, as new problems are continuously pushing the limits of current approaches. Consider, for instance, the technical challenges faced by a web search engine when dealing with the size of the web and its continuous growth. A recent study reports that the size of the internet doubles every 5.32 years [88].

1.2 Time-Dependent Collections

The classic setting in information search, commonly defined as *ad-hoc retrieval*, deals with an established number of fixed documents and a varying set of queries (i.e. information needs). For each query, the system constructs a list of documents found on the collection that are considered to be relevant. In many scenarios, this is still a valid assumption, for instance when searching over frozen documents or collections, such as science abstracts or historic magazine archives. However, this is rapidly becoming the exception since most collections of documents change over time. A noteworthy example of a very dynamic collection is the World Wide Web. The web exhibits strong dynamic properties both at the document level (e.g. document revisions) and at the collection level (e.g. additions or removal of documents).

The dynamic characteristics of the web have been studied in detail, predominantly in the context of web crawling. Web crawling is a critical and expensive task where efficient resource allocation is decisive. A good understanding of document's creation, update or deletion rates is of key importance to a proper scheduling of the operations. With this information, web search engines have to estimate when to re-visit a previously crawled page and also estimate when to re-crawl a domain in search of new documents. However, this understanding of *web dynamics* has had little application in the context of document weighting and ranking. Information retrieval systems normally treat dynamic collections as static collections that are regularly replaced (i.e. fetched).

Baeza-Yates et al. [7] draw an interesting analogy between web search and the task of an astronomer watching the sky. Web search engines take periodic snapshots of the web and then use the information available in these snapshots to weight web documents. Like an occasional observer of the sky, classic web information retrieval techniques do not take into account the motion of the web. Studying the impact of information dynamics in information retrieval is on the basis of this work.

1.3 Proposed Thesis

As noted in the previous section, a fundamental assumption in most approaches is the existence of a static underlying collection of documents that is regularly updated. Classic systems view the collection as fixed regardless of previous changes to the documents. The removal of this assumption opens numerous research opportunities in the study of the dynamic side of document collections. One of our initial findings was that time is ubiquitous in information retrieval settings, it pervades all tasks and stages of the process. As a result, this is a comprehensive dissertation that deals with various aspects of information retrieval. Our thesis can be outlined as follows:

> The dynamic facet of document collections can be explored to improve current information retrieval tasks.

Additionally, it is important to stress that our thesis is focused on improving the response to current information needs by using temporal information in current collections, not retrieving or finding historical information. To illustrate the range of problems that we are tackling, consider the following exploratory questions:

- Is it possible to obtain rich temporal features from documents?
- Has the history of a document impact on its current importance?
- Is the evolution of scoring features a stronger signal than snapshot observations of those features?
- Is a document's past content relevant to present ranking?

Addressing these questions required work in different information retrieval areas, more specifically in feature characterization, query analysis, link analysis and keyword extraction. The web was considered an obvious choice as the primary document collection for our investigations. The importance of the web as a communication medium, its dynamic nature, and the easy access to its data justify this choice. The web was used as a central document collection, both as a whole and as parts by focusing on specific components like wikis and blogs. Wikis and blogs have an important characteristic for our goals — most documents contain explicit timestamps identifying the date of each revision.

1.4 Summary of Contributions

This work introduces several contributions to the topic of information retrieval in dynamic collections. Overall we have provided concurring evidence that supports the proposed thesis — the use of temporal information improves current information retrieval tasks. In this dissertation we present several original and independent contributions related to various stages of the information retrieval process in the context of dynamic collections. These contributions are outlined next.

We present a **characterization of the use of temporal expressions** in generic search tasks. Based on a large sample of web searches, we have found that temporal expressions are used in approximately 1.5% of all queries. Moreover, we have found that these expressions are primarily related to current events and more frequent within some topics, specifically in automotive, sports and news.

We started our investigation by studying the possible sources of temporal information in a dynamic collection. We **propose a taxonomy to classify features containing time-dependent data** according to two axis in the context of the web, namely document-based evidence and web-based evidence. On the web, a document's last update date can be determined by inspecting its HTTP headers. However, many web servers do not provide reliable timestamp information. In this context, we have **proposed a technique to improve the quality of temporal information** by considering a document's neighborhood. With this approach we are able to increase the amount of temporal evidence available for each document.

Looking at the link activity in a large collection of web documents, we observe that the distribution of citations to a document over time contains valuable information. We **present evidence showing that revision dates are strongly related to each document's topic**. For instance, a peak in a document's revision activity suggests the existence of an important event. Based on several examples we show that time-agnostic algorithms are unable to capture the correct popularity of sites with high citation activity. For this reason, we study the impact of a time-dependent scoring function in a standard citation-based rank, making it more sensitive to recent events. Results show that **a time-aware authority estimation algorithm can produce better estimates of document relevance by taking into account the progression of citations over time**.

Finally, we study the evolution of document content over time and show that **there is a rapid progression in a document's text towards a stable version**. More important,

we demonstrate that **by having access to the revision history of a document, keyword extraction can be significantly improved**. We propose measures that outperform the classic term frequency weighting approach. Finally, we find that **relevant terms that are otherwise inaccessible, can be extracted by using methods that skim through a document's history**.

1.5 Original Publications

The work presented in this dissertation was originally published in research papers in several peer-reviewed international conferences. These papers are listed here for reference, ordered by publication date.

1. Sérgio Nunes. **Exploring Temporal Evidence in Web Information Retrieval**. *BCS IRSG Symposium: Future Directions in Information Access 2007*. August 2007. Glasgow, Scotland.
2. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **Using Neighbors to Date Web Documents**. *CIKM'07 9th annual ACM International Workshop on Web Information and Data Management 2007 (WIDM'07)*. November 2007. Lisbon, Portugal.
3. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **Use of Temporal Expressions in Web Searches**. *30th European Conference on Information Retrieval 2008 (ECIR'08)*. March 2008. Glasgow, Scotland.
4. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **WikiChanges - Exposing Wikipedia Revision Activity**. *4th International Symposium on Wikis (WikiSym 2008)*. September 2008. Porto, Portugal.
5. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **FEUP at TREC 2008 Blog Track: Using Temporal Evidence for Ranking and Feed Distillation**. *17th Text REtrieval Conference (TREC 2008)*. November 2008. Gaithersburg, Maryland, USA.
6. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **Using Temporal Evidence in Blog Search**. *ECIR'09 Workshop on Information Retrieval over Social Networks*. April 2009. Toulouse, France.
7. Sérgio Nunes. **Sources of Temporal Web Evidence**. *Informer*, Nr. 30, pp. 7-8, ISSN 0950-4974. IRS Group, British Computer Society. Spring 2009.
8. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **Improving Web User Experience with Document Activity Sparklines**. *EPIA'2009 - 14th Portuguese Conference on Artificial Intelligence*. October 2009. Aveiro, Portugal.

9. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **FEUP at TREC 2009 Blog Track: Temporal Evidence in the Faceted Blog Distillation Task**. *18th Text REtrieval Conference (TREC 2009)*. November 2009. Gaithersburg, Maryland, USA.
10. Sérgio Nunes, Cristina Ribeiro and Gabriel David. **Term Frequency Dynamics in Collaborative Articles**. *10th ACM Symposium on Document Engineering (DocEng2010)*. September 2010. Manchester, United Kingdom.

1.6 Dissertation Outline

This dissertation is organized in four major parts: (1) background review, (2) domain characterization, (3) contributions and evaluation, and (4) conclusions and future perspectives. Each part and its corresponding chapters are outlined below.

Part 1: Background Review. The first part includes a broad survey of previous work.

Chapter 2: "Retrieval in Dynamic Collections" (p. 11) reviews and organizes the research in this area in three categories depending on the main source of temporal evidence being used: link-based approaches, content-based research, and metadata-based research.

Part 2: Domain Characterization. The second part presents a characterization of the problem's domain.

Chapter 3: "Temporal Features on the Web" (p. 23), the possible sources of temporal information in a dynamic collection are identified and arranged in a taxonomy. These sources are organized in two main groups, namely document-based features and web-based features.

Chapter 4: "Temporal Expressions in Web Searches" (p. 29) presents a study on the use of temporal expressions in search queries is documented. It is important to highlight that our thesis is focused on generic search queries, either including temporal expressions or not. This study provides an overall impression on the direct use of temporal expressions in users' search queries.

Part 3: Contributions and Evaluation. The third part of this dissertation includes the original contributions that support the central thesis. Each contribution is organized as an individual chapter where a problem is described and the proposed approach is discussed and evaluated.

Chapter 5: "Using Neighbors to Date Web Documents" (p. 35) includes the description of a technique that explores a document's neighborhood to better estimate the creation date of each web page. Results show that it is possible to use a document's vicinity to estimate its most recent update date.

Chapter 6: "Link Authority over Time" (p. 49) presents a study on the impact of time in link-based ranking algorithms. Using a large collection of blog posts, a traditional link-based ranking algorithm is compared with a temporally biased alternative. Results show that including temporal information in authority estimation algorithms improves the quality of the rankings.

Chapter 7: "Content Dynamics in Retrieval" (p. 67) studies the importance of a document's terms in face of each document's revision history. We propose two alternative measures for term extraction that outperform the classic term frequency measure.

Part 4: Conclusions and Future Perspectives. The fourth part presents a summary of the dissertation.

Chapter 8: "Conclusions and Future Work" (p. 95) reviews and summarizes the main contributions of this work and discussed several research directions for future work.

Appendixes. Two appendixes are included in this dissertation.

Appendix A: "WIKICHANGES" (p. 109) presents the web-based system developed to extract and draw citation profiles to Wikipedia articles.

Appendix B: "TREC Blog Track Participations" (p. 115) consists of a detailed description of our participations on the TREC Blog Track. Given the fact that blogs are rich in temporal information, we have tried to explore this characteristic in this track's tasks. However, although positive improvements were observed, the results obtained in these participations were below our expectations.

Chapter 2

Retrieval in Dynamic Collections

Previous work in information retrieval over dynamic collections has been almost exclusively focused on web documents. The dynamic properties of the web have been explored to improve different retrieval tasks. Research in this area can be organized in three categories depending on the main source of temporal evidence being used. Link-based approaches uses links, both in-links and out-links, in a temporal context to refine information retrieval. Content-based research examines documents' content from a temporal point of view. Finally, metadata-based research uses evidence gathered from external sources.

This chapter provides an overview of the state of the art in information retrieval over dynamic collections. We start with a brief survey of papers that highlight the strong dynamic nature of the World Wide Web. Then, previous work is organized according to the characteristics of the temporal properties used. The three main approaches identified are presented in independent sections.

2.1 Introduction

The classic setting in information retrieval involves a fixed collection of documents and a stream of arriving information needs, typically as text queries. Traditional models and theories in this field have adopted this point of view. However, most document collections have a pronounced dynamic nature. In many cases, documents are added to or removed from the collections over time. As an example, consider a patent repos-

itory where new applications are processed and stored. A more elaborated scenario can be found on the web, where documents' content also changes over time. In this case, the evolving nature of the collection is evident both at the structural level and at the document level. This generic setting — where documents are added to a collection over time — has been modeled as an *information filtering* problem. In this approach, a predefined set of rules is compared against a stream of arriving documents. Information filtering systems are commonly used to categorize or extract information from incoming documents [40]. SPAM detection mechanisms for email are an example of this type of systems. Contrary to information retrieval systems, the collection is viewed as inherently dynamic and the information need tends to be static over time.

The study of the temporal properties of a collection for information retrieval tasks, such as ranking, is an interesting and still largely unexplored topic. Although the web is not the only dynamic document collection available, it has been used as a reference in many of the previous works in this area. Its rich temporal facet and indisputable popularity explains this choice. There are several known difficulties in using temporal information from the web, most notably in obtaining reliable information and accessing historic information. Some of these problems are investigated in this dissertation. It is important to restate that our focus is on the use of temporal information available in the collection for current retrieval tasks (e.g. ad-hoc search). As opposed to search with temporal restrictions, the use of temporal information for interface improvements [4] or search in web archives [17].

Two specific types of web sites tend to be used recurrently in this area, specifically blogs and wikis. This can be understood by considering the fact that temporal information is explicitly available in these sites — in blogs each post has a precise timestamp, while in wikis the revision history of each document is kept by the system. The Wikipedia in particular is a controlled environment appropriate for rigorous, replicable research. Also worth noting is the fact that Wikipedia, more specifically its graph evolution, has many similarities with the web [22]. On the other hand, blogs are much more dynamic than the conventional web and have a somewhat different behavior [25]. More precisely, the lifetime of blogs' content (i.e. links and text) is short and links have a strong temporal locality.

Table 2.1: Overview of studies in web dynamics.

Study	**Size (∼pages)**	**Duration**	**Frequency**
Cho et al. (2000) [26]	720,000	18 weeks	Daily
Brewington et al. (2000) [20]	2,000,000	30 weeks	Daily
Koehler (2004) [53]	343 sites	325 weeks	Weekly (with gaps)
Ntoulas et al. (2004) [67]	150,000,000	52 weeks	Weekly
Fetterly et al. (2004) [36]	151,000,000	11 weeks	Weekly
Kim et al. (2005) [50]	34,000 sites	14 weeks	Two-day intervals
Grimes et al. (2008) [39]	23,200	n/a	Hourly
Adar et al. (2009) [2]	55,000	5 weeks	Hourly and sub-hourly

2.2 World Wide Web Dynamics

The dynamic nature of the web has been studied with varying degrees of detail. Overall, the web is considered a very dynamic and heterogeneous source of information. Using temporal evidence to improve information retrieval on the web is an area with growing popularity. First studies explored the temporal dimension with the main goal of reducing costs by focusing on crawler scheduling to optimize resource allocation (e.g. Cho and Garcia-Molina [26], Brewington and Cybenko [20]). The large majority of research in this topic is centered on identifying change and creation patterns to better allocate the available resources for crawling. In this context, the optimal situation is to only fetch a web page immediately after it has changed, and never download a page that has not changed since the last crawl. More recent works have looked at web dynamics as a possible source of evidence to improve other web information retrieval tasks. As an example, Adar et al. [2] study the dynamics of the pages that are visited, and suggest the use of a web page's stable content to improve document ranking. A summary of the more significant studies in web dynamics is presented in Table 2.1. A more detailed description of some of these works is discussed next.

Cho and Garcia-Molina [26] conducted one of the first experiments to study the dynamic nature of the web and its implications for the design of incremental crawlers. During a period of 4 months, a total of 720,000 pages were crawled from 270 popular web sites on a daily basis. Changes were detected by computing checksums for the whole content of each page. With this method, it was possible to determine the frequency of change but not the degree of change. Altogether, it was found that 40% of all web pages changed within a week and that, after 50 days, half of the pages had

changed. However, it is important to note that, due to the rough change detection technique used, even the smallest changes are considered. Brewington and Cybenko [20] used data gathered from an university clipping service to study the characteristics of web dynamics for the design and maintenance of fresh indices. This service downloaded approximately 100,000 documents a day during a period of 7 months. In total, more than 2 million web documents were analyzed. Collected data was then used to analyze frequency (*how often*) and degree of update (*how much*). Overall, approximately 56% of the pages never changed during this period, while 4% changed in all observations. In addition, the authors suggested that changes to web pages follow a *Poisson distribution*. It is important to note that, contrary to other studies, the pages gathered for this article were not randomly sampled from the web. Also, a sampling bias is introduced given the fact that daily downloads were limited to 100,000 documents. The thesis that changes to web pages follow a Poisson distribution was recently challenged by Grimes and O'Brien [39]. In this paper, the authors analyze a subset of fast moving pages extracted from a random sample from the Google crawl. Although results were inconclusive, relatively few pages in their sample followed a strict Poisson model. An important finding is that update volume depends on the local time and day of the week.

Ntoulas, Cho and Olson [67] analyzed the evolution of both content and link structure of web pages, specially focusing on aspects of potential interest to search engine designers. Data was gathered from weekly crawls of 150 web sites selected from the top ranked sites in Google Directory. Crawls were performed over the course of a year, starting in late 2002. Each weekly snapshot had an average size of 65 GB, resulting in a final dataset with more than 3.3 TB. Both change frequency and change degree were analyzed. The shingling technique was used to quantify the degree of change. Authors found that only 20% of web pages last one year and that, after a year, 50% of the content on the web is new. Link structure was observed to be more dynamic than content — 25% new links are created every week, while only 8% new pages are created and 5% new content is produced. Evaluations based on the *tf-idf* metric revealed that most new content comes from the creation of new pages, instead of updates to the existing ones. Content shift in existing pages is small, i.e. current and previous versions of a page exhibit small *tf-idf* differences. It was also found that frequency of change was not a good predictor of degree of change. However, current degree of change was correlated with previous degree of change. This is still one of the largest studies conducted addressing the dynamic nature of the web and one of the few to address web page creation

rates.

In a similar large scale experimental study, Fetterly et al. [36] examined the evolution of web pages. A sample with more than 150 million URL was produced using a breadth-first crawl strategy. Next, each URL was crawled weekly over the period of 10 weeks (between November and December 2002). The full text was retained for only 0.1% of all downloaded documents and changes in these documents were measured using 5-word shingles. Changes across top-level domains were observed and it was found that there is a strong correlation between frequency of change and a document's top-level domain. In contrast, there is a much weaker relationship between degree of change and top-level domain. It was observed that larger pages changed more often and more frequently than smaller pages. Further exploration revealed that most of these changes occurred in the attributes of URL values. Manual inspection showed that session identifiers were responsible for most of these changes. Overall, most changes detected were either in the document's markup or of trivial nature. Concurring Ntoulas et al.'s study, past frequency of change was found to be a good predictor of future change. However, authors alert that "there are pages that change only to a small degree, but where the changing portion is the most salient component of the page". Thus, both frequency and degree of change should be combined to determine crawling frequency.

The longest running study was conducted by Koehler [53]. In this study, a random sample of 343 web sites was crawled periodically over the course of 6 years and 12 weeks. The author concludes that the web is not a stable medium for long-term content preservation. Nonetheless, it is important to note that recent studies have suggested that lasting contents tend to be referenced by different URL during their lifetime [38]. In a recent paper, Gibson et al. [37] studied the volume and evolution of web page templates. They found that 40% to 50% of the content on the web is template content and that, over the last 8 years, this ratio has doubled. Additionally, using Ntoulas et al.'s data combined with crawls from the Internet Archive, the degree of change was reevaluated after template removal. Authors found that *templated* and *detemplated* regions of a page exhibit similar change degree over time.

More recently, Adar et al. [2] conducted a detailed study on the dynamics of web content in pages that are actually visited. Contrary to other analyses, the sample of web documents used in this work is not random. Instead, the authors have selected a set of roughly 55,000 pages based on information obtained from Microsoft's Live Search toolbar. As the authors note, the intention was to capture a sample that represented

pages that are actually revisited and consumed by users in different ways. This sample of pages was crawled hourly over 5 weeks. Additionally, a subset of fast-moving pages was identified and crawled with sub-hourly frequency. Compared to previous studies, this sample of pages displayed a higher change rate, with approximately 40% of pages changing nearly every hour. An important contribution of this work was the distinction between stable and dynamic content in web documents. The stable part of a document is defined as the content that remains the same over time. Using *change curves* plots, the authors showed that the stable content of a page tends to stabilize after a short period of time. As the authors suggest, this distinction between stable and dynamic content might be used to improve web document ranking by giving different weights to each type of content.

2.3 Link-Based Approaches

In *link-based approaches*, the time-dependent information used in the algorithms is obtained from the documents that point to or are pointed by each document. Berberich, Vazirgiannis and Weikum [19] propose an algorithm, named T-Rank, which extends PageRank to improve page ranking by exploring the freshness and activity of both pages and links. A temporal model for link analysis is proposed where each node and edge in the web graph is annotated with timestamps. These timestamps represent different types of events (i.e. creation, modification, deletion). Also, to capture the user's temporal focus, a window of interest is defined based on two timestamps that limit the relevant period. Given this temporal window of interest, freshness is defined by a linear function that is maximum if timestamps occur within the user defined period and that decreases linearly if they occur within the tolerance intervals. The activity of an object (page or link) is defined as the sum of all modifications occurred within the time window. The T-Rank algorithm extends PageRank by modifying the underlying probabilities of the random walk to favor certain nodes. The transition probability from node x to node y is defined as a weighted combination of the freshness of the target node y, the freshness of the link between x and y, and the average freshness of all incoming links of y. On the other hand, the random jump probability of a target page x is a weighted combination of the freshness of x, the activity of x, and the average freshness and average activity of the pages that link to x. Just like PageRank, T-Rank values are computed using the power iteration method. The quality of this proposal was assessed using the DBLP corpus, Amazon's products pages and user studies. Overall,

users tended to prefer the ranks produced with the time-aware approach.

Yu et al. [87] argue that popular ranking algorithms (e.g. PageRank and HITS) miss an important dimension of the web by not considering its temporal characteristics. Since these algorithms rely largely on the total number of accumulated links, older pages tend to be favored. The PageRank algorithm is adapted by weighting each citation according to the citation date. The new technique, dubbed TimedPageRank, uses an exponential decay function to weight citations. An aging factor is also included in the final formula so that nodes' score decline linearly with time. Since the evaluation was based on a bibliographic dataset, the algorithm is further refined by also including TimedPageRank scores for authors and journals. Preliminary results based on a catalog of physics publications show a better overall performance over a set of 25 queries.

2.4 Content-Based Approaches

In *content-based approaches*, the document is the sole source of temporal evidence used by the algorithms. Jatowt and Ishizuka [46] address the problem of web document summarization by exploring previous versions of the documents. This temporal web page summarization approach analyzes the content retrieved from temporally distributed versions of a single web document to produce a summary of the main concepts addressed over a given time period. This method is applicable to documents that are frequently updated. The reasoning behind this approach is that, by using historical versions of a document, more content becomes available, in turn leading to improvements to the final summary. To observe each term's evolution, regression analysis is used to summarize the relationship between time and the frequency of a term. A formula combining the slope of the regression line, its intercept and the variance of term frequency is used to rank terms. Finally, sentences are selected and ranked based on their average term scores. However, this proposal was not thoroughly evaluated, only a single example of a manually selected homepage is used to illustrate the benefits of the strategy. This technique was later expanded to address multi-document summarization [47].

Liebscher et al. [55] analyze content over time to identify rising or falling terms. The authors try to identify lexical dynamics so that current search results can be amplified or dampened. Dissertation abstracts and discussion board posts were used in an experimental setup. Each large corpus was split into multiple independent corpora, each

associated with documents occurring within a specific time interval. Temporal trends were then computed using simple time series analysis techniques based on linear models. The most dynamic topics were identified by comparing the slopes of the regression lines. The authors argue that this information can be exploited to improve information retrieval. The argument is that previously popular terms, belonging to documents that capture more fundamental aspects of a topic, should be amplified. On the other hand, currently popular terms that were once rare should be dampened so that new topics are not overemphasized. Based on several anecdotal examples from the dissertation abstracts and discussion boards, the authors show that terms decline or rise over time as expected. For instance, in a set of approximately 5,000 Ph.D. and Masters thesis in artificial intelligence published between 1986 to 1997, one of the top rising bigrams was *neural networks*, while top falling bigrams included *expert system(s)* and *rule based*. This work is related to the field of Topic Detection and Tracking (TDT) [3], an active community within the Information Retrieval discipline that deals with problems such as first story detection, topic tracking and story segmentation.

This line of research was further explored by the same authors to tackle text categorization problems [56]. In this context, the authors identified terms that are *temporally perturbed*, i.e. terms whose distribution over fixed categories varies over time. The paper refers to *schwarzenegger* as an example of a temporally perturbed term, since over time it has steadily moved from the entertainment to the politics category. Based on these findings, a feature modification technique was proposed to account for these lexical changes. Experiments performed using an ACM document collection show that this approach clearly improves the retrospective categorization task.

The work by Elsas and Dumais [33] is one of the first to analyze the use of the temporal dynamics of document content to improve relevance ranking. Using a collection of top ranked web documents, the authors establish a relationship between content change patterns and document relevance. They observe that highly relevant documents are more likely to change than documents in general, both in terms of frequency and degree. Based on this finding, the authors propose two methods that improve document ranking by leveraging content change. In the first approach, a query-independent method, they find that favoring dynamic pages leads to performance improvements. In the query-dependent technique, it is shown that favoring a document's static content also results in performance improvements.

2.5 Metadata-Based Approaches

In *metadata-based approaches*, temporal evidence is gathered from sources other than the document content or its vicinity. Examples include HTTP headers and external services, such as PageRank values or repositories containing archived information. Berberich et al. [18] propose BuzzRank, a method that is complementary to time-agnostic ranking algorithms. By analyzing time series of importance scores (like PageRank scores), BuzzRank identifies growth trends in these scores using generic growth models and curve fitting techniques. For instance, in a reported experiment using the DBLP bibliographic dataset, BuzzRank highlighted papers that, though not having the highest PageRank value, had the highest PageRank growth. Since PageRank scores over time are not directly comparable, graphs were normalized treating missing nodes as dangling nodes. The presented version of BuzzRank is very expensive both in terms of computing power and storage requirements since it needs to store the entire graph and perform PageRank computations for each time interval. The authors conducted a simple experiment using the DBLP database to evaluate the proposed approach. They found that BuzzRank produces results different from those obtained with PageRank, and tends to highlight topics that were hot during the specified periods.

Amitay et al. [6] disclose significant events and trends by observing HTTP headers of web documents. The Last-Modified field is used to approximate the age of the page's content. Using this information to timestamp web resources, several interesting applications are explored. It is clearly shown that real life events can be exposed mainly due to what authors call *fossilized content*. In a nutshell, three phases are involved in the discovery of fossilized content. First, by issuing one or several queries to public search engines, a small topical collection of web documents is assembled. These queries are manually selected so that they characterize the desired topic. Then, a second collection is built based on the links that point to the URL in the first collection. Finally, the Last-Modified values of the pages in the second collection are gathered and a histogram is plotted. The authors report several experiments where clear patterns are visible. While exploring the notion of timestamped web resources, the authors introduce the concept of timely authorities, as opposed to simple link-based authorities. This idea is illustrated with the adaptation of the HITS [52] and SALSA [54] algorithms, by adjusting vertices weights to include a time-dependent bonus. To demonstrate the different results obtained with this approach, two authority rankings are computed for two queries, namely a basic authority ranking and a timely authority ranking. In the second rank-

ing, results with a declining number of citations over time tend to be ranked in lower positions. On the contrary, resources with a large percentage of recent citations are ranked higher.

To address the problem of question answering on the web, Yamamoto et al. [85] use the Internet Archive [45] to assess the trustworthiness of user given statements. The rationale is that if a phrase has been continuously stated for a long time on the web, its reliability is higher. On the contrary, if occurrences of a sentence decline over time, its trustworthiness also declines. For each possible proposition, temporal profiles using a public web archive are produced representing the growth (or decline) of occurrences over time. Using time series analysis, these profiles are used to estimate the statement's accuracy. The effectiveness of this approach was illustrated using two time-dependent propositions selected by the authors. In both cases, the system was able to correctly determine the veracity of the statements.

Recently, Yang et al. [86] proposed TemporalRank, a new algorithm that combines the current PageRank score with an historic score for each web document. The historic score is a combination of previous PageRank scores computed using preceding snapshots of the web graph. This proposal is evaluated using five large sub-graphs based on snapshots of the web and a collection of user feedbacks collected from a toolbar software used by a commercial search engine. Results show that historic information directly improves the quality of the ranking. Moreover, its quality improves as more snapshots are included. This is one of the few works in this area that was tested over a collection of real web documents and evaluated with user judgements.

2.6 Summary

The World Wide Web is an extraordinary resource and one of the symbols of the modern information society. It is a vast repository of data and an increasingly important information source for millions of users worldwide. It is also an excellent example of a large dynamic collection of documents. Current research on web dynamics shows that the web is very active, exhibiting both high decay and high creation rates. Content is being created mostly through the production of new pages (rather than updating existing ones) and, once created, these pages exhibit very rapid decay rates. Although there is a strong connection between past update frequency and future update frequency, the correlation between frequency of update and degree of update is small. Page templates

represent approximately 50% of the content on the web. However, studies indicate that they have little impact on the change degree of a page's content, as regions with templates and without templates exhibit similar change degree over time. Finally, recent work has shown that web pages tend to have stable areas, containing content that remains unvarying over time.

The first works in web dynamics were almost exclusively focused on understanding document update rates to improve the crawling stage. The information derived from the evolution of content and structure is typically not included in most information retrieval tasks. Even simple temporal information, such as web page decay measures, have not been included in result ranking, as noted by Bar-Yossef et al. [12]. More recently, the use of temporal information has been investigated in the context of various tasks, such as result ranking [19], web mining [6], clustering [56], text classification [9], summarization [47] or question answering [85]. Given the pervasiveness of time and its significance to users, the proliferation of work in time-dependent features is inevitable. Based on an analogy first mentioned by Baeza-Yates et al. [7], we can state that it is not possible to understand a dynamic collection with a single snapshot in the same way that it is not possible to understand the universe by simply looking at the sky.

Chapter 3

Temporal Features on the Web

Sources of temporal information on the web can be organized in two main groups, namely document-based features and web-based features. The former encompasses all evidence extracted from individual documents, while the latter includes evidence obtained from the whole web.

This chapter presents an original taxonomy to organize potential sources of time-dependent features on the web. First, document-based features are structured according to the source of the data, more specifically by document's content, URL address and HTTP protocol. The same approach is adopted for web-based features, which are arranged by neighbors, external archives and web logs.

3.1 Document-Based Features

Document-based temporal evidence is obtained by exploring the characteristics of single web documents. These characteristics are limited in scope and, typically, user-generated. Thus, they are subject to direct manipulation and might be engineered. Document-based temporal evidence is classified according to three types: content, URL address and HTTP protocol.

3.1.1 Content

A web document's content might be explored in various ways to gather temporal evidence. Using natural language processing techniques, specifically information extraction methods, it is possible to identify words or expressions that convey temporal meaning (e.g. "today", "a long time ago", "October 18") and use these to date documents [81]. On the other hand, the evolution of a document's content through time can be viewed as a single temporal feature. This new feature can incorporate multiple metrics derived from well-known information retrieval concepts (e.g. term frequency, term count). Observing HTML markup code might also be of value to obtain temporal evidence (e.g. evolution of out-links, stability of terms). Also, content differences between versions of the same document can be viewed as document-based temporal evidence.

Research in finding near-duplicate web pages [43] has led to the development of multiple algorithms for this task. For instance, comparing content checksums (or signatures) of web documents is a simple but naïve technique that is rarely used. Very similar documents, differing only in a word, character or number, have disparate signatures. This scenario is very common on the web, where documents typically contain a timestamp or a simple visits counter. Alternatively, the *shingling technique*, proposed by Broder et al. [21], is one of the most popular algorithms for this problem. In a nutshell, documents are structured in sets of fixed length consecutive term sequences (or shingles) and then the similarity of the resulting sets is compared. This algorithm produces a single value that can be used to compute a similarity degree.

3.1.2 URL Address

All public web documents have at least one unique address. Occasionally, documents might have multiple addresses that resolve to the same web page. The different segments of an URL, namely host, path and search part, might be used as a source of temporal information. For instance, a current New York Times URL is structured in the following way — `http://www.nytimes.com/2009/10/19/world/asia/19books.html`. It is possible to derive the document's year, month and day of publication by parsing this URL. Based on this information, it would be possible to estimate the document's inception date.

3.1.3 HTTP Protocol

The Hypertext Transfer Protocol (HTTP) is an application level protocol used to request and transmit hypertext documents and components between user applications and online servers. Clients submit standard requests to servers identifying specific resources. Servers reply to each request, sending standard headers and, if available, the requested resource (body of the message). HTTP headers are comprised of fields, part optional and part required. Of interest is the Last-Modified field representing the date and time at which the resource was last modified. This HTTP header field might be used to add temporal information to a web resource, specifically a web document.

However, this field is not always available and, even when available, may not return a valid date. This erratic behavior is generally attributed to incorrectly configured web servers. Several independent studies have estimated that 35% to 80% of web documents have valid Last-Updated values [20, 6, 38]. Despite this problem, it is possible to explore this information with very positive results as described by Amitay et al. [6] (see Capter 5).

3.2 Web-Based Features

The entire web is an important source of information about individual web documents. In other words, multiple independent sources can be combined to produce information about a particular web resource. One distinct advantage of this approach is that it is hard to influence or tinker. On the web, a public document is integrated in a wider context. A document is connected to other documents and resources through a hypertext mesh (the document's neighborhood). On the other hand, in a web information system there are inherent components that are continuously gathering information, such as external information repositories or web server logs. As shown below, each of these components might be explored as a source of web-based temporal evidence.

3.2.1 Neighbors

Links are intrinsic to a hypertext system like the web. Web resources, or graph nodes, are connected by referencing other nodes or by being referenced by them. Inside this network it is possible to define the concept of a web document's neighborhood, the

set of nodes that point to a document or are pointed by it. Considering a document's vicinity it is possible to derive multiple features. Regarding temporal evidence, a document's in-links and out-links may be observed through time to reveal trends or unexpected patterns (e.g. link farms for spamming).

3.2.2 External Archives

There are several services with the mission of archiving public web data. The Internet Archive [45], a non-profit organization, is the leading initiative in this field. This group has been archiving the web since 1996 and its collection is open to public access. Typically, the data is collected through periodic web crawls and gathered using standard format archives. For each web resource found, multiple snapshots are collected through time. After acquiring old copies of specific web pages, content matching algorithms can be applied to produce temporal evidence. There are other services from where it is possible to extract dated copies of web resources, most notably search engine's caches and large web crawls available for research.

3.2.3 Web Logs

A web information system is a combination of multiple smaller subsystems that work together to provide a seamless experience to the users. These subsystems are very heterogeneous, ranging from a user's browser to complex HTTP server software capable of handling millions of requests per second. Most of these subsystems' activities are recorded in a persistent fashion, usually in flat file logs. Two types of web logs can be explored as sources for temporal evidence, namely query logs and access logs. Query logs record users search intentions, while access logs record users access to web resources. Both types are filled with temporal markup that can be used to annotate web resources. The major problem of web logs research is the private nature of the data, making these logs very difficult to obtain and raising significant privacy issues.

3.3 Conclusions

In this chapter we present a taxonomy to frame temporal features according to their genesis, more specifically we propose a classification based in two groups — document-

based and web-based features. Alternatively, these features can be arranged according to their *temporal order*. First order features are those that are direct sources of temporal information, such as the Last-Modified value in HTTP response headers, or temporal expressions in a document's content. Second order temporal features are obtained from the observation of standard (non-temporal) features over time, such as document's size evolution, or the number of citations over time. Considering the survey of related work presented previously in Chapter 2, it is possible to see that most works in this area have used web-based features as a source of temporal data. This is an expected result considering the fact that this class of features is easier to obtain. Document-based features tend to require a direct observation of the evolution of a web document, a process that demands important crawling and storage resources.

We opted to study the web in detail because of its importance, relevance and distinctive nature. The dynamic nature of the web, dissected in Section 2.2, is also a strong argument for this analysis. Although the web is a source of rich and complex temporal information, there are subgroups of web pages where this type of information is more abundant and more easily accessible, namely blogs and wikis. An example is the revision history typically available for each document in a wiki and its importance for studies focused on document dynamics. On the other hand, time is deeply entrenched in the typical structure of a blog, both in individual posts and in users' comments.

Chapter 4

Temporal Expressions in Web Searches

While trying to understand and characterize users' behavior online, the temporal dimension has received little attention by the research community. This exploratory study uses two collections of web search queries to investigate the use of temporal expressions. Using standard information extraction techniques we identify temporal expressions in these queries. We find that temporal expressions are used in 1.5% of all queries and are commonly related to current and past events.

This chapter describes the investigation conducted to understand the use of time-dependent expressions in user issued web search queries. First, our experimental setup is discussed, including details about the datasets and software libraries used to extract the temporal expressions. Then, a detailed analysis of the use of these expressions over time and across categories is advanced. Finally, these results are discussed in the last section.

4.1 Introduction

Query log analysis is currently an active topic in information retrieval. There is a significant and growing number of contributions to the understanding of online user behavior. However, work in this field has been somewhat limited due to the lack of real user data and the existence of important ethical issues [11]. The recent availability of large

datasets has specially contributed to a growing interest in this topic. On the other hand, temporal information extraction has reached a point of significant maturity. Current algorithms and software tools are able to extract temporal expressions from free text with a high degree of accuracy. This chapter contributes to the characterization of the use of temporal expressions in web search queries, by combining work from these two areas. Our main goal is to provide a better understanding of how users formulate their information needs using standard web search systems. Our focus is on a particular facet of this behavior, namely the use of temporal expressions. It is important to note that the central thesis of this dissertation is focused on generic search queries — i.e. our goal is to improve the retrieval of documents regardless of the use of temporal expressions in search queries. While not directly related to the central thesis, this study provides important insights on the direct use of temporal expressions by online users.

We found no previous work on the specific topic of identifying and characterizing the use of temporal expressions in web search queries. Thus, the related work presented here is divided in the two parent topics: *temporal expression extraction* and *query log analysis*. In recent years, Temporal Information Extraction emerged from the broader field of Information Extraction [81]. Most works in this field are focused on the study of temporal expressions within semi-structured documents [59]. In our study we apply these techniques to short text segments that represent information needs. A large part of the research in query log analysis has been devoted to the classification and characterization of queries. An example of a detailed work in this area is from Beitzel et al. [14]. In this work the authors perform a detailed characterization of web search queries through time using a large topically classified dataset. Our work differs from this since we are interested in how temporal expressions are used within queries to express information needs.

4.2 Experimental Setup

We used two publicly available datasets containing web search queries. The first dataset includes a collection of manually classified web search queries collected from the AOL search engine [15]. Each one of the 23,781 queries has been manually classified by a team of human editors using a set of predefined topics. The classification break down is summarized in Table 4.1. The second dataset is also from AOL [72] and includes more than 30 million (non-unique) web queries collected from more than 650,000 users

Table 4.1: Break down of the categories in the AOL dataset.

Category		**Category**	
Autos	2.9%	Personal Finances	1.4%
Business	5.1%	Places	5.2%
Computing	4.5%	Porn	6.0%
Entertainment	10.6%	Research	5.7%
Games	2.0%	Shopping	8.6%
Health	5%	Sports	2.8%
Holidays	1.4%	Travel	2.6%
Home & Garden	3.2%	URL	5.7%
News & Society	4.9%	*Misspellings*	5.5%
Organizations	3.7%	*Other*	13.2%

over a three month period. This dataset is sorted by user ID and sequentially ordered. For each request there is also information about the time at which the query was issued and, when users followed a link, the rank and the URL of the link. An important feature of this second dataset is the availability of the query issuing time, making possible the positioning of temporal expressions. For instance, we are able to determine the specific date of a search for *"a week ago"* because we have access to this information. However, and unlike the first dataset, this one isn't classified.

Temporal expressions were extracted from each query using free, publicly available, natural language processing software. First, text was tagged using `Lingua::EN::Tagger` [28], a part-of-speech tagger for English. The output was then redirected to TempEx [60], a text tagger that is able to identify a large number of temporal expressions. This tagger covers most of the types of time expressions contained in the 2001 TIMEX2 standard [35]. In Table 4.2 several examples of this process are presented, showing that TempEx is able to detect a wide range of temporal expressions (e.g. explicit dates, implicit dates, periods).

4.3 Use of Temporal Expressions

First, we investigate how temporal expressions are distributed within distilled web queries. For this task we used the first dataset since it is manually annotated with

Table 4.2: Examples of tagged search queries.

olympics 2004 olympics <TIMEX2 TYPE="DATE" VAL="2004">2004</TIMEX2>
easter 2005 <TIMEX2 TYPE="DATE" ALT_VAL="20050327">easter 2005</TIMEX2>
monday night football <TIMEX2 TYPE="DATE">monday night</TIMEX2> football
us weekly us <TIMEX2 TYPE="DATE" SET="YES" PERIODICITY="F1W">weekly</TIMEX2>

classes. The topics containing the higher percentage of queries with temporal expressions are: Autos (7.8%), Sports (5.2%), News & Society (3.9%) and Holidays (2.5%). Examples of queries containing temporal expressions are: *"1985 ford ranger engine head"* (Autos), *"chicago national slam 2003"* (Sports) and *"los angeles times newspaper april 1946"* (News & Society). Manual inspection reveals that the higher number of temporal expressions in the Autos class is mostly due to searches for vintage cars.

In a second experiment, we analyzed the overall distribution of temporal expressions in web search queries. Our first finding is that the use of these expressions is relatively rare. In the first AOL dataset the total number of queries including temporal expression was 347 (1.5%). Remarkably, on the second dataset, we found 532,989 temporal expressions resulting in an equal percentage of 1.5%. Removing duplicate queries results in a small increase in these values, specifically 1.6% for the first dataset and 1.9% for the second. To evaluate the quality of the TempEx tagger when applied to web search queries, we manually classified a random subset of 1,000 queries from the large AOL corpus (including duplicates). Within this subset, the automatic process found 14 temporal expressions and the manual method found 19 temporal expressions. Standard IR measures were computed: accuracy (0.99), precision (0.92) and recall (0.63). The low recall value indicates that the tagger is being conservative, missing some temporal expressions. The non-parametric McNemar test [64] was used to compare the two classifications. The test confirms that the automatic classification is equivalent to the human classification ($p > 0.05$).

Restricting our analysis to the second AOL dataset, we performed additional measurements. Since query issuing time is available, we are able to precisely date a large fraction of the temporal expressions found. Using this information we measured the number of expressions referencing past, present and future events. TempEx automatic-

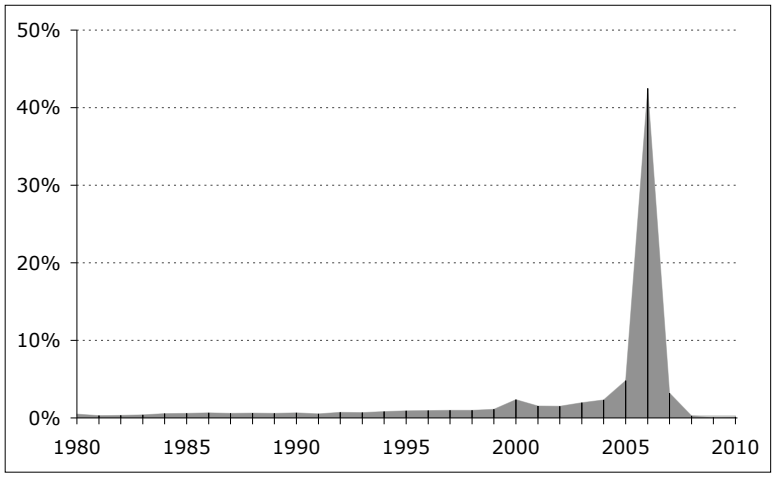

Figure 4.1: Years Mentioned in Temporal Expressions

ally detects some of these expressions. This automatic tagger was able to identify generic references to the past (e.g. *"once"*, *"the past"*) (1.25% of the temporal expressions), to the present (e.g. *"now"*, *"current"*) (4%) and to the future (e.g. *"future"*) (0.82%). The vast majority of temporal expressions (94%) do not include explicit references like these. We've extracted the year from all dated expressions and counted the occurrences. Taking into account that all queries were issued between March and May 2006, we see that the majority of dated temporal expressions is related to current events. The frequency distribution is positively skewed with a long tail toward past years as shown in Figure 4.1. Summarizing, in all temporal expressions identified, 42.5% indicate a date from 2006, 49.9% include a reference to dates prior to 2006, and 4.2% are from dates after 2006.

To better understand which temporal expressions were being used, we manually inspected a list of the top 100 more common expressions. These expressions account for slightly more than 80% of the queries containing temporal expressions. We grouped all references to a single year and to a single month in two generic expressions (i.e. *<Year>* and *<Month>*). The top 10 expressions used in queries are: *<Year>* (45.7%), *easter* (5.6%), *daily* (5.4%), *<Month>* (4.6%), *now* (2.3%), *today* (2.1%), *mothers day* (1.8%), *current* (1.2%), *christmas* (1.1%) and *weekly* (0.8%). It is important to note that seasonal res-

ults (e.g. "Easter") are artificially inflated in the 3 month of data (March to May).

One of our first hypothesis was that temporal expressions were regularly used to fine-tune queries. To investigate this hypothesis we did a rough analysis on query refinements within the AOL dataset. Our algorithm is very simple and only identifies trivial query reformulations. In a nutshell, since the dataset is ordered by user and issuing time, we simply compare each query with the previous one to see if there is an expansion of the terms used. For instance, *"easter holidays"* is considered a reformulation of *"easter"*. With this algorithm we found 1,512,468 reformulated queries (4.2%). We then counted the presence of temporal expressions in this subset and verified that only 1.4% of these queries contained temporal expressions.

4.4 Conclusions

Based on the previous experiments, we find that the use of temporal expressions in web queries is relatively scarce. Using two different datasets, we've found that temporal expressions are used in approximately 1.5% of the queries. We speculate on four reasons that might explain this situation: (1) information needs of web users are mostly focused on current events or time-independent topics; (2) users are generally happy with the results obtained using short text queries [74]; (3) users resort to more advanced interfaces when they have dated information needs; (4) there is an implicit idea that web pages are regularly updated and only the most recent version of the content is important. Investigating these hypotheses is left for future work. Focusing on the small subset of temporal expressions extracted, we've found that most temporal expressions reference current dates (within the same year) and past dates, exhibiting a long-tailed behavior. Future dates are rarely used. Finally, we've shown that these expressions are more frequently used in topics such as: Autos, Sports, News and Holidays.

Although temporal expressions appear in only a small fraction of all queries, the scale of the web translates this percentage into a large number of users. Temporal expression extraction might be used in public search engines to improve ranking or result clustering. We think that, as the web grows older, and more content is accumulated in archives (e.g. Internet Archive) the need for dated information will rise [5]. Search engine designers can respond to this challenge by incorporating temporal information extraction algorithms or by developing specialized search interfaces.

Chapter 5

Using Neighbors to Date Web Documents

Time has been successfully used as a feature in web information retrieval tasks. In this context, estimating a document's inception date or last update date is a necessary requirement. Classic approaches have used HTTP header fields to estimate a document's last update time. The main problem with this approach is that it is applicable only to a small part of web documents. We evaluate an alternative strategy based on a document's neighborhood. Using a random sample of web addresses, we study each document's links and media assets to determine its age. If we only consider isolated documents, we are able to date 52% of them. Including the document's neighborhood, we are able to estimate the date of more than 86% of the same sample.

This chapter presents an experiment conducted to evaluate the richness of a web document's neighborhood to estimate last update dates. The experimental setup is first described, with an emphasis on the characterization of the data used in this study. A web document's neighborhood is then analyzed in terms of incoming links, outgoing links and media assets. Based on real word examples, the proposed approach is compared with the traditional method. In the final section we summarize the main findings and suggest future research avenues.

5.1 Introduction

In information contexts, time is viewed as an important feature [32]. Contrary to the number of incoming links to a page or the term frequency of a web document, a document's age is an easily perceivable characteristic. The concept of *new* and *old* is well understood by users. Currently, in state-of-the-art information retrieval systems, time plays a small role in serving users' information needs. Web documents tend to be treated equally despite their age. However, as the web grows older, we speculate that time will play a significantly higher role in the overall web information context.

One way to explore this information would be to include temporal evidence in the ranking of web resources. A first step towards this *temporal-sensitive ranking* is the task of dating web resources. Classic approaches have used document-based features to achieve this goal. HTTP is a stateless request/response protocol. Clients submit standard requests to servers identifying specific resources. Servers reply to the request sending standard headers and, if available, the requested resource (body of the message). HTTP headers include a Last-Modified field representing the date and time at which the resource was last modified. This field has been used with success as an indicator of a document's date in real world applications. However, this approach has an important drawback — a very high percentage of the response headers do not contain reliable information. Most servers do not return any information at all, while others always return the current date or simply a wrong value [27]. Previous studies have reported percentages of valid Last-Modified values ranging from 34% to 79% of the tested URLs.

In this study, we explore ways to improve these values by looking at web-based features, namely neighbors of the original documents. The main rationale is that, by looking at web resources connected to the original document, more information can be gathered, improving the chances of retrieving valid HTTP values. In a nutshell, our hypothesis is that we can improve the dating of a single document if we look at its vicinity. This assertion has an implicit assumption — connected web resources tend to have similar update patterns. The overall context is illustrated in Figure 5.1. Observing the web from the point of view of a single document, we consider three types of web resources: documents containing links to the selected document (incoming links), documents pointed to by the selected document (outgoing links) and, finally, the media assets associated with the document (e.g. images). For each of these sets, the HTTP headers are available. We expect to find more reliable information in the header of

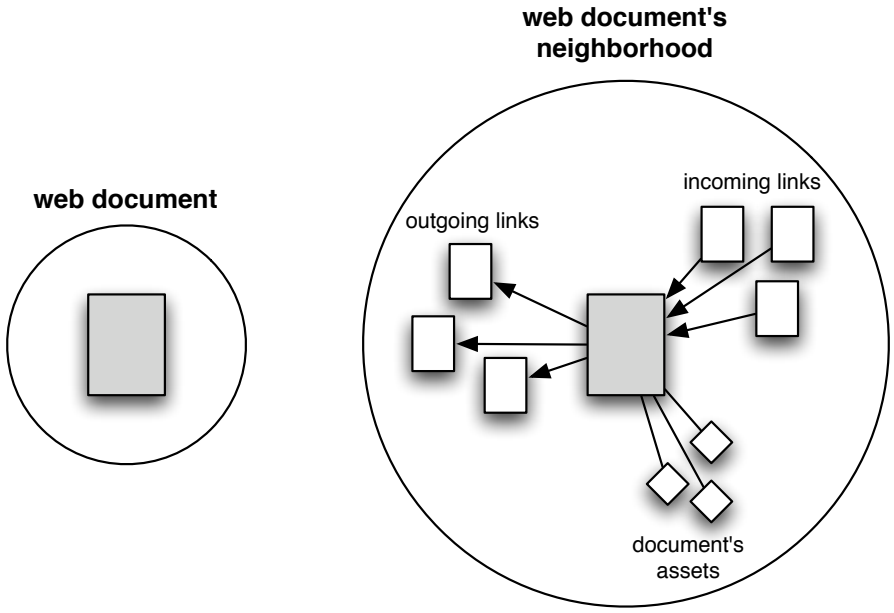

Figure 5.1: Using Neighbors to Extract Data.

static resources such as images than in the documents including them, mostly dynamic resources.

Web document's dates have been successfully used by Amitay et al. [6] to produce temporal profiles of popular topics. By observing HTTP headers, particularly the Last-Modified field, significant events and trends are revealed. The Last-Modified field is used to approximate the age of the page's content. This work clearly shows that real life events can be exposed mainly due to what authors call fossilized content. A tool, named *Timestamped Link Profile* (TLP), is used to characterize the distribution of timestamped links within a topical community — see Section 2.5 for a description on how these plots are assembled. Despite the lack of the Last-Modified field in more than 50% of the URLs, the authors report several experiments where clear patterns are visible.

For web crawling, dating web resources is an extremely important task. Due to its increasing size and dynamic nature, it is impossible to frequently crawl the entire web. Search engine designers have implemented strategies to cope with this problem. For example, regularly updated documents tend to be more frequently crawled [26]. Bar-Yossef et al. [12] have explored a document's neighborhood to estimate its decay rate. Instead of limiting the observation to the number of dead links of a single document, the document's neighborhood is studied to develop a better estimative of the decay

rate. A document's decay rate is not only related the number of dead links in it, but also to the number of dead links in documents at distance $>= 1$ from it. Using a similar approach, neighborhoods have also been successfully used to estimate properties of individual documents. Sugiyama et al. [75] have developed a technique to refine the *tf-idf* scheme for a target document by using the contents of its hyperlinked neighboring pages. Chen et al. [24] proposed a method to estimate the PageRank of a particular web page by only looking at the local neighborhood of a node. The authors conducted several experiments and concluded that it is possible to produce a reasonable PageRank value by only looking at a moderate number of nodes. This method can be used to efficiently estimate individual PageRank values without having to perform a large-scale computation on the entire web graph.

In a distinct line of research, Wong et al. [81] describe methods based on temporal information extraction for document dating. Using natural language processing techniques, specifically information extraction methods, it is possible to identify words or expressions that convey temporal meaning (e.g. "today", "a long time ago"). Although we are not aware of any experimental work following this approach, this technique can also be used to date web documents.

5.2 Dataset Characterization

Our initial data contains 10,000 URLs from the Yahoo! Directory [84]. This sample was obtained using Yahoo! Random Link (YRL) service [82]. The YRL service returns a random link from Yahoo!'s index via HTTP GET requests. Table 5.1 shows the distribution of top-level domains (TLD) in the sample. In Table 5.2, depth is defined as the level of the hierarchy of the URL within the domain. For example, the depth of a homepage is 0, while for a URL pointing to a document within the root folder depth equals 1.

Although there is no public official information about the YRL service, we think that the sample is based on Yahoo! Directory and not on its search engine index. The high percentage of homepages found in the sample (84.4%) corroborates this hypothesis. There is a small number of duplicates in our sample (0.66%). This has no impact in our analysis because it occurs in a very small percentage of URLs. It is important to note that this dataset cannot be seen as a random sample from the Yahoo! search engine. Obtaining random samples of URLs from the web is an active research problem [13].

The Last-Modified header field was available in 52% of the requests in our sample.

Table 5.1: TLD distribution in sample.

TLD	Count	Percentage
.com	6,012	60.12%
.org	920	9.20%
.uk	643	6.43%
.au	435	4.35%
.edu	430	4.30%
.net	389	3.89%
.us	230	2.30%
.ca	177	1.77%
.nz	110	1.10%
.sg	46	0.46%

Table 5.2: Depth distribution in sample.

Depth	Count	Percentage
0 (homepage)	8439	84.39%
1	751	7.51%
2	428	4.28%
3	223	2.23%
4	97	0.97%

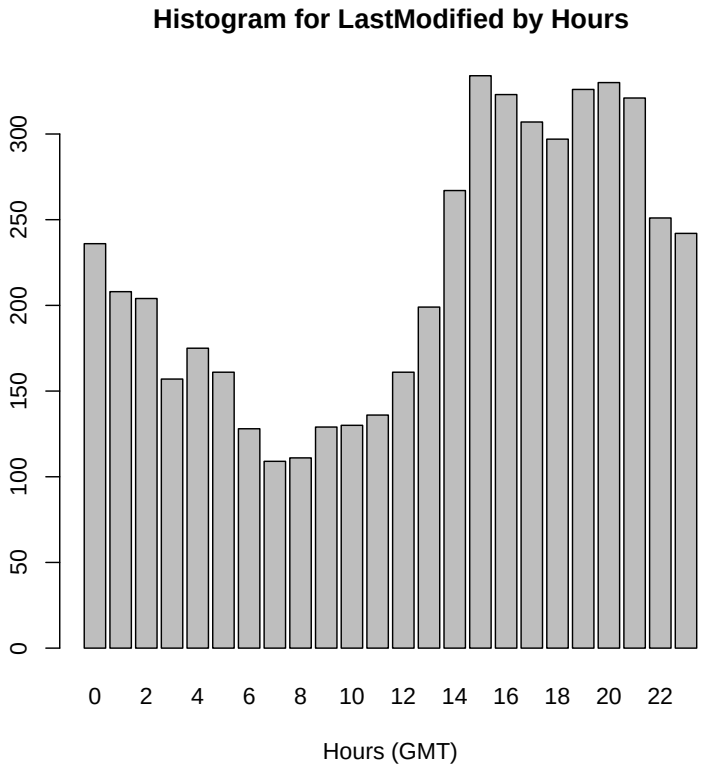

Figure 5.2: LastModified value by hour.

Since many servers return the current date for every HTTP request, we excluded these cases. The hourly distribution of Last-Modified values is presented in Figure 5.2. It is interesting to note that most updates occur within USA working hours (8 AM to 8 PM (EST)). The percentage of hosts returning Last-Modified dates has been measured in previous studies (see Table 5.3). Gomes and Silva [38] attribute their low score to the fact that "the lifetime of contents is getting shorter". Although these studies report a relatively wide range of values, the value of 52% found in our sample is compatible with those.

5.3 Exploring Neighbors

As illustrated in Figure 5.1, we've defined three types of neighbors for a given web document: incoming, outgoing and assets. Incoming neighbors corresponds to the

Table 5.3: Percentage of Retrieved Last-Modified in Previous Studies.

Study	**Year**	**% Retrieved**
Douglis [31]	1997	79%
Brewington [20]	1999	65%
Bent [16]	2003	56%
Gomes [38]	2006	36%
Our Sample	2007	52%

set of web documents having a link to the selected document. Outgoing neighbors is the set of web documents pointed to by the selected document. Assets are all the web resources, namely images, objects, CSS files or JavaScript files, that are pointed to by the original document. More formally, the assets collection includes all URLs that are referenced in HTML `src` attributes. To obtain these collections of URLs, we used the Perl programming language and the Yahoo! API. Perl was used to download each document and parse its contents. Both the links to other documents (out-links) and the links in HTML's `src` attributes (assets) were collected. The Yahoo! API [83] was used to retrieve the links pointing to each URL. This API does not return more than 1,000 links per request. In our sample, 10% of the tested URLs were above this limit. For these URLs, only the first 1,000 returned links were considered. It is worth noting that results from the Yahoo! API are different from those obtained using Yahoo!'s public web interface [63].

For each set of neighbors, an average Last-Modified value was calculated. As defined before, requests returning the current date were considered invalid and were not included in the average. In the end, for each URL, we had its Last-Modified value, the average Last-Modified value of the incoming links, the average Last-Modified value of the outgoing links and the average Last-Modified value of its assets. In Table 5.4 the percentage of valid responses for each type of neighbor is presented. As expected, the number of valid answers is higher for media assets (typically static files). These results confirm that it is possible to use a document's vicinity to improve the percentage of valid answers. However, are these valid results useful? To answer this question, we observed the correlation between a document's Last-Modified value and the Last-Modified average from its neighbors. The strongest correlation occurs with a document's out-links (r=0.74), followed by media assets (r=0.6) and finally in-links (r=0.28). Figures 5.3 and 5.4 represent the correlation plots for in-links and out-links respectively.

Table 5.4: Valid Last-Modified by neighbor type.

Source	% **Valid**
In-Links	47%
Out-Links	48%
Assets	83%
Neighbors Combined	94%
Original URLs (baseline)	52%

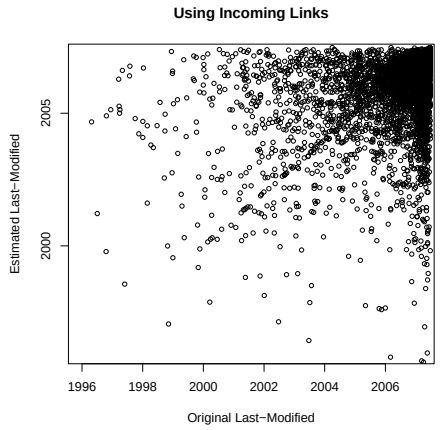

Figure 5.3: Correlation of Last-Modified between document and incoming links.

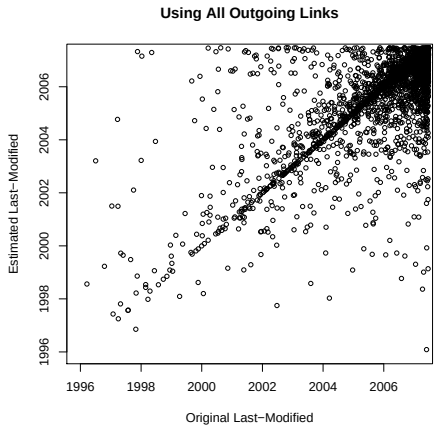

Figure 5.4: Correlation of Last-Modified between document and outgoing links.

To further investigate the nature of these correlations, we separated the out-links in two distinct sets, the outgoing links to other domains (out-out) and the outgoing links to the same domain (out-in). These two subsets exhibit different correlation values (see Table 5.5). It is possible to observe that the high correlation found with out-links is mostly attributable to outgoing links to the same domain, and likely to the same server. Finally, averaging all neighbors except incoming links, we get a correlation value of 0.73.

Table 5.5: Statistics for outgoing links subsets.

Source	% Valid	Correlation (r)
Out-In Links	45%	0.82
Out-Out Links	38%	0.51

5.4 Experimental Evaluation

To validate our findings we prepared a real world experiment. We built a Timestamped Link Profile (TLP) (see Section 5.1) for a popular event, specifically the *FIFA World Cup 2006 in Germany*. Our first step was to manually select queries that return relevant results for this topic, namely: "Germany World Cup", "2006 FIFA World Cup" and "World Cup 2006". Then, we selected two sites that were references (i.e. authorities) for this topic: `http://wm2006.deutschland.de` and `http://www.dfb-kulturstiftung.de`. Using the Yahoo! API, we retrieved all incoming links to these domains (not only to the homepage), resulting in a total of 1,000 URLs after the elimination of duplicates. We then applied our techniques, as described before, to retrieve Last-Modified values from HTTP headers. We were able to get 258 valid answers from this list. Despite this very low percentage, the TLP plot clearly revealed the expected pattern (see Figure 5.5). Following the steps described in Amitay et al. [6] we framed our plot to the period of interest, namely from the 1st of January of 2004 to the 1st of January of 2007.

To validate our approach, we then produced Last-Modified estimates based on media neighbors and outgoing neighbors. Incoming links neighbors were excluded due to their low correlation values, as shown in the previous section. We then plotted another TLP based exclusively on these estimates (see Figure 5.6). It is important to note that we've only used URLs that returned invalid HTTP headers. In other words, this TLP is based solely on neighbors from URLs that did not return a valid HTTP header. A visual comparison of these figures reveals a similar pattern. The time frame of the depicted event (i.e. "FIFA World Cup 2006") is clearly visible in both plots. There is a small shift to the right in the estimated plot that can be explained by the time needed to update the neighbors. We've conducted the same experiment using other topics and have achieved comparable results. We are confident to state that our neighbors-based approach returns valid results.

In a second experiment, we tried to explore a collection containing a very small number

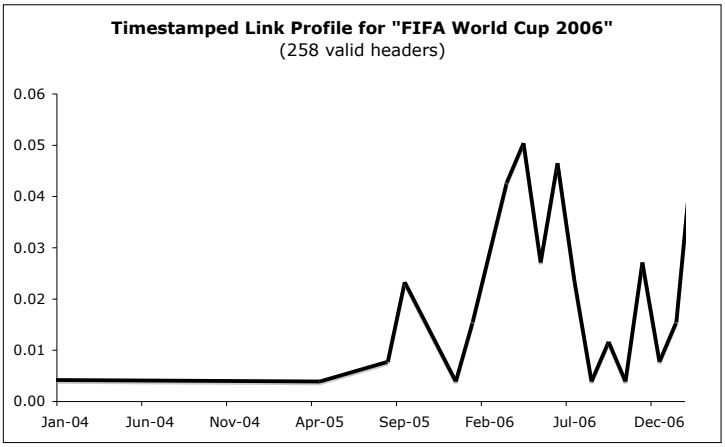

Figure 5.5: TLP for FIFA World Cup 2006 based on valid Last-Modified values.

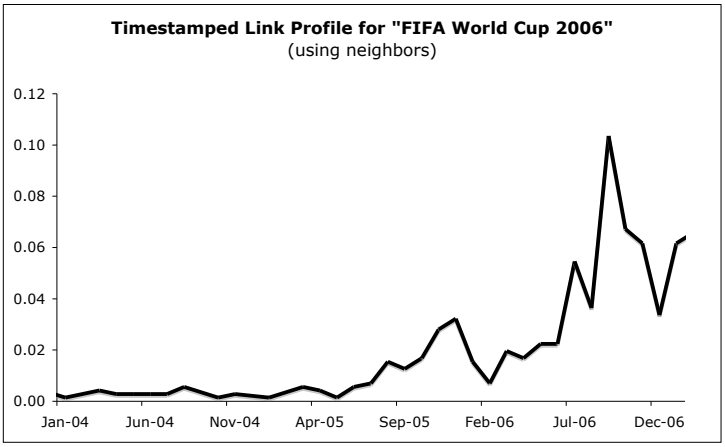

Figure 5.6: TLP for FIFA World Cup 2006 based on estimated Last-Modified values.

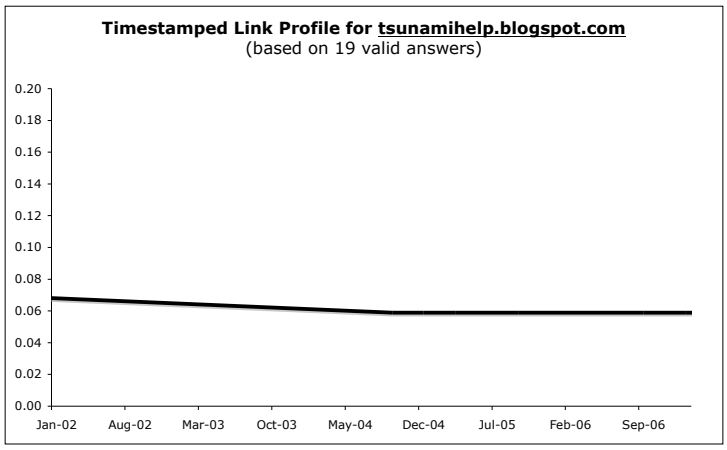

Figure 5.7: TLP for the tsunamihelp blog based on valid Last-Modified values.

of URLs. In this case, we decided to build a TLP for a single URL. We choose `http://tsunamihelp.blogspot.com` because it is an regularly updated blog about events that have significant impact and discussion (i.e. tsunamis). First, we collected 100 incoming links using the Yahoo! API. Then, we tried to fetch the Last-Modified field from each URL in this collection. The small number of valid answers (19%) is plotted in Figure 5.7 for the period between January 2002 and December 2006. Clearly, no pattern emerged from this small set of results. Using our technique, we produced an estimate for all the 100 URL in the base collection. We were able to get a very high response rate of 99%. The resulting plot for the same period (Figure 5.8) is significantly different from the previous one. A peak erupts from the plot in September 2006, corresponding to a period after a major earthquake followed by a tsunami occurred in Indonesia [78]. Based only on a very small collection of web documents we were able to uncover information.

5.5 Conclusions

Dating information resources is a valuable task. There are multiple use cases where this is important, such as finding up-to-date information, organizing web resources, understanding the flow of conversations or crawler scheduling. For web information

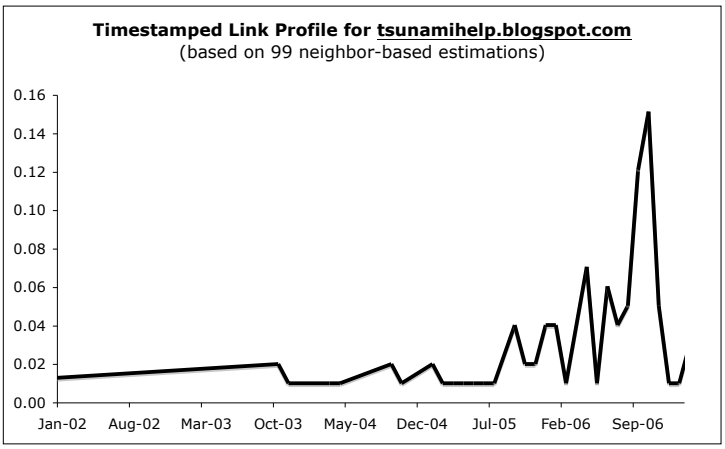

Figure 5.8: TLP for the tsunamihelp blog based on estimated Last-Modified values.

resources, estimates of the last update date have typically relied on the observation of HTTP headers, on period content analysis looking for changes, or on natural language processing. In this study we explore and evaluate a complementary technique to date web documents, specifically HTML documents. The proposed technique complements the traditional approach based on HTTP headers by also looking at a document's vicinity. The HTML structure of a document is parsed to derive its vicinity. Then, the set of neighbor resources is analyzed and an average update date is calculated. This average is then used to date web documents for which HTTP headers are not available. In the end, we are able to improve the percentage of dated web documents using HTTP headers. While alternative approaches try to detect changes by periodically inspecting documents, this technique only requires a snapshot of the web and doesn't require any historical information.

In our sample of 10,000 URLs, we were able to retrieve valid HTTP headers in 52% of the requests. As expected, by including neighbors, we were able to retrieve valid headers from 86% of the URLs. There were differences between the various types of neighbors. While for media assets we had 83% valid answers, for incoming links we only had 47% answers. With the inspection of static web resources (e.g. images) it is possible to achieve a higher percentage of valid answers. We found that there is a strong correlation between a document's Last-Modified field and the average Last-Modified

value from its neighbors (r=0.73). To validate these findings, we've conducted several experiments using real-word data. We've built Timestamped Link Profiles for several topics, following the original idea by Amitay et al. [6]. For each topic we produced two TLPs, one based on real data and the other based on estimates from media and outgoing links neighbors. The results confirmed that neighbor-based estimates are valid and accurate. Despite these positive results, it is important to note that estimates should be interpreted with critical judgement. During the experimental validation phase we observed occasional "noise" in the data caused by a non-homogeneous vicinity. For instance, while producing the TLP for one of the topics, we noticed that a large number of neighbors were from a specific web server with badly configured headers. This situation caused a significant shift in the final estimated TLP. Removing these pages from the neighbor's set was sufficient to greatly improve the final estimates. Overall, our results show that it is possible to use a document's vicinity to estimate its most recent update date.

Chapter 6

Link Authority over Time

The dynamic nature of documents directly impacts the high-level structure of a collection. For instance, changes in web documents result in changes to the layout of the web graph. Based on a large collection of blog posts, we observe that the distribution of citations to a web document over time contains valuable information. We show that time-agnostic algorithms are unable to capture the correct popularity of sites with high citation activity. Moreover, we show that a time-biased approach to authority detection performs better than alternatives discarding temporal information.

This chapter presents a study on the impact of time in authority estimation algorithms based on link analysis. We use a large collection of blog posts spanning over more than 3 years as a test bed for experimentation. First, we compare a traditional link-based ranking algorithm with a temporally biased alternative, providing some insights on the evolution of link data over time. Next, we use temporal profiles of link activity to analyze sites with high citation activity. Finally, we experimentally evaluate a time-biased scoring function for authority estimation on the web.

6.1 Introduction

The World Wide Web is a particularly dynamic medium. Several studies have observed and documented this dynamic nature of the web [67, 36, 2]. A recent study by Adar et al [2] has shown that popular pages exhibit a very high change rate, with approximately 40% of the pages in the sample changing nearly every hour. In a previous study,

Ntoulas et al. [67] found that, after one year, 50% of the content on the web is new, reflecting a high degree of change (see Section 2.2). However, and despite this intrinsic dynamic facet, research on web dynamics has not been incorporated into mainstream web research. As an example, consider the well-known link-based ranking algorithms, such as PageRank [71] or HITS [52], where the web is modeled as a directed graph without any temporal information attached. These algorithms have been used as an important signal to determine the authority and relevance of documents on the web or, more precisely, on snapshots of the web.

We believe that the temporal dimension of the web is a potentially rich data source for information retrieval tasks. In this context, we measure and investigate the impact of time in link-based authority estimation algorithms. This study is based on a large collection of blogs spanning over more than 3 years. We start by observing the dynamics of global ranks as older and newer information is removed. Next, we observe how citations to web documents evolve over time. In this process we identify several cases where standard, time-agnostic, algorithms fail to identify emerging trends. To overcome these limitations, we study the distribution of citations over time as a possible signal for information retrieval tasks. We design two experiments to evaluate the use of temporal features in authority estimation algorithms. In a first experiment, we compare time-agnostic and time-sensitive ranking algorithms with a reference rank based on the total number of visits to each blog. In a second experiment, we use feedback obtained from domain experts to contrast different rankings of Portuguese news web sites. Both experiments confirm that time-dependent features obtained from link-analysis are able to capture valuable information that can be used to improve ranking tasks.

6.2 Document Collection

As a result of a protocol established between the University of Porto and Portugal Telecom, we have had access to a large collection of Portuguese blogs. The criteria originally used to determine if a blog is Portuguese or not are unknown to us. One important aspect of this collection is the fact that it includes all the blogs registered at SAPO Blogs (hosted at `http://blogs.sapo.pt`), a service operated by SAPO, a subsidiary of Portugal Telecom. Since this is a complete collection of a single and large blog service provider, independent of crawling policies or problems, it can be seen as a good sample of Portuguese blogs. The blogosphere (or blogspace) is very rich in temporal informa-

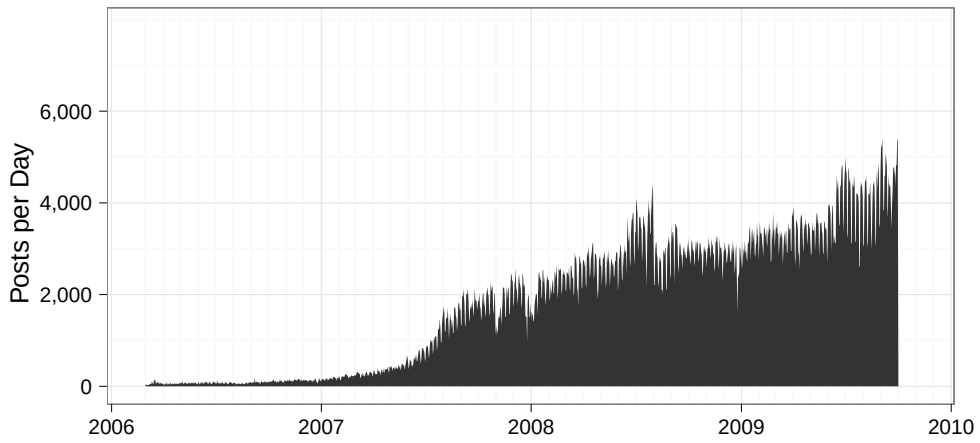

Figure 6.1: Posts published over time.

tion because, by default, all posts have an attached timestamp. This provides a context especially well suited for time-dependent analysis. The remainder of this chapter is based exclusively on the blogs from this service, except where otherwise noted.

The current version of the SAPO Blog service was launched in March 2006 and, as of September 2009, had more than 100,000 registered blogs and approximately 2.4 million posts. Figure 6.1 depicts the number of posts published per day over time. To obtain an overall picture of the linking activity in this collection, we extract all HTML links from each post, together with its publishing date. Figure 6.2 shows the total number of links found per month (labeled 'original'). This figure reveals an anomalous pattern in link activity occurring in mid 2008 and mid 2009. After manual observation of the atypical periods, we found that these peaks can be attributed to fake blogs, commonly used as *link farms* to artificially promote selected sites. To address this problem we implemented a simple filter based on three signals: the ratio between the total number of links in a blog and the total number of posts; the average number of posts per day; and the average number of links per day. We removed all blogs according to the following criteria: an average number of links per post higher than 10, an average number of posts per day higher than 50, and an average number of links per day higher than 500. Additionally, we also removed all blogs with less than 3 posts and all blogs with an overall temporal span smaller than 3 days. A total of 62,346 blogs were identified and removed (49.4% of the initial set), resulting in a drastic reduction of the original collection. The revised link activity over time is included in Figure 6.2 with the label

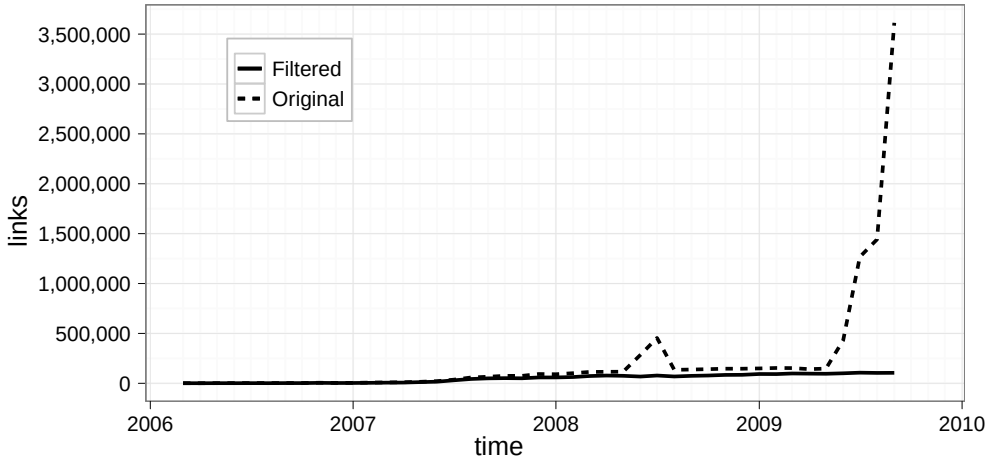

Figure 6.2: Links found over time.

'filtered'. It is clear from this figure that the unexpected peaks were most likely due to artificial links.

6.3 Link Authority Dynamics

In this section we present the results of our investigations on the dynamics of link-based algorithms over time. We use a simple measure based on the total number of incoming links (i.e. citations) to determine a web document's link authority. This is commonly referred to as a document's in-degree when modeling the web as a directed graph. Previous work in this area has shown that basic in-degree counts outperforms PageRank using several information retrieval measures (i.e. NDCG, MRR and MAP) [66]. Additionally, computing a document's in-degree is a simple procedure when compared to PageRank or other sophisticated measures. Also, we focused our study on host-based rankings rather than document-based rankings because, contrary to isolated web document, they are more stable over time and, therefore, more suitable to our goal. Moreover, host-based rankings are easier to evaluate by human assessors. In addition, there is a much higher number of citations for each host, resulting in richer sets of temporal information. To obtain a host-based ranking, we simply coalesced individual web documents by host.

A ranking based solely on the total number of citations is vulnerable to manufactured

scenarios where a small number of hosts originate the large majority of the citations. This situation is particularly sensible when working with a relatively small subset of the web, such as in our case. We found several notorious examples in our dataset — e.g. one host had an in-degree of 1,056 but all citations were from only 2 distinct sites. To account for these situations, instead of using the raw number of citations, we limited the number of citations from one host to another to 1. In other words, several citations from one host to another host (typically from different pages) are combined and only count as one, more specifically the first one. In the end, we obtain a weightless directed host-graph without multiplicity. It should be noted that this is an oversimplified approach to produce a link-based rank. Nonetheless, manual observations of top ranked items confirm that the ranking obtained is satisfactory for this exploration. The global rank obtained is dominated by media hosting sites like YouTube, Flickr and Blogger's cache servers. The top positions of the rank also include reference sites like the English and Portuguese versions of Wikipedia. The list is then filled with Portuguese news sites, reflecting the notorious attention given by blogs to mainstream media. This rank is based on the entire collection of blog posts, ranging from March 2006 to September 2009 (43 months).

To investigate the impact of the temporal dimension in link authority measures, we trimmed the initial collection and computed new ranks. First, we trimmed the data by removing one month from the beginning of the collection and only counting the links found within the remaining 42 months. The collection was further trimmed one month at a time, until only one month of data was left (Sep 09). This approach was intended to simulate the common setting where a search engine only has access to a limited period of historical data. Additionally, using a similar method, we also trimmed the newest months from the collection. This scenario was intended to simulate the situation where a search engine doesn't have access to the freshest information — e.g. there is a temporal lag between the published data and the crawled copy of this data. Finally, we also produced 43 independent ranks, each based on the data from a single month.

To evaluate the impact of removing data from the collection, we compared the percentage of common items between the global rank (the baseline) and each of the other ranks based on partial information. The adopted baseline represents the ideal scenario, where an algorithm has complete knowledge of the reality. We limited each rank to a maximum of 100 references. Figure 6.3 shows the rank intersection for each scenario. We also tested Kendall's *tau* distance [49] to measure the correlation between ranks and obtained very similar results. Given that the idea of *common items* is easier

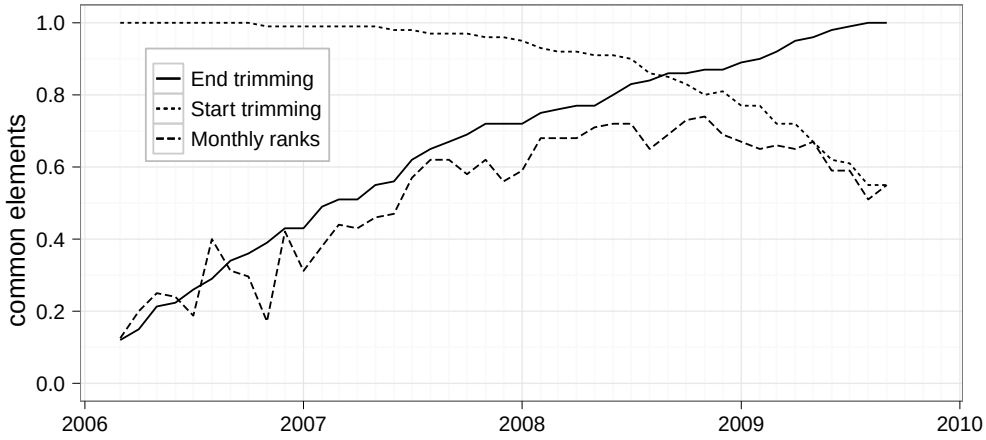

Figure 6.3: Common items with baseline in top 100 ranks for different trimmings.

to visualize and understand, we opted for this measure. From the figure we can see that, when trimming the newest months (end trimming), the percentage of common items increases from $\sim 20\%$ (when only one month is considered, March 2006) to 100% (when the complete collection is considered). On the contrary, when oldest months are trimmed (start trimming) the percentage of common items decreases from 100% (complete collection) to $\sim 55\%$ (only September 2009 is considered). The monthly ranks, also depicted in the figure, show the expected behavior, with a significant maximum of $\sim 75\%$ in November 2008. In other words, if we only had access to one month worth of data, this month would produce the best approach to a rank based on the entire collection (43 months).

Overall, the curves behave as expected — as more months are removed, and less information becomes available, the number of common results decreases. At first glance one could think that historical information is less determinant given that removing older months from the complete dataset results in a slower decrease in similarity when compared to removing newer months. As can be seen in the figure, the similarity between ranks has a steeper decay when new months are removed (top right area of the chart). However, if we recall that the SAPO Blog Service only gathered significant attention after mid-2007 (see Figure 6.1), we see that, between this date and the final date of the collection, ranks based on start and end trimming are similar. Until mid 2007, the correlation between the top 100 lists when older months are removed is very high and stable, a situation that can be attributable to this reduced overall activity.

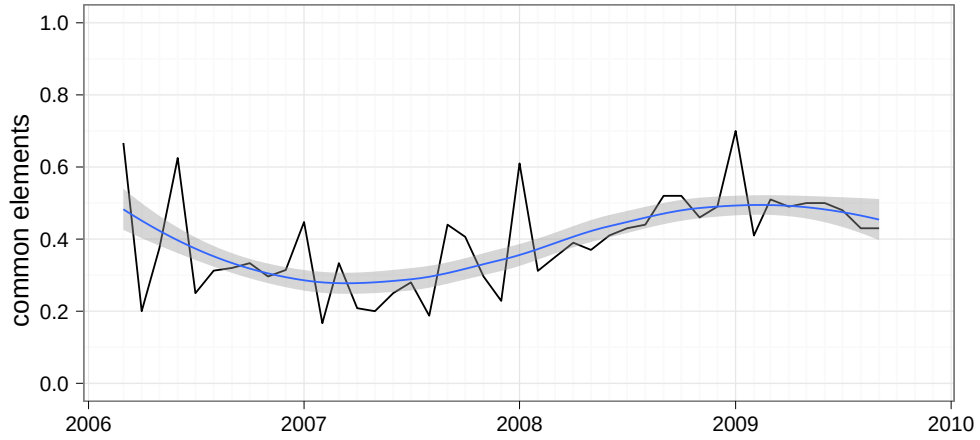

Figure 6.4: Common items in top 100 ranks for each month versus previous month.

To have a more fine-grained perspective about the internal dynamics of link-based ranks over time, we compared the rank obtained when using data from a single month with the rank obtained using data from the previous month. As depicted in Figure 6.4, the percentage of common items oscillates between 20% and 60%, with an apparent tendency to stabilize in more recent months. We add a smoother function to the figure to better illustrate this pattern. This shows that, even when using data from a single month, there is an important fraction of items that prevail.

Overall, this analysis shows that, for standard cumulative link-based ranking algorithms, access to historical information is important but not critical. We can see that even without 6 months of the latest data we can produce a rank that is very similar to the baseline rank obtained with complete knowledge of the collection ($> 90\%$ common items). As mentioned before, similar plots were obtained when using Kendall's *tau*. This means that, when using a time-agnostic algorithm, the freshness of the index is not decisive to obtain accurate estimations of citation-based rankings. The presented data also seems to indicate that there is no evident advantage of historical data over contemporary data, nor the other way around.

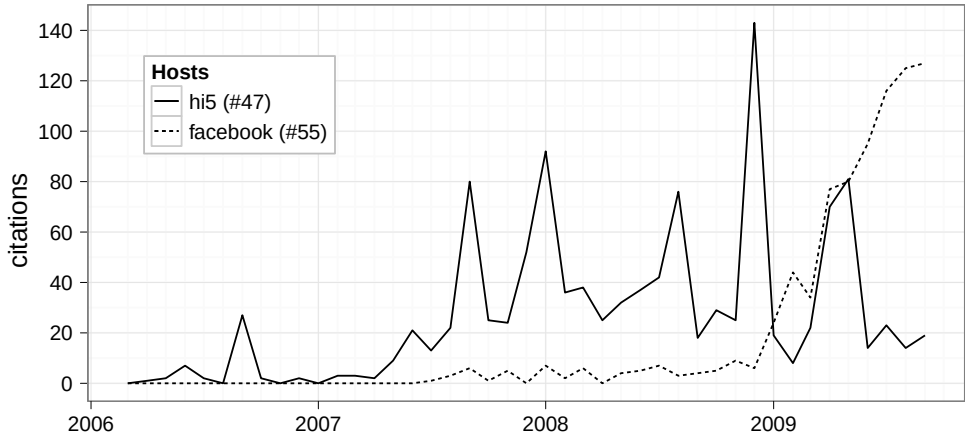

Figure 6.5: Hi5 and Facebook monthly citations over time.

6.4 Temporal Profiles

As discussed in the previous section, time-agnostic ranks based on in-degree counts tend to be relatively stable over time. To better understand the evolution of citations for a given host, we extract *temporal profiles* using the dates of each individual citation. A temporal profile of a web page or site corresponds to the distribution of citations to that page or site projected over time. Figure 6.5 shows the monthly evolution in citations for two well-known social networking web sites — Hi5 and Facebook. The Hi5 service has historically been very popular in Portugal compared to other competitors. This is reflected in the host's temporal profile. On the other hand, Facebook has been mostly unknown in Portugal until early 2009. As is very clear from the figure, Facebook had a very significant growth during 2009. However, when comparing the total number of citations, Hi5 (position 47) is placed ahead of Facebook (pos. 55). Given the fact that Facebook has consistently collected a higher number of citations in all months during 2009, this ordering can be seen as outdated. A second example is presented in Figure 6.6. In this case, Twitter's temporal profile is presented side by side with two other sites: MySpace, a social networking website, and Slide, a media-sharing service. This comparison illustrates that, although Twitter has a significantly higher number of citations in the last months, it is still ranked below the other two sites when comparing the raw number of accumulated citations.

Temporal profiles are a useful tool to unveil part of the history behind simple in-degree

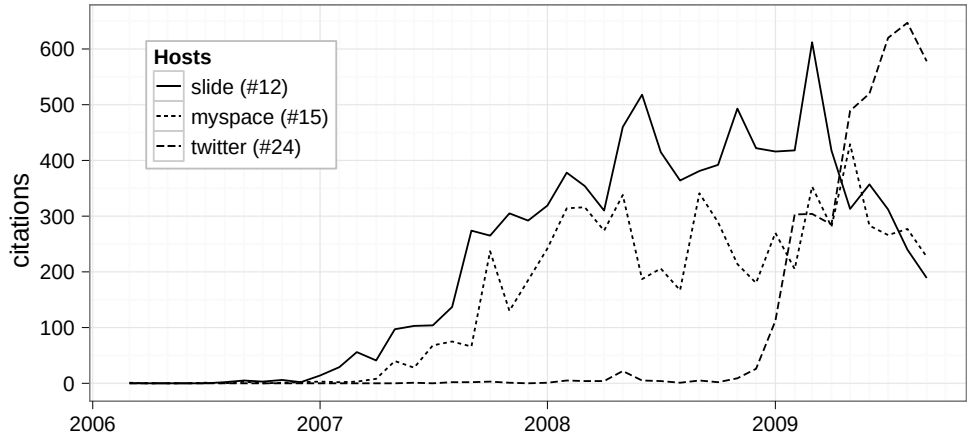

Figure 6.6: Twitter monthly citations over time.

counts. They can reveal growing or fading citation activity over time. Thus, the temporal profile of a web resource can be regarded as an original signal that captures time-sensitive information about these resources. It is worth noting that, for the two examples presented here, the rise in citations occurs in the most recent months of the collection. This indicates that, although access to the most recent information may not be determinant for overall ranks (as discussed in the previous section), it can be critical for particular sites such as these. To conclude, it is important to keep in mind that these charts suffer from a bias introduced when reducing the multiplicity of citations between two hosts. Since we are considering only the first citation when multiple citations occur between the same hosts, it is likely that these figures are under-evaluating recent data.

6.5 Experiments and Results

This section documents the experiments conducted to evaluate the use of temporal profiles as signals for host ranking. In the following subsection we define a simple time-dependent link authority measure and show its advantages in comparison to the previously discussed approaches. This measure is also evaluated with positive feedback from domain experts in an experiment designed to rank Portuguese news websites.

6.5.1 A Time-Sensitive Link Authority

We explore an approach that takes into account temporal information together with citation counts to produce a time-sensitive ranking. In a nutshell, instead of considering that all citations have the same weight, we attribute a lower weight to older citations. In the previously presented time-agnostic approach, each citation counted as one vote with weight 1. Here, we use the formula defined in Equation 6.1 to compute the value of each citation as a function of its age (in months). A citation's weight decays as its age increases.

$$w(age) = \frac{1}{(age+1)^p}, \;\; p > 0 \tag{6.1}$$

The parameter p can be adjusted to define the rate of decay, for instance a higher p value means that older citations loose value more quickly. The age of a citation was measured as the number of months to September 2009. For example, if $p = 1$, then each citation made during Sep 09 would have a weight of 1, and each citation made during Aug 09 would have a weight of $\frac{1}{1+1} = 0.5$. A citation made one year earlier (Sep 08) would have a weight of $\frac{1}{12+1} = 0.077$. This approach is similar to the previously discussed work from Yu et al. [87]. The previously defined *original* rank is obtained with $p = 0$, i.e. all weights are equal independent of age.

We produced two ranks considering the complete dataset using $p = 0.5$ (named *decaysoft*), and $p = 1.5$ (*decayhard*). A comparison between these ranks and the previous time-agnostic rank (named *original*) is depicted in Figure 6.7. This figure presents the percentage of common items (*y-axis*) between each pair considering a different number of top items (*x-axis*). There is a high number of common items between each rank's top items since the most cited sites tend to be consistently very cited over time. This means that top 10 (or lower) ranks tend to be very similar, even when older citations are discarded. This figure also shows that by decreasing the value of older citations, we obtain different ranks. Moreover, we can see that this difference increases, as expected, in proportion to the weight given to older citations. With $p = 0.5$ the percentage of common items converges to 87%, while with $p = 1.5$ this percentage decreases towards 60%.

To better understand the impact of modifications to the parameter p, we compared the original rank with different time-sensitive ranks, at different p values. The results of this experiment are depicted in Figure 6.8, with p varying between 0.1 and 4. While in the

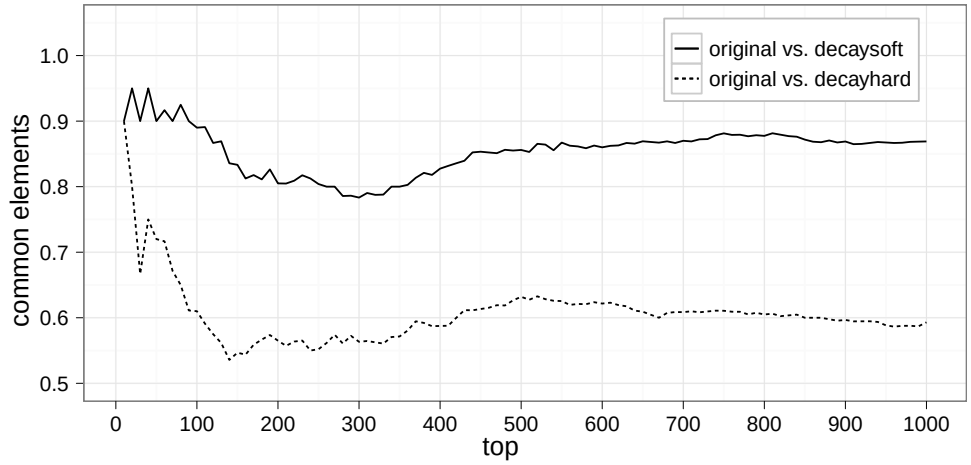

Figure 6.7: Rank intersections for original ranking versus time-sensitive alternatives.

previous analysis we compared different top ranks keeping fixed p values (i.e. *decayhard* and *decaysoft*), here we compare three fixed tops (i.e. 50, 500 and 1000) and change p in the *x-axis*. The percentage of common items between the original rank and the time-biased ranks reaches a plateau near $p = 3$. This happens because the range of possible weights obtained with Equation 6.1 decreases as p increases. In other words, the weight attributed to posts with different ages becomes 0 for higher p values, resulting in almost identical ranks.

The above experiment shows that, by altering the weight of links as a function of time, we can produce different rankings. We now evaluate the quality of these ranks by comparing them to the previously defined time-agnostic ordering. In the following, except where otherwise noted, we compare the baseline rank with the time-dependent rank where $p = 0.5$. First, we revisit the cases discussed in Section 6.4 in light of the time-sensitive rank. Given the higher value given to recent citations in this new approach, Facebook (#39) is ranked ahead of Hi5 (#54) as expected, and Twitter (#10) clearly jumps to the top, in front of Slide (#11) and MySpace (#12). This result is a better reflection of the current Portuguese web. In a nutshell, sites that exhibit a quick and unexpected increase in recent popularity are discernible when using a time-sensitive approach. Another point worth noting is the fact that the temporal ranking is able to quickly detect changes in a site's address. One of the major television networks in Portugal (SIC) has changed its URL two times in recent years, from `sic.sapo.pt` to `sic.aeiou.pt` (Apr 2008 to Aug 2009) and then back to its original `sic.sapo.pt`. This evolution is

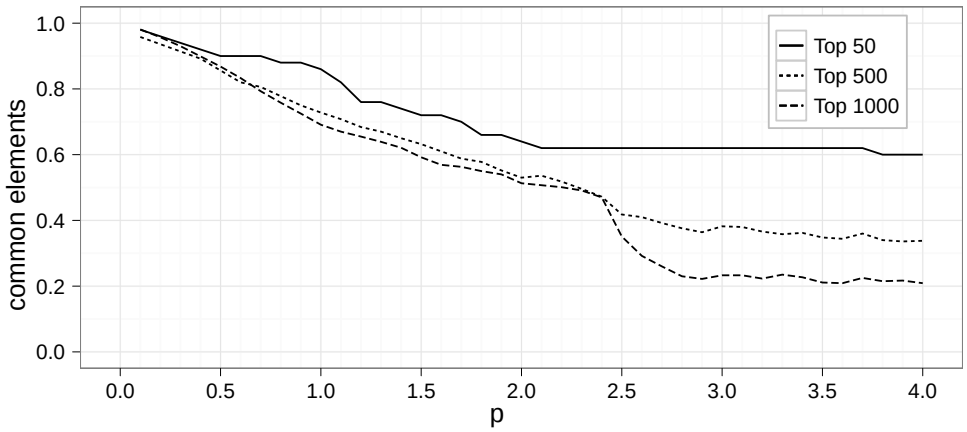

Figure 6.8: Common items between original ranking and a time-sensitive alternative with different values for the p parameter, for a distinct number of top items.

very explicit in the temporal profile shown in Figure 6.9. While the baseline rank still puts `sic.aeiou.pt` (#31) ahead of `sic.sapo.pt` (#41), the time-sensitive rank with $p = 1.5$ inverts this order (#23 vs. #53). We have found several other cases where the same happens.

6.5.2 Ranking Blogs

In the context of the previously mentioned protocol between the University of Porto and Portugal Telecom, we were also given access to a sample of SAPO Blogs's web access logs. This is an extremely relevant and unique resource that can be used to estimate a global abstract ranking for SAPO's blogs. These log files span over 27 days between late March and mid April 2010. Consequently, there is a lag of approximately 6 months between the collection of blog posts studied and the corresponding web access logs. To assess the quality of this dataset, we first estimate the consistency of these web access logs by extracting ranks for each of the days found in the log files. In this process we only include HTTP GET requests, POST requests being excluded because they represent post or comment submissions, not directly related to page visits. In addition, we also remove all requests made by the top web crawlers, namely Google's, Yahoo's and Bing's. These crawlers account for a significant number of the total hits (approx. 10% in our case), introducing a bias that we try to minimize. We observe that

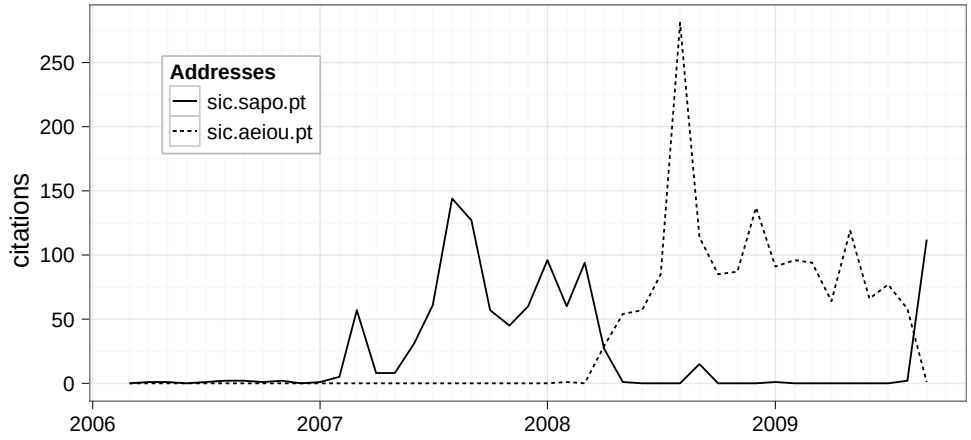

Figure 6.9: `sic` domains and citations over time.

the total number of hits per day for all blogs exhibits a stable value across all days of our sample (in the order of millions). Next, for each day of access logs, we extract a list of blogs ordered by the total number of hits in that day filtering crawlers' accesses. We produce a global rank that combines all 27 days, aggregating all hits per blog. The correlation between the rank in each day and this global rank shows high stability over time — for top 10, 100 and 1,000 ranks, the percentage of common elements is stable and close to 85%.

To evaluate the impact of time in authority estimation algorithms, we compare the baseline, *decaysoft* and *decayhard* ranks defined in Section 6.5.1 with this global rank based on the total number of visitors per blog. First, to obtain comparable lists, we filter our ranks to include only SAPO's blogs. The rank of blogs ordered by the total number of visitors has more than 85,000 elements, while the three alternative ranks we have created have less than 20,000 elements. Table 6.1 summarizes this information and also includes the overlap between the reference rank (based on the number of visitors) and each one of the other ranks. In this table we see that the percentage of common elements between the reference rank and the ranks to be evaluated is considerably low. Taking this into account, we adopt the approach proposed by Bar-Ilan [10] — i.e. we intersect and re-rank the two lists being compared and use Spearman's *rho* as a measure of rank correlation. Spearman's *rho* ranges between -1 and 1, where 0 stands for no correlation, -1 for complete disagreement and 1 for complete agreement between the two ranks. The coefficients of correlation, for different number of top items, between

Table 6.1: Size and overlap of SAPO's blogs ranks.

Rank	Size	Overlap (%)			
		@100	@500	@1000	full set
Visitors	87,914	100	100	100	100
Original	17,132	24	23	25	38
Decaysoft	14,874	22	21	25	34
Decayhard	14,874	15	16	20	34

Table 6.2: Correlations, for different top items, between the reference visitors-based rank and the time-agnostic plus time-sensitive ranks.

Top	Original		Decaysoft		Decayhard	
	rho	p	rho	p	rho	p
10	**0.22**	0.54	-0.10	0.78	0.09	0.81
20	0.20	0.38	**0.28**	0.24	0.22	0.35
100	0.12	0.25	**0.15**	0.12	0.14	0.16
500	0.16	0.00	**0.22**	0.00	0.18	0.00
1,000	0.17	0.00	**0.22**	0.00	0.21	0.00
5,000	0.25	0.00	**0.30**	0.00	0.27	0.00
10,000	0.32	0.00	**0.38**	0.00	0.30	0.00

the reference rank and the time-agnostic and time-aware alternatives are outlined in Table 6.2. The highest correlation value in each row is highlighted in bold. Also included in this table is the significance value of each correlation test (p-value). When comparing ranks of size equal or greater than 500, all results are statistically significant at 0.01 level. Examining this table we see a consistently higher correlation value in the *decaysoft* rank, except when comparing the top 10 elements (not statistically significant). Moreover, for most of the top values observed, the *decayhard* rank also exhibits a higher correlation compared to the *original* rank. These results show that, when compared to a time-agnostic alternative, a rank including temporal information has a higher correlation with an independent reference rank based on the number of visitors.

6.5.3 Ranking Portuguese News Websites

Obtaining relevance judgements for abstract top rankings, such as the ones we are producing, is very challenging. Without a context (e.g. a query) we cannot select users, nor ask them to simply rank the theoretical importance of sites. Thus, to address this problem and reduce the uncertainty of an evaluation, we design an experiment that narrows down the context to a particular topic, specifically *Portuguese news websites*. First, for each one of the three ranks being evaluated, we hand-pick all the news sites and produce new ranks keeping the information about the original positions. Then, for each set of two ranks being compared, we identify all pairs of sites where the ordering in each rank was reversed. Next, for each pair, we calculate the distance between the positions in each rank. Finally, we add the two distances (one for each rank) and order these values, obtaining a list representing the pairs where the difference between the ranks is more noticeable. To illustrate this procedure, consider the following case: in the time-agnostic rank site `A` is ranked in position 10 and site `B` is ranked in position 22. In the time-dependent rank the ordering is reversed and site `A` is ranked in position 8 and `B` is in position 6. The final value for this pair is 14, obtained by adding the differences found in each rank ($12 + 2$).

After this initial procedure, we select 9 pairs of popular news sites from the top of this ordered list and conduct two experiments to assess the relative value of each ranking. Given that most media outlets manage various web hosts (e.g. `news.example.com`, `www.example.com`), we decided to use the principal URL for each site (typically the `www`). In a nutshell, we have singled out pairs of popular news sites where the ordering found in the two ranks exhibits a more striking difference. It is worth restating that we try to reduce the ambiguity of the evaluation by focusing on the more extreme cases.

We conduct a first assessment of this data using the information published in Netscope's monthly rank [62]. Netscope is an opt-in service that measures websites' audiences. This is the most popular service in Portugal, frequently cited when discussing the typical profile of the Portuguese internet user. All sites included in our pairs could be found in Netscope's top 100 rankings. These rankings are published monthly and, to match the end of our dataset, we base this evaluation on the September 2009 rank. The Netscope ranking supports the baseline rank in 7 (out of 9) cases, while it endorses the temporal-sensitive ranking in 2 of the pairs. While this experiment clearly attests the importance of a global, time-independent rank, it also suggests that the temporal

profile of citations contains valuable information.

To complement this first experiment we organized an additional evaluation using human experts. We contacted 11 experts in the area of communication media — mostly teachers and editors in reference publications — and asked them to express their opinions about the set of 9 pairs defined before. More specifically, for each pair, we asked them to indicate which of the sites they view as more popular for the Portuguese bloggers in general. We asked them to skip the pair if they had doubts about the relative popularity of the sites. From the collected answers, the experts agreed with the baseline rank in 5 cases and with the temporal rank in the 4 remaining cases. This experiment reinforces the idea that a rank biased towards newer citations contains valuable information. Overall, our experiments show that, while historical data is indispensable to determine the value of resources, contemporary data contains important information that reflects current trends familiar to users.

6.6 Conclusions

This work is based on a large sample of blogs from a single service provider, spanning a period of 43 months. Although the collection is limited to a single provider, it corresponds to a complete copy of the data from a large service without any sampling. This results in an organic collection without the biases typically introduced during the crawling phase. Using this data we study the temporal properties of a ranking obtained with a simple algorithm based on the total number of citations. The evidence collected suggests there is no significant difference between the value of historical information and contemporary information in time-agnostic rankings — i.e. the impact of removing either older or newer data from the collection is alike. Using several informal examples we also observe that the standard time-independent ranking is unable to capture the correct popularity of sites with very high citation activity in recent periods. This is a problem likely to occur more frequently as the size of a collection grows and more value is accumulated in past citations, i.e. ranks will tend to crystallize and be less vulnerable to subsequent changes. We compare these results with an alternative link-based ranking algorithm where newer links are given a higher weight. In other words, a citation's value decays as it gets older. As expected, we observe that time-sensitive and time-agnostic ranks become more divergent as the value attributed to older links decays more rapidly.

To evaluate and compare both time-independent and time-dependent approaches in ranking we designed several experiments. First, based on several hand-picked examples, we see that an algorithm that favors recent citations, more quickly discards abandoned web sites and more rapidly identifies popular present trends. Also, we find that a time-sensitive rank can be used as a good indicator of the correct address of web sites that had multiple URLs over time. Next, we designed a detailed experiment to assess the value of the temporal rank versus the classic temporal-agnostic rank. We establish a reference rank based on the number of visitors to each individual blog and measure the correlation between this rank and the alternatives being evaluated. The time-biased alternatives exhibit the highest correlation with the reference rank. This experiment clearly confirms the advantages of combining temporal information in a link-based authority estimation algorithm.

As a final experiment, we asked a group of experts to give feedback on a selected number of pairwise comparisons. While the standard baseline rank was preferred in 5 out of 9 cases, in the remaining 4 cases the time-dependent rank was favored. This suggests that valuable information can be found in both ranks. While the time-agnostic rank is still a vital source of quality data, the time-sensitive rank captures important information that is otherwise invisible.

It is clear that access to fresh data is essential but not sufficient to identify current trends, algorithms are of major importance. Temporal profiles are a valuable tool and offer a richer picture of a web resource's authority when compared to raw citation counts. Although both time-agnostic and time-aware approaches are based on the same raw data, the experiments conducted indicate that they can be treated as complementary signals for relevance assessment by IR systems. We show that temporal information on the web can be used to derive stable time-dependent features, which can be successfully used in the context of web document ranking. We conjecture that the use of time-dependent features will be conditioned by the retrieval task being addressed. In summary, the major contributions of this chapter include a detailed analysis and characterization of a realist, large-scale collection of web document from a time-aware point of view and the evaluation of the impact of a time-dependent scoring function in a standard citation-based rank.

Chapter 7

Content Dynamics in Retrieval

In real-world collections, the content of documents is routinely revised over time, either to add, correct or delete information. This dynamic character of document content is an important source of signals containing rich temporal information. We use this information to characterize the dynamics of content over time and to develop new measures for term weighting. The measures proposed are successfully evaluated, outperforming the classic term frequency measure in several experiments.

This chapter presents our investigations on the use of content dynamics for retrieval tasks. We use documents from Wikipedia in all evaluation experiments described here. First, we look at update profiles of articles to show the relationship between revision activity and topic popularity. Then, we extract the content from revisions made to each article and observe the progression of several properties. Finally, we present new term weighting measures that incorporate a document's history in their calculations.

7.1 Introduction

In real-world information retrieval systems, the underlying document collection is rarely stable or definite. For instance, in personal systems, such as files or e-mails stored in a computer, documents are routinely added, removed or edited. Similarly, in enterprise and public environments, the existence of shared repositories of information is a standard scenario, resulting in active collections of documents which are continually updated. In this chapter we focus on the study of content-based features over time, i.e.

signals extracted from the content of documents at different points in time.

We use the English version of Wikipedia as a document collection in all experiments described in this chapter. Wikis are unique software systems that allow users to easily create and maintain web documents. The most distinctive feature of a wiki is the ease with which web pages can be created or updated. Typically, no authentication is needed to perform actions on a wiki system, making them ideal tools for online collaborative information systems. Another important feature of wikis that is present in most implementations is the preservation of the entire revision history of every web page. The Wikipedia is a prime example of a service made possible by the use of wikis. Wikipedia is an open web-based encyclopedia developed on top of the free MediaWiki software. Each article is the result of the collaboration of many volunteers from around the world. Currently (July 2009), Wikipedia has over 15 million articles written in more than 250 languages and is ranked among the top ten most-visited sites in the world [79]. The English Wikipedia alone has more than 3.3 million articles.

Since the complete revision history of each single article is preserved, there is an enormous quantity of information available from Wikipedia's past. The size, scope and popularity of Wikipedia, together with the fact that all information kept by Wikipedia is easily available via a public application programming interface (API) [1], make this collection a unique and appropriate resource for this investigation. Finally, the fact that all content from Wikipedia is public guarantees that this study is reproducible by others.

7.2 Document Revision Activity

We start our investigations by looking at the revision activity of individual documents over time. The concept of *update profile* of an article is defined as the distribution over time of the revisions made to that article. We take into account all changes made to an article, even those that were subsequently removed (such as vandalism). Figure 7.1 show two side-by-side plots of the complete revision history of Wikipedia's articles *2005* and *2008*. These articles are hubs that point to the events reported in Wikipedia each year. We can see that the number of events being reported has been growing over the years. The article *2005* had 5,714 revisions, while the article *2008* had 11,494 revisions.

1. `http://en.wikipedia.org/w/api.php`

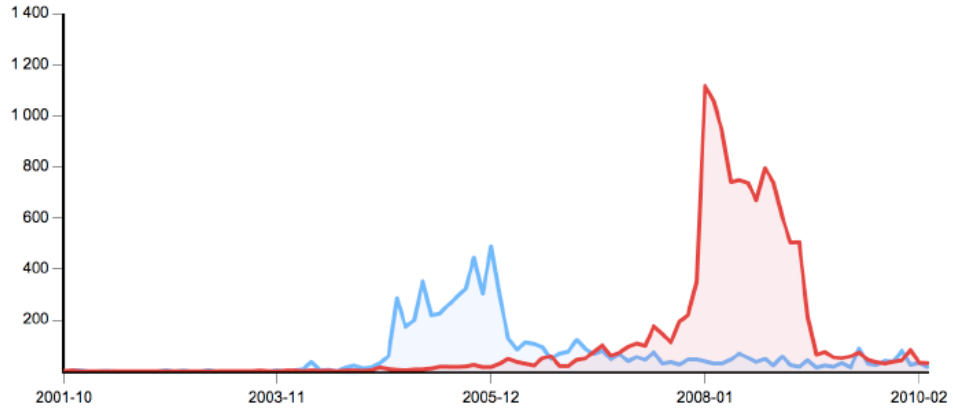

Figure 7.1: Revision history plots for Wikipedia articles *2005* (bright) and *2008* (dark).

Building an update profile for a Wikipedia article is a simple process. For a given concept expressed as a text string, we first perform a disambiguation step where we match the text to a specific Wikipedia article. Then, using Wikipedia's API we iterate through all revisions of the article and extract each revision's date. Finally, we count the number of revisions by period (e.g. day, month, year) and produce a standard time series. Due to the existence of missing values, we need to fill the time series with empty "slots" to have an uniformly sampled series. To illustrate this, consider an article that has been created in January and then updated in March. To get a consistent time series we need to automatically insert February with zero updates.

We've built update profiles for several pages based on popular subjects. We found a striking correlation between the popularity of the topic (as perceived by mainstream media) and its profile. The number of updates to the article grows significantly while the event is taking place, both due to the flow of new information and to the increased attention given to the article. The update profile for the article *Tour de France* (Figure 7.2) is an example of a recurring, predictable event clearly captured in the revision activity timeline. In Figure 7.3 the reaction to unexpected events, specifically the disappearance of *Steve Fossett*, is shown. A very significant peak occurred in this article in September 2007, when the adventurer was reported missing during a flight. Later, a smaller burst occurred in February 2008 when he was declared dead after months of searches. We found similar patterns by manually inspecting more than 80 individual Wikipedia articles about unexpected events. Even when the topic is only of regional interest (e.g. Portuguese presidential elections), thus containing fewer revisions, activity bursts are clearly identifiable.

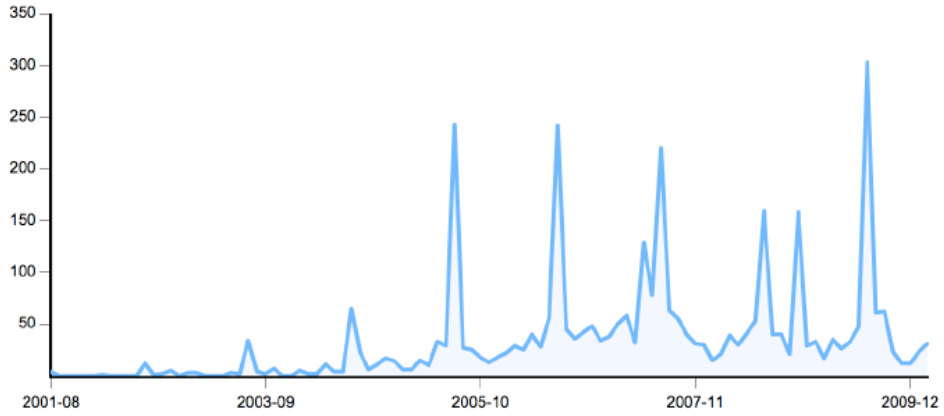

Figure 7.2: Revision history plot for the Wikipedia article on *Tour de France*.

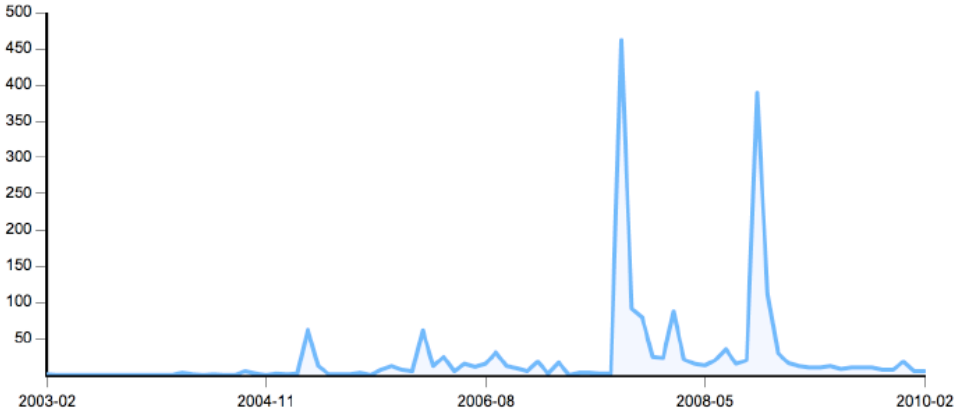

Figure 7.3: Revision history plot for the Wikipedia article on *Steve Fossett*.

To enable a quick and easy access to the update profile of Wikipedia articles, we developed a web-based application named WIKICHANGES. This tool uses the Wikipedia API to download all the revisions made to an article and then projects an aggregated count of these onto a temporal scale. WIKICHANGES also introduces several innovative features, most notably the summarization of changes for a given period, side-by-side comparison of update profiles, and embedded sparklines for the live Wikipedia. The application is available at `http://irlab.fe.up.pt/p/wikichanges`. Since its launch in 2008, this page has been visited by more than 1,000 unique visitors. Several screenshots of this application are included in Appendix A.

7.3 Term Frequency Dynamics

In this section, we investigate the frequency dynamics of terms in web documents over their lifespan. Specifically, we investigate the progression of similarity values over three projection variables. We have seen that documents on the World Wide Web are dynamic entities. Contrary to other communication media, where different versions of the same document are treated as separate documents (e.g. print), on the web this distinction is less obvious. Web documents are rarely static, exhibiting a high degree and rate of change. To understand the internal dynamics of web documents over time, we observe and measure the changes between different versions of the same document. We focus our study on a particular type of web documents — collaboratively written documents, more specifically Wikipedia articles.

We start by modeling each version of a document as a term frequency vector and use the cosine similarity measure to quantify the differences between past versions and the current version of each article. To investigate the progression of similarity values we consider three different projections axes, namely revision order, revision date and document size. Depending on the projection variable used, the similarity curves exhibit different shapes. We also contrast the internal dynamics of high quality documents with a sample of ordinary documents.

In a related line of work, Adar et al. [2] conducted a detailed observation of a large collection of popular web pages and were able to clearly distinguish between stable and dynamic content. The stable part of a document is defined as the content that remains the same over time. Using *change curves* plots, the authors showed that the stable content of a page becomes steady after a short period of time. The results presented

are consistent with our findings. However, our work is different since it focuses on the complete lifespan of web documents, from inception to its current version, as opposed to the observation of subsequent changes in popular documents. In other words, while Adar et al.'s work is focused on the changes made to existing documents, we try to measure the internal dynamics of a web document since its creation. Moreover, we look at typical documents from a wiki, instead of popular pages with frequently updated content (e.g. news sites, portals, forums).

The investigation presented here has similarities with the work published by Thomas and Sheth [76] on the content dynamics of Wikipedia articles. These authors model each revision to an article as a "tf-idf vector" and use a cosine similarity measure to evaluate the convergence of content. We explore a similar idea to model changes between revisions. However, these authors are focused in a classification problem — distinguish high quality articles from lower quality articles by looking at content evolution. They found no statistically significant difference between both types of articles in terms of edit history. Our work has a different context, it is focused on the measurement and characterization of the internal dynamics of a document's content for information retrieval tasks. Moreover, while these authors use an absolute revision-based timeline to observe the evolution of content, we use normalized projections to overcome the problem of article comparability.

7.3.1 Document Collection

We use the English version of Wikipedia to assemble a collection of documents for analysis. We select all articles currently classified as *featured article* (i.e. belonging to *Category:Featured articles*) and a parallel set of random articles. The featured article category contains articles identified by the community as high quality documents, frequently singled out in Wikipedia's frontpage. The most noticeable difference between these two groups of documents is in the total number of revisions, with featured articles having a significantly higher number of total revisions. Table 7.1 summarizes the distribution of the number of revisions in both sets. For each set of documents we present the value for the 1st and 3rd quartile, the median and the mean number of revisions. It is worth noting that, to avoid sampling documents with a very small number of total revisions, we discard all random articles with less than 50 revisions.

Table 7.1: Document collection overview.

		Number of Revisions			
	Articles	1st Q.	Median	Mean	3rd Q.
Featured	2,710	348.5	645.5	1,363	1,534
Random	2,430	65	100	226	188

7.3.2 Term Frequency Analysis

As mentioned in the previous section, we model each version of a document as a *term frequency (tf) vector* and use the cosine of the angle between the two vectors to quantify the similarity of two document versions. With this approach, similarity values vary between 1 for identical vectors and 0 for orthogonal vectors. We remove all wiki-markup and stop words from each version of the document before assembling the tf vectors.

To observe how content evolves within a document, we compare each version of a document with its current version. As an example, consider Figure 7.4 which shows the evolution of this similarity measure by revision for two featured articles. The Wikipedia article about "35 mm film" currently has slightly over 700 revisions, while the article about the "1896 Summer Olympics" has approximately 1200 revisions. As can be seen in the figure, the *similarity profile* of both articles steadily converges to 1, although at different paces. This is the expected result — each revision made to an article tends to move it closer to its current version. Worth of notice is the fact that content similarity tends to evolve quite rapidly, reaching high levels after a relatively small number of revisions. For instance, in the "35 mm film" the cosine similarity between the version at revision 200 and the latest version of the article (over revision 700) is over 0.9. The abrupt drops observed in both profiles are due to vandalism, a well-known problem in Wikipedia. Finally, we have highlighted with a circle the revision where each article was added to the *featured articles* category.

Although this figure reveals quite similar trends in two different articles, it also shows that it is difficult to compare the similarity evolution of articles with a distinct number of revisions. For this reason, we propose the normalization of the horizontal axis based on a quantile discretization approach. This way we are able to observe, side-by-side, articles with a different number of revisions and obtain a comparative picture of the internal dynamics of content across a broad group of documents. We explore three

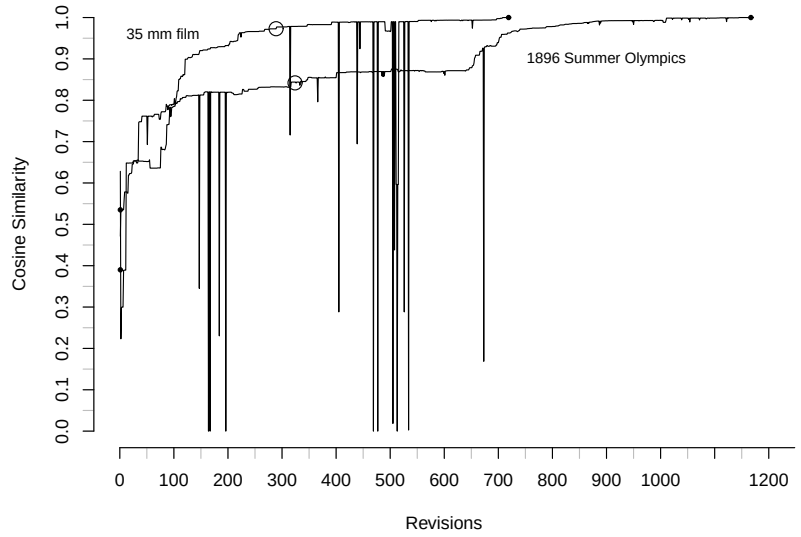

Figure 7.4: Similarity by revision order in two Wikipedia articles.

different projection variables for a normalized horizontal axis — revision order, revision date and document size. The following sections describe each approach.

Discretizing by Revision Order

We establish 25-quantiles for each article's revision history and extract the content for each of the 25 bins. For instance, in an article with a total of 50 revisions, we first extract revision 2 (1st bin), then revision 4 (2nd bin) and so forth. The content found in each of the 25 revisions (bins) is compared with the content available in revision 50 (the current version of the article). Figure 7.5 depicts a series of boxplots, each one summarizing the similarity values found between each bin and the current version of the document, for all the 2,710 featured articles. Note that the *x-axis* is represented in a $[0, 1]$ scale for consistency. This picture shows a clear pattern about the evolution of content similarity over revision order. As can be seen with the help of the horizontal lines added, the median similarity is over 0.8 since the 4th bin, i.e. at 16% ($\frac{4}{25}$) of the revision history of an article. Moreover, halfway the revision history of featured articles, the 1st quartile of similarity values is higher than 0.9. In other words, in more than 75% of all featured articles the intermediate revision is already very similar to the current

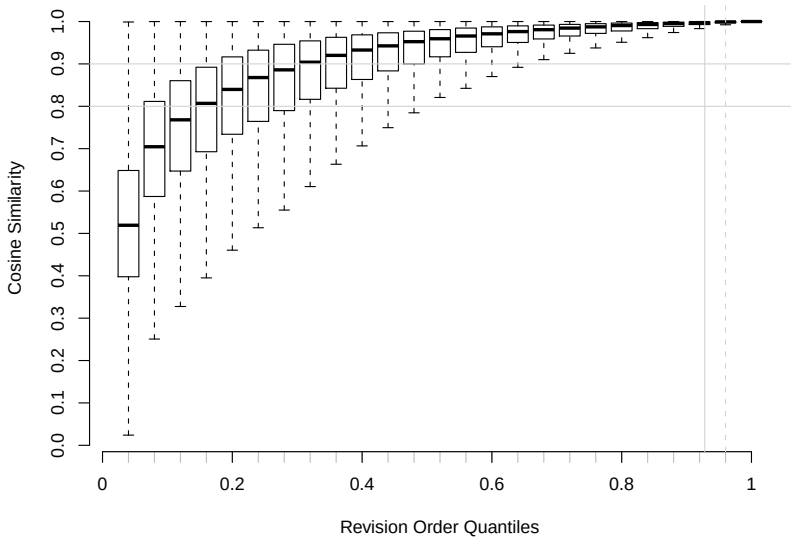

Figure 7.5: Cosine similarity for featured articles, discretized by revision order.

version. Subsequent changes to an article have very little impact in a document's term frequency vector. This figure clearly shows the rapid progress of content similarity in featured articles. When looking at a comparable plot based on the random collection of articles mentioned before, a somewhat different picture appears as seen in Figure 7.6. Although the similarity values also move consistently towards 1, there is an higher dispersion of values in each bin when compared with featured articles. The height of each boxplot indicates the spread of values in each bin. Contrasting the median cosine similarity for both types of documents highlights the differences between the two datasets (see Figure 7.7).

Discretizing by Revision Date

In this approach, we consider the date information that is available in each revision made to a document. Based on this information, we can view the progression of similarity values as a function of *time* instead of *order*, as presented in the previous section. First, we discard short-lived documents by removing all articles with a total time span lower than 50 days. We also establish 25-quantiles for each article based on the complete temporal span of the article, from its inception to its current version. For instance,

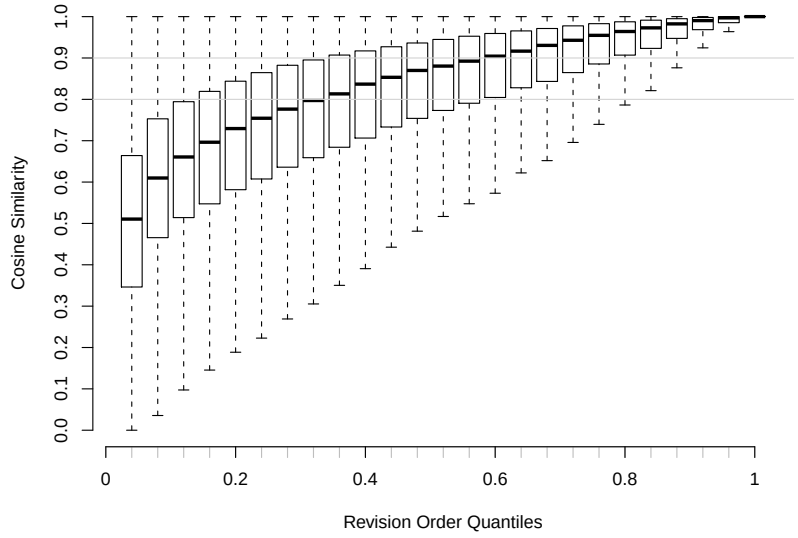

Figure 7.6: Cosine similarity for random articles, discretized by revision order.

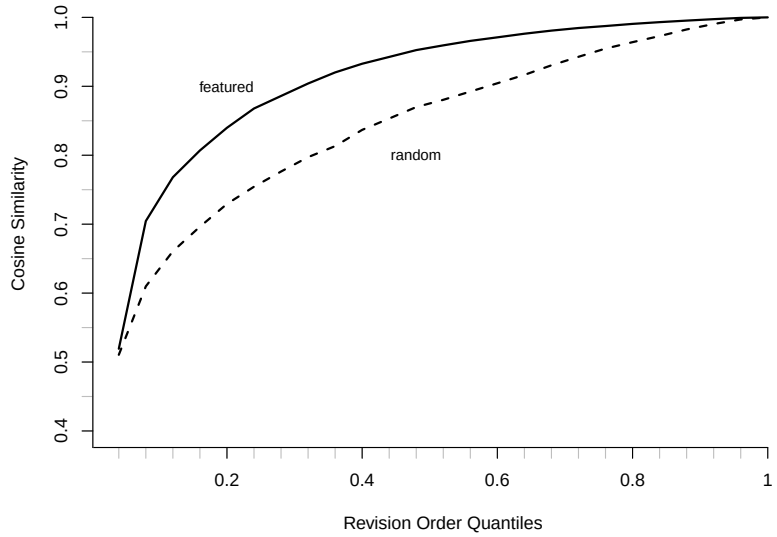

Figure 7.7: Cosine similarity of featured and random articles, by revision order.

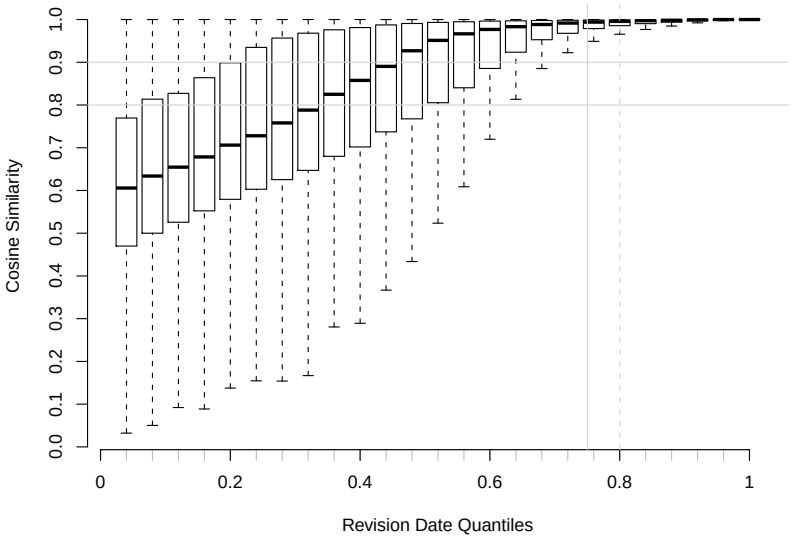

Figure 7.8: Cosine similarity for featured articles, by revision date.

for an article spanning over 100 days, we first extract the revision made in the 4th day (1st bin), then the revision made in the 8th day (2nd bin), and so forth. If no revision was made on a specific day, we consider the most recent previous revision that was active on that day. The content found in each of the 25 days (bins) is compared to the article's latest revision, corresponding to the 100th day. Figure 7.8 represents a series of boxplots, each one summarizing the similarity values found between each bin and the current version of the document, for all featured articles.

The overall profile exhibits a more irregular evolution when compared with the projection based on revision order (contrast with Figure 7.5). In this case, the initial progression of similarity values is less delimited. The height of each boxplot shows a high dispersion in the values of each bin over a large part of the initial quantiles. A very fast convergence is noticeable close to the middle of the overall lifespan. As presented in Figure 7.9, the set of random articles shows a more regular progression in similarity values. Comparing Figures 7.7 and 7.9, we can see that a projection based on revision dates results in more visible differences in the set of featured articles.

To better understand the impact of being added to the *featured article* category, we determine the average bin that represents the moment when an article is associated to this

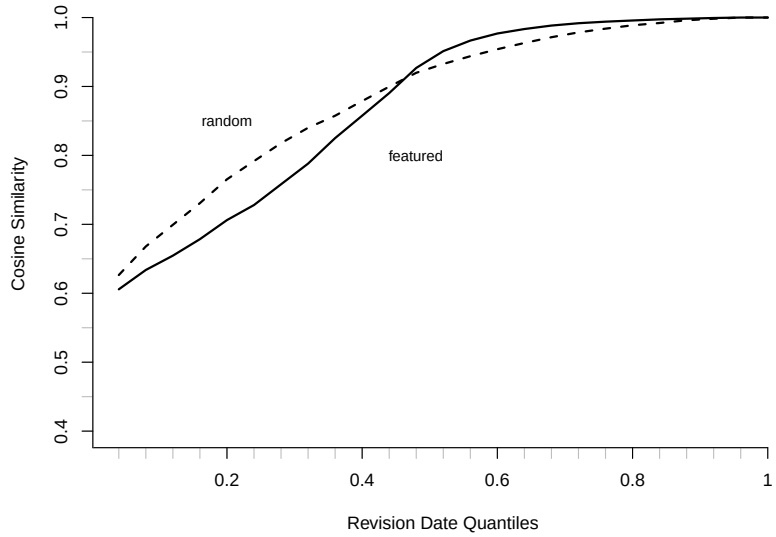

Figure 7.9: Cosine similarity of featured and random articles, by revision date.

category. When the bins on the horizontal axis are based on the revision order, the mean value is 23.2 and the median value is 24. This means that the large majority of revisions are made before the article is added to the *featured article* category. When bins are based on the revision date, the mean is 18.8 and the median is 20. The two vertical lines in Figures 7.5 and 7.8 represent the mean (solid line) and the median (dashed line) for each case. The lower value found in the second case indicates that the revisions made after being added to the *featured article* category are more dispersed through time. It is important to note that this information is not error-free. As mentioned in Wikipedia's documentation, if an article is vandalized and the category information is removed, the original dates for each category association are lost. Thus, it is likely that in reality these values are lower.

Discretizing by Document Size

After revision order and revision date, we test a third projection variable — document size. Specifically, we use the document size at each revision as a projection axis to observe the progression of changes to an article. We discard short sized documents by defining a lower limit of 100 bytes for each article's current version. Again, we establish

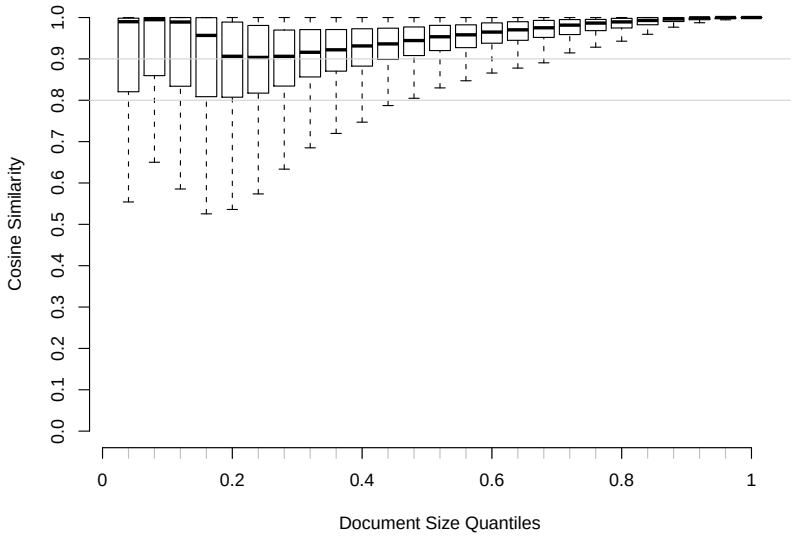

Figure 7.10: Cosine similarity for featured articles, discretized by document size.

25-quantiles for each article based on the document's length in bytes at each revision. Consider for instance an article with the current size of 75 bytes, we first extract the revision which results in an article with 3 bytes (1st bin), then the revision resulting in an article with 6 bytes (2nd bin), until the last revision (25th bin) corresponding to an article with 75 bytes. Figure 7.10 represents a series of boxplots, each one summarizing the similarity values between each bin and the current version of the document, for all featured articles.

Contrary to the previous plots, the median similarity value in featured articles is consistently very high from the beginning. It is possible to see that this value is clearly above 0.9 across all size-based bins. As can be seen in Figure 7.11, the side-by-side comparison between featured and random articles projected by document size shows quite distinct patterns. While featured articles exhibit high similarity values across all bins, random articles have lower similarity scores in the initial bins. This can be explained by the substantial difference in the total number of revision between featured and random articles (see Table 7.1), which results in a much larger number of revisions separating featured articles' bins.

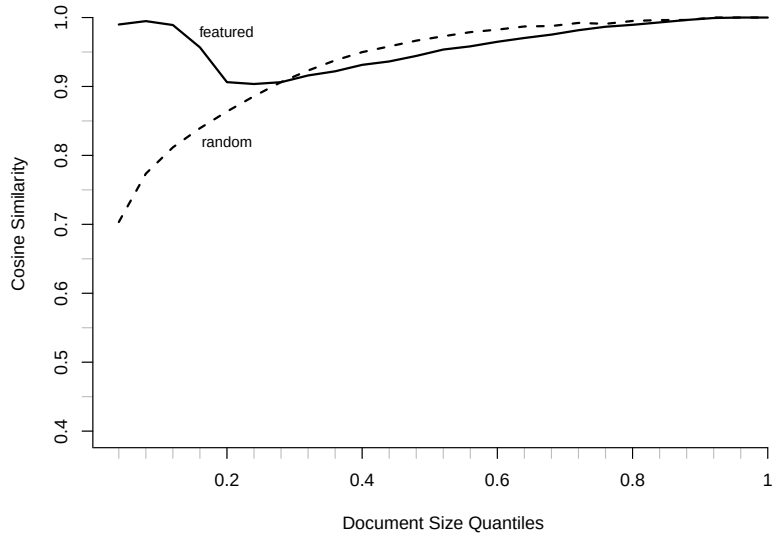

Figure 7.11: Cosine similarity of featured and random articles, by document size.

7.4 Term Weighting and Document History

Based on the previous investigations, we find that term frequency in encyclopedic documents — i.e. comprehensive and focused on a single topic — exhibits a rapid and steady progression towards the document's current version. In this section we investigate the use of document history information for document term weighting. The basic idea behind our approach is to give more weight to terms that have existed for a longer time in a document. For instance, it is our intuition that a term that has subsisted in a document since its first version should be valued higher than a term that was introduced only in the latest revision made. In other words, our hypothesis is that a term's prevalence over time is a good measure of pertinence. To evaluate this theory, we conduct several experiments using a collection of documents from Wikipedia retrieved in July 2010. We evaluate the proposed measures using three independent methods. In the first approach, we use data from Wikipedia itself to judge each set of terms. In the second method, we use an external collection of tags from a popular social bookmarking service as a gold standard. Finally, with the third method, we use feedback gathered from users to evaluate and compare our proposals against classic measures.

To incorporate the temporal dimension of documents in a scoring function, we consider

that each document d is composed of a set of revisions defined as $R_d = \{r_1, r_2, \ldots, r_n\}$. The first version of a document is represented as r_1 and the latest as r_n. Additionally, the set of revisions of a document d containing the term t is given by

$$R_{t,d} = \{r : r \in R_d \text{ and } tf_{t,r} > 0\}$$

, where $tf_{t,r}$ represents the frequency of term t in revision r. Except where otherwise noted, we treat the words *version* and *revision* as synonyms, both representing a specific instance of a document at a given point in time. A document's individual revision (or version) is represented as a tuple $(ts, terms)$, where ts is a date corresponding to the instant when the revision was published, and $terms$ denotes the contents of the document at that moment. The content is modeled as a *bag-of-words* ordered by term frequency. Consider the Wikipedia article on 'Information retrieval' as an illustrative example. This article has more than 650 words in its latest version. A bag-of-words representation of its content, ordered by term frequency, would be as follows: $terms = \{\{information, 45\}, \{retrieval, 44\}, \{documents, 32\}, \{relevant, 17\}, \ldots\}$.

7.4.1 Revision Frequency

A weighting function incorporating a term's *revision frequency* (rf) is defined in Equation 7.1. Basically, a term's rf weight for a given document is given by the ratio between the number of revisions containing that term and the document's total number of revisions. A term occurring in all versions of a document would have a rf score equal to 1. This measure ignores the frequency of terms at each revision, and only considers the presence or absence of the term. For instance, a term occurring multiple times at a given revision is weighted equally to a term appearing only once at that same revision. To incorporate a term's frequency at a given revision, we extend the previous formula and obtain a term's *revision term frequency* (rtf), as defined in Equation 7.3. In this case, we incorporate in the final score the *relative term frequency* (rel_tf) at each revision as defined by Equation 7.2. In a nutshell, the rel_tf of a term in a document is given by the ratio between the frequency of the term and the total number of terms in that document.

$$rf_{t,d} = \frac{|R_{t,d}|}{|R_d|} \tag{7.1}$$

$$rel\_tf_{t,d} = \frac{tf_{t,d}}{\sum_{t' \in d} tf_{t',d}} \tag{7.2}$$

$$rtf_{t,d} = \frac{\sum_{r \in R_{t,d}} rel\_tf_{t,r}}{|R_d|} \tag{7.3}$$

7.4.2 Revision Span

The previously defined term weighting measures view the revision history of a document as a set of evenly distributed document versions. However, the lifespan of each version varies widely, ranging from extremely short-lived versions (spanning over a few minutes) to long-lived versions that exist over many days. Taking this into account, we introduce the concept of *revision span* (rs), where the lifespan of each specific revision is taken into account in the weighting formula. The function *ls()*, defined in Equation 7.4, is used to obtain a revision's duration. The function *ts()* simply returns the time at which a given revision occured.

$$ls(r_i) = \begin{cases} \text{current time} - ts(r_i) & \text{if } r_i \text{ is the latest revision} \\ ts(r_{i+1}) - ts(r_i) & \text{otherwise} \end{cases} \tag{7.4}$$

The term weighting measure is defined in Equation 7.5, where the weight of a term in a document is given by the ratio between the period when the term was in the document and the document's total lifespan. The numerator gives the complete lifespan of a term in a document's revision history by adding the durations of all revisions containing the term. Finally, we extend this formula to also consider the frequency of each term in each revision. This measure, named *revision term frequency span* (rtfs), is presented in Equation 7.6.

$$rs_{t,d} = \frac{\sum_{r_i \in R_{t,d}} ls(r_i)}{\sum_{r_i \in R_d} ls(r_i)} \tag{7.5}$$

$$rtfs_{t,d} = \frac{\sum_{r_i \in R_{t,d}} \left(rel\_tf_{t,r_i} \times ls(r_i) \right)}{\sum_{r_i \in R_d} ls(r_i)} \tag{7.6}$$

Table 7.2: Results obtained with each method for different documents.

Article	tf	rf	rtf	rs	rtfs
Information retrieval	information	ir	information	ir	information
	retrieval	retrieval	retrieval	acm	retrieval
	documents	information	documents	science	documents
	relevant	science	ir	retrieval	ir
	precision	system	text	databases	text
Research	research	research	research	information	research
	hypothesis	information	hypothesis	knowledge	basic
	scientific	basic	basic	science	knowledge
	academic	applied	academic	applied	applied
	work	generally	scientific	research	information
Data mining	data	mining	data	people	data
	mining	data	mining	correlations	mining
	patterns	large	analysis	mining	analysis
	analysis	patterns	information	investment	people
	information	analysis	patterns	large	information

7.4.3 Preliminary Comparison of Measures

In the previous sections we introduced four term weighting functions that are based on a document's revision history. Two kinds of functions were presented: the first kind does not take into account the effective time span of each revision; in the second case, the lifespan of each revision is included in the weighting formula. In addition, we considered two approaches with respect to the frequency of a term at each revision. First, we only consider if a term is present or not in each revision, next we consider the relative term frequency at each revision. We perform a first exploratory examination of these weighting functions and compare them with the classic *term frequency* measure (tf) by looking at a few illustrative examples, represented in Table 7.2. This table presents the 5 best scoring terms obtained using each approach. We see that there are clear differences between each method, even when just the top 5 terms are considered.

7.4.4 Experimental Evaluation

In this section, we present the methods designed to evaluate the proposed weighting measures. We adopt three independent approaches, the first based on Wikipedia data, the second based on a reference external collection and a third approach based on direct user feedback. We start by analyzing the document collection used and present some

Table 7.3: Summary statistics for each set of documents.

	N	Revisions	Age (days)	Words (current)
Featured	100	1199	2053	1199
Random	100	47	1199	140
Social	100	2415	2376	744

descriptive statistics. Then, we evaluate the impact of each measure in terms of result diversity. Finally, in the last three sections, we document the evaluation experiments and discuss the corresponding results.

Document Collection

To evaluate the value of the proposed measures, we use a sample of documents from the English version of Wikipedia. We define three reference sets of documents for this investigation. The first set contains a random sample of Wikipedia's featured articles, i.e. articles randomly sampled from the 'Featured articles' category. A second set includes random articles obtained via the 'Random article' feature available on Wikipedia. The third set is based on the most popular Wikipedia articles bookmarked at a well-known social bookmarking web site. This set was prepared using the Wiki10+ dataset released by Zubiaga [89], which contains more than 20,000 unique Wikipedia articles, all of them with their corresponding social tags. Each set comprises a total of 100 distinct articles. A brief summary of the main properties of each set is presented in Table 7.3. The numbers included in the table represent the mean value for each attribute. The total number of words was calculated based on each article's current version. Comparing the different properties, we see a significant difference in the number of revisions between the random set and the other two sets. Interestingly, although articles in the social set have the highest number of revisions and age, they have less words than the articles in the featured set. This can be explained by the fact that featured articles need to meet certain criteria before being labelled as such. On the other hand, the social set includes articles that attract significant attention, which can explain the high number of revisions.

Table 7.4: Mean percentage of common items between measures in featured articles.

	rf			**rtf**			**rs**			**rtfs**		
top	10	50	100	10	50	100	10	50	100	10	50	100
tf	26	36	41	**83**	**78**	**77**	17	30	36	70	63	63
rf		—		30	45	52	**59**	**75**	**78**	36	53	59
rtf	30	45	52		—		20	38	47	**79**	**75**	**75**
rs	**59**	**75**	**78**	20	38	47		—		26	50	59
rtfs	36	53	59	**79**	**75**	**75**	26	50	59		—	

Divergence in Scoring Functions

To observe the differences between each proposed measure, we compute the number of common terms in the rankings obtained with each pair of measures over all featured articles. The results are outlined in Table 7.4. Although this table is symmetric, we have included all redundant values to facilitate reading. For each pair of scoring functions, we determined the ratio of common items for a fixed number of top terms (10, 50 and 100), averaged over all 100 features articles. For instance, looking at this table, we can see that, on average, there are only 17% of items in common between the top 10 items extracted with tf and rs. In each row, we have highlighted the pairs with highest similarity. We can see that the use of term frequencies versus simple term existence is determinant. Finally, the relatively low overall ratios suggest that the proposed measures introduce a noticeable number of new terms. Even with rtf, which has the highest overall similarity with tf, approximately 20% new terms are introduced.

Evaluation with Wikipedia Data

We can use Wikipedia itself to evaluate the quality of each set of terms. The idea is to use an article's lead as a summary of the body of the article. As stated in Wikipedia's Manual of Style [80] — *"The lead should define the topic and summarize the body of the article with appropriate weight."*. Given that featured articles are more likely to comply with Wikipedia rules, we assume that these articles have the best leads. Thus, we base this evaluation on the collection of featured articles. For each article in this set, we extract its lead (i.e. first paragraph) and, for each approach, determine the number of terms found in it. We conduct this procedure for different number of top terms, as depicted in Figure 7.12. The number of terms used is in the *x-axis* and in the *y-axis* is the ratio of

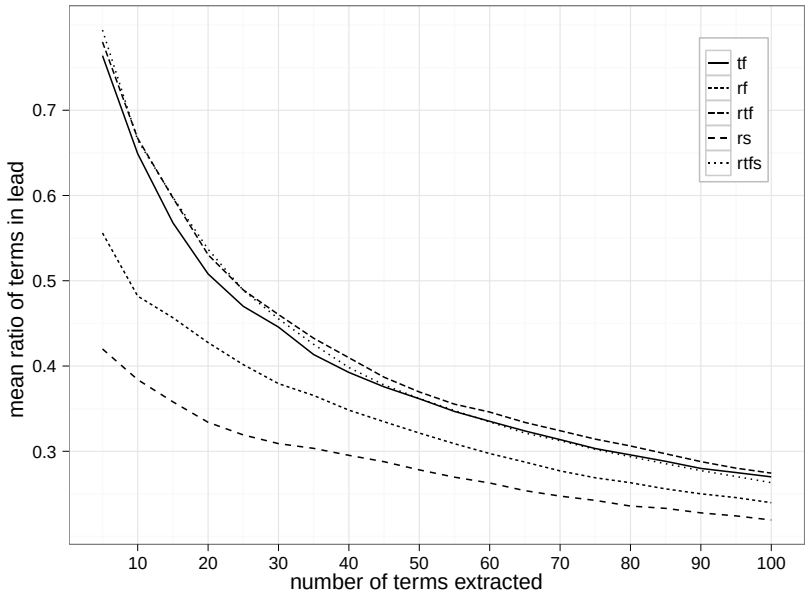

Figure 7.12: Mean ratio of terms found in articles' lead.

terms found in the article's lead. The values presented are the mean values over all 100 articles in the featured set.

From this figure we can see that the measures with best performance are those based on the frequency of terms, as opposed to those based on the occurrence of terms. More important, we can see that both rtf and rtfs outperform classic tf, when up to 50 terms are being tested. For more than 50 terms, the results obtained with rtfs decay slightly more rapidly than those obtained with tf. To evaluate the significance of these results, we use two sample paired t-tests for the rtf and rtfs measures with tf. Results are presented in Table 7.5, where each line represents a test using a specific number of top terms. From this table we can see that most results for rtf are significant, either at 99% or 95% — indicated using ** and * respectively. For the rtfs measure, we only include the values where rtfs outperforms tf (up to 50 terms). Contrary to the rtf measure, the improvements obtained with rtfs are not significant (except for 20 terms). Summing up, with this experiment, we can conclude that rtf is consistently better than tf for term extraction.

Table 7.5: Paired t-test results for rtf and rtfs versus tf using articles' leads.

	rtf		**rtfs**	
terms	t(99)	p-value	t(99)	p-value
10	1.767	0.040*	1.122	0.132
20	2.839	0.003**	2.506	0.007**
30	1.862	0.033*	0.995	0.161
40	2.772	0.003**	0.723	0.236
50	1.531	0.064	0.055	0.478
60	2.150	0.017*	—	—
70	2.400	0.009**	—	—
80	2.597	0.005**	—	—
90	2.118	0.018*	—	—
100	1.311	0.096*	—	—

Evaluation with Social Annotations

Wikipedia articles are very popular among internet users. A significant number of articles is shared by users, either by e-mail, blog posting or social bookmarking. This observation is supported by a simple analysis of the Wiki10+ dataset released by Zubiaga [89]. This dataset was prepared in April 2009 and includes all articles from the English version of Wikipedia that were bookmarked in Delicious [2] by at least 10 users. Delicious, currently a Yahoo! property, was a pioneer service in the area of social bookmarking and is still considered one of the references in this area. The Wiki10+ dataset contains 20,764 unique URLs and, for each URL, all corresponding Delicious tags. A simple analysis based on the histogram shown in Figure 7.13 reveals that the dataset only includes up to 30 tags for each bookmark. This can be explained by the fact that Delicious only displays the 30 most popular tags for each bookmark and offers no other way of obtaining the complete set of tags. Table 7.6 presents the 10 most popular tags found in this dataset for the three articles considered earlier. It is worth noting that some of the tags used are simple graphical variations of each other (e.g. data-mining and data_mining). We make no effort to consolidate or correct these instances.

To evaluate each term weighting approach using the Delicious external reference set, we measure the number of common items pairwise. First, we select the 100 bookmarks in this dataset with the highest number of users — i.e. those that were bookmarked by

2. `http://delicious.com`

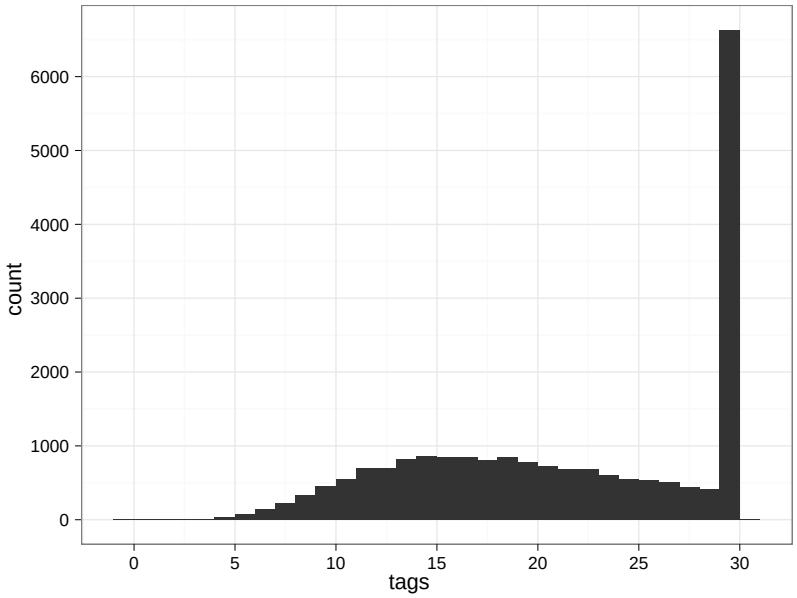

Figure 7.13: Distribution of bookmarks by number of tags.

Table 7.6: 10 most popular tags on Delicious for different articles.

Article	Top Delicious Tags	
Information retrieval	search	reference
	ir	informationretrieval
	information-retrieval	retrieval
	information	research
	wikipedia	recall
Research	research	dissertation
	wikipedia	terminology
	science	overview
	definition	researching
	info	science_technology
Data mining	datamining	database
	wikipedia	programming
	data	statistics
	mining	data_mining
	reference	data-mining

Table 7.7: Mean cosine similarity between Delicious tags and each method's terms.

	top 10	top 50	top 100	all
tf	.441	.428	.422	.408
rf	.273	.199	.168	.127
rtf	**.459**	**.437**	**.436**	**.436**
rs	.185	.183	.164	.130
rtfs	.444	.419	.419	.419

more users. Then, we compare the tags available for each bookmark with the terms extracted using each method. Figure 7.14 summarizes the results obtained, presenting the percentage of common items found for different number of top terms. We can see that both rtf and rtfs have a higher number of terms in common with the Delicious set. The superiority over tf is consistent across all number of terms considered. Again, the worst performing measure is rs.

Given that, for each term extraction method, we have weights associated with each term, we can use this information to make a more precise comparison with each tag's weight found in Delicious. Thus, for each one of the 100 articles, we produce a weighted term vector using all tags found on the Wiki10+ dataset. Then, for each term extraction method and for each article, we also create term vectors considering a different number of top terms. Specifically, we build four vectors for each article and method, one including all terms and the others considering only the top 10, 50 and 100 terms. Finally, we calculate the cosine similarity between the reference vector based on Delicious data and each one of the other five vectors. The results, averaged over all articles, are presented in Table 7.7. Highlighted in bold is the best performing measure for each number of top terms. We see that the rtf method outperforms all other methods, including the classic tf measure. We use a two sample paired t-test to evaluate the significance of rtf's performance over tf. We find that rtf's better performance when using all terms is significant at 99% ($t(99)=3.78$, $p=0.0001$), and significant at 95% when restricting the vector to the top 10 terms ($t(99)=2.24$, $p=0.014$) and the top 100 terms ($t(99)=1.96$, $p=0.026$). Again, we see that a time-aware measure exhibits better results than an approach that discards historical information.

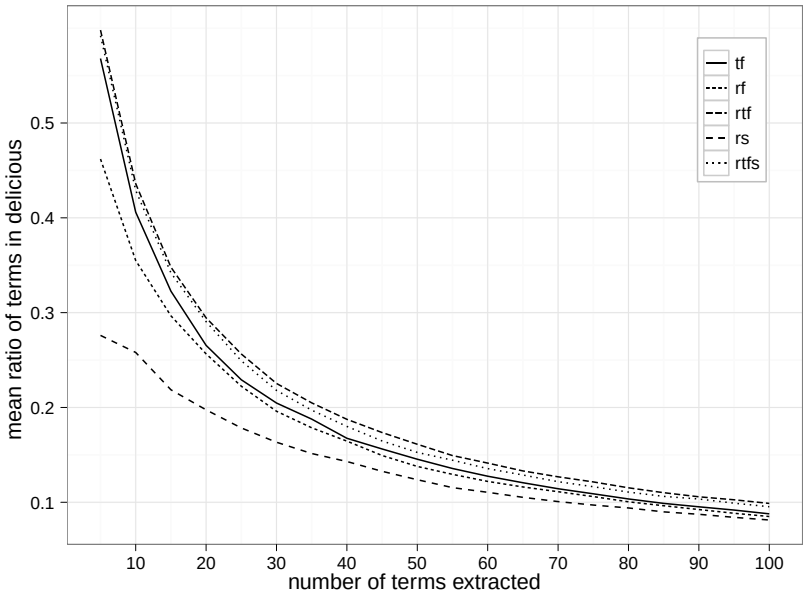

Figure 7.14: Mean ratio of terms found in top Delicious tags.

Evaluation with User Feedback

The previous evaluation methodologies are based on indirect measures, i.e. no direct user feedback is collected. In this section we describe an evaluation experiment designed to obtain direct user judgements. Basically, for each article in the evaluation set, we present two alternative lists of terms and ask the user to choose the most relevant to the article. We do some basic stop word removal and then extract an ordered list of 10 terms using each algorithm. We use the crowdsourcing [44] service CrowdFlower [3] to design this experiment and collect user feedback. CrowdFlower is a service that redirects user designed tasks to 'labor-on-demand' marketplaces, such as Amazon's Mechanical Turk. These tasks, known as Human Intelligence Tasks (HITs), are distributed across internet users (i.e. workers) that execute them in exchange of monetary payment. Given the lack of direct supervision, the execution of individual tasks offers no assurance in relation to quality control. It is well known that task design and indirect control mechanisms, such as qualification tests, are paramount when crowdsourcing jobs [51]. To improve the quality of our results we try to eliminate low-value work by using two different strategies: request multiple judgements for each task and define some tasks as ground truth.

3. `http://crowdflower.com`

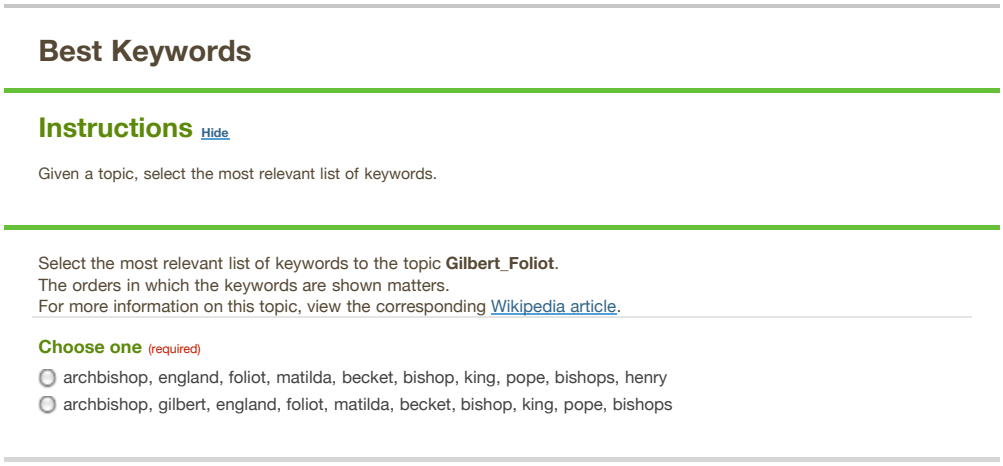

Figure 7.15: Interface design for evaluation task in CrowdFlower.

A screen capture of the interface presented to workers is presented in Figure 7.15. For each individual assessment task, we require a minimum of 5 independent judgements. Using this information, we only consider valid answers those where the most voted option wins by at least $\frac{2}{3}$ of the votes. Additionally, we define 10% of the tasks of a given pairwise comparison as ground truth, known as *gold* in CrowdFlower. Setting gold tasks can substantially improve the quality of the answers [30]. CrowdFlower's proprietary algorithms use this information, together with worker's historical record, to automatically accept or reject submissions. Given that we are conducting subjective tasks, there is no correct answer to use as ground truth. Thus, we create artificial tasks that we use as ground truth. To produce these tasks we simply replace one of the term lists with a list of keywords obtained from an unrelated article. For instance, when evaluating an article on the NeXT computer system, the user is presented with a list containing correct terms (e.g. nextstep, computers, jobs) and another with off topic terms obtained from a completely unrelated article (e.g. lacelot, merlin, excalibur). We mark the first option as the correct choice and define this task as gold in CrowdFlower's interface.

Considering the monetary costs associated with this experiment, we select two subsets of 50 articles from the original collections of featured and social articles. After running the experiments, we simply count the number of wins for option 1 versus option 2. We present the confidence intervals at 95% for the true proportion of wins of each new

Table 7.8: Confidence intervals at 95% for preferences against term frequency.

	rf	rtf	rs	rtfs
Featured	(0.105, 0.335)	(0.513, 0.821)	(0.089, 0.311)	(0.538, 0.809)
Social	(0.105, 0.335)	(0.317, 0.599)	(0.138, 0.382)	(0.433, 0.710)

measure over the term frequency baseline in Table 7.8. These intervals are calculated approximating the binomial distribution to the normal distribution. For instance, when considering featured articles, we are 95% confident that the interval 51%–82% contains the true proportion of wins of rtf over tf. As the lower confidence limit of this interval is higher than 50%, we can state that the rtf measure is preferred over the tf one, outperforming it. The same happens with the rtfs measure when compared with tf.

Overall, the quality of the proposed measures is clearer when considering the set of featured articles. In addition, we see that the rtf measure performs worst than tf for the set of social articles. We think that this can be explained by the fact that the articles in the social set are more vulnerable to vandalism and subsequent reverts. Thus, a measure that ignores the duration of the revisions (like rtf) is likely to be affected by this. We can see from Table 7.3 that the articles in the social set have a much higher number of revisions, despite a similar age and a significantly lower current number of words. To conclude, we can say that this results are clear and consistent with those reported in the previous experiments. Again, we see that the use of document history in term weighting algorithms consistently improves the results.

7.4.5 Relevance of Former Terms

Following the previous investigations, one question worth exploring is the relevance of former terms, i.e. terms that where once included in the content of an article but currently aren't. This is an interesting question worth addressing — is the past content of a document relevant to current retrieval tasks? To obtain a list of former terms ordered by importance, we simply remove all the current terms from the lists obtained with the temporal-dependent measures. Table 7.9 shows the top 5 former terms for each of the articles mentioned earlier. There are noticeable differences between each of the 4 measures. Although many of the proposed terms are broad and not directly related to each article, we can find terms that are clearly relevant (e.g. intuition, investigate, faculty). To estimate the relevance of former terms, we conduct an experiment using the set of

Table 7.9: Top scoring former terms for different documents.

Article	rf	rtf	rs	rtfs
Information retrieval	keeps	engine	stand	draws
	stores	written	bodies	engine
	engine	draws	confusion	stand
	stand	platform	networked	bodies
	bodies	gpl	keeps	confusion
Research	institutions	nih	thoroughly	reach
	corporations	institutions	literal	labor
	holder	mind	investigate	hard
	received	intuition	institutions	mind
	faculty	investigate	relations	nih
Data mining	interpreted	intelligence	interesting	intelligence
	interesting	artificial	interpreted	implies
	network	network	danger	card
	intelligence	danger	intelligence	drugs
	year	properly	concerning	danger

tags acquired from Delicious. For the same set of 100 articles used in Section 7.4.4, and for each proposed measure, we count the number of former terms that also exist on the set of tags used in Delicious. In 19 of the 100 articles we found 1 former term that was also used as a tag in Delicious. In one particular case, we found 2 former terms being used as tags. To conclude, in 20% of the articles analyzed we found evidence that former page content contains relevant keywords.

7.5 Conclusions

In this chapter we show that the history of document content contains relevant information that can be successfully used to improve information retrieval tasks. First, we have seen that the distribution of the revision activity over time is highly correlated with the popularity of a document's topic over the same period. Then, we present an investigation about the internal dynamics of collaborative web documents. We analyze the complete lifespan of real web documents by comparing the progression of similarity values over three projection axes — revision order, revision time and document size. Also, using the categories defined in Wikipedia, we contrast high quality document with regular documents. Based on this study we find that content in early versions of a document quickly becomes very similar to the present version of the document. This

contrasts with Adar et al.'s [2] work on the stable and dynamic content of popular documents on the web. While popular documents (e.g. portals, news sites) have a stable content core and a dynamic portion corresponding to the page sections that are regularly updated, in-depth focused articles don't have this structure and dynamics. The evolution of content is cumulative and centered on a principal theme.

Next, we study the influence of document history in term weighting. We define and extensively evaluate four new measures for document term weighting. All the proposed measures explore the document's revision history as an additional signal to improve term discrimination. Based on different evaluation experiments we show that document history is a useful source of information to improve document term weighting. We demonstrate that temporally aware measures, specifically the proposed *revision term frequency* and *revision term frequency span*, outperform the classic term frequency measure. Although we have used Wikipedia, and the full revision history of its articles as a document collection, this work can be easily adapted to other contexts. Consider the case of web search. Given that web search engines periodically crawl the web, they have access to historical information about web documents. This information can be used without difficulty to incorporate time-dependent signals on term weighting functions.

It is worth noting that traditional measures like term frequency are based on a single version of a document (i.e. the current version), thus directly dependent on the latest updates. On the contrary, the proposed time-dependent measures are based on multiple versions of the same document. This results in more robust weighting measures, which are less vulnerable to sporadic changes. This is a valuable quality in the context of shared or public repositories because of the higher resistance to SPAM or other malicious modifications. Nonetheless, this robustness can be seen as a drawback when dealing with naturally fast changing documents like homepages that are continually updated with the latest information.

Finally, we would like to highlight the full reproducibility of the work presented in this chapter. All data, except for the human assessments, is public and freely available.

Chapter 8

Conclusions and Future Work

In this dissertation, we have investigated the impact of time in various information retrieval tasks. The central thesis guiding this work was that the dynamic facet of document collections can be explored to improve current information retrieval tasks.

This chapter outlines our main contributions to the use of temporal features for information retrieval problems. We conclude by discussing some promising directions for future work.

8.1 Summary of Contributions

The discipline of Information Retrieval deals with the problem of finding or devising answers to user informational needs, such as locating a piece of information or preparing a short summary from a group of documents. At the core of most retrieval systems, a multitude of signals extracted from the collection is combined to derive document or term scores. In the large majority of cases, the collection is viewed as a static, fixed group of documents. We relax this assumption and identify, extract and incorporate time-dependent signals into document and term scores. We show that temporal features contain valuable information and can be successfully incorporated into the standard retrieval process.

We have made original contributions to the field of information retrieval by studying the impact of using the dynamic dimension of collections on retrieval tasks. These contributions have been published in peer-reviewed publications (see Section 1.5). Our

contributions encompass several information retrieval tasks, namely: feature extraction, query analysis, link analysis, document authority ranking and term weighting. A summary of the main contributions presented in this dissertation is included next.

In Chapter 3, we survey the temporal features that exist on a collection of web documents and organize these in two principal classes, namely document-based and web-based. The first class includes all features that can be extracted from single documents, ignoring the surrounding context. This class was further structured according to the source of the data, more specifically document-based features can be extracted from a document's content, the document's URL address, or from the HTTP protocol. The class of web-based features is also arranged in sub-classes, specifically temporal features can be extracted from nearby documents, from independent external archives, or from server-side logs.

In Chapter 4, we study a large collection of web search queries and find that explicit temporal expressions exist in 1.5% of the queries. Furthermore, by mapping each temporal query in relation to the user's time of submission, we show that the large majority relates to current (within the same year) or past events. Finally, we show that these expressions are more frequently used in topics such as: autos, sports, news and holidays.

In Chapter 5, we present a method that improves the dating of web documents. We use the header information existing in web documents on the vicinity of the original document to estimate its date. We find that within a collection of randomly sampled web documents, it is possible to determine the date of 50% of them. If we include information from each document's neighborhood, we can date more than 85% of the documents in that same sample.

In Chapter 6, we measure the impact of time in link-based authority estimation algorithms. We base this investigation on a large sample of blogs from a single service provider, spanning a period of 43 months. We find that, in a time-agnostic ranking, the impact of removing either older data or newer data from the collection is alike. Additionally, we show that standard time-independent rankings are unable to capture the correct popularity of web sites with very high citation activity in recent periods. We show that an alternative time-sensitive algorithm, where the weight of links decay over time, produces rankings that are closer to ranks based on the number of visitors to each blog.

In Chapter 7, we present four measures for document term weighting that incorporate

temporal content-based features in their formulas. Two of the measures, revision term frequency and revision term frequency span, consistently outperform the classic term frequency measure in several evaluation experiments. Additionally, we find that the distribution of the revision activity over time is highly correlated with the popularity of a document's topic over the same period. Looking at the progression of document content in a collaborative collection, we find that content in early versions of a document quickly becomes very similar to the present version of the document. Finally, we show that former document content contains relevant keywords that are inaccessible to the retrieval system if only the latest version is considered.

To conclude, we review the proposed thesis defined in Chapter 1. The work presented in this dissertation supports this thesis. We can assert that *the dynamic facet of document collections can be explored to improve current information retrieval tasks*. Next, we comment on the exploratory questions also outlined in the beginning of this dissertation.

Is it possible to obtain rich temporal features from documents?
Yes. In Chapter 3, "Temporal Features on the Web", we make a detailed survey on the possible sources of temporal information on the web, which includes several examples of *document-based* features. In Chapter 5, "Using Neighbors to Date Web Documents", we show that temporal features from a document's neighborhood can be positively used to enrich the original document with temporal data.

Has the history of a document impact on its current importance?
Yes. In Chapter 6, "Link Authority over Time", we see that the history of a document, specifically the progression of citations to it, contains valuable unique information that has direct impact on the document's importance.

Is the evolution of scoring features a stronger signal than snapshot observations of those features?
Yes. In Chapter 7, "Content Dynamics in Retrieval", we propose and evaluate a new term weighting measure that incorporates a document's previous versions. We show that this measure, named revision term frequency, outperforms the classic tf measure which is based on a single observation of the document (the latest).

Is a document's past content relevant to present ranking?
Yes. In Chapter 7, "Content Dynamics in Retrieval", we show that terms that existed on a document but are no longer visible in the document's current version are still important to characterize the document.

8.2 Future Work

We have shown that the historical information available in document collections can improve current retrieval tasks. However, it is our belief that we have only explored a small part of the possibilities opened up by considering document collections as dynamic entities. There are several possible directions for future developments based on the work presented here. These are briefly discussed in the remainder of this section.

Temporal Features in Blogs We have explored the impact of temporal features in two participations on the TREC Blog Track. These experiments are fully described in Appendix B. TREC Blog Track Participations. In these participations we have proposed and studied several temporal features for blog collections. We think that our team's results can be substantially improved by focusing on the feature combination algorithms.

Progress of Signals Over Time We have studied the progression of two important signals over time, namely document in-degree and term frequency. Berberich et al. [18] did a similar study considering the progression of PageRank values. These are just a tiny fraction of all the signals that have been documented in the literature. The study of the progression of IR signals over time is an interesting research path for future work.

Relevance of Former Content We have briefly investigated the importance of former page content for current retrieval tasks. This is an interesting research question worth of a more detailed investigation. We have only looked at the presence of former page content in social bookmarking services. It is important to collect direct user feedback and investigate on the value of this information for different IR tasks, such as document summarization.

Changes Summarization Task While looking at the update profile of Wikipedia documents, we have identified an interesting task worth addressing — the automatic summarization of a partial set of changes. While traditional summarization tasks focus on the summarization of the complete document, this task is focused on the summarization of the changes made during a given period. This would be a helpful aid to a user trying to understand why a given document had a peak of revisions at a given time. As collections grow older and bigger, we think that understanding what happened at a given period will be increasingly important.

Temporal Evidence in Real-Time Media The adoption of mobile devices and personal publishing services (e.g. blogs, microblogs, social networking) has contributed to the emergence of the 'real-time' media. Content published in these sites is intrinsically rich in temporal information. Additionally, content is published in real-time — the time lag between real-word events and media coverage has been drastically reduced. We think that real-time media can become an important source of temporal information. Services such as Twitter or Facebook can be used as external sources of temporal information, adding layers of temporal hints to data.

Interactive Manipulation of Temporal Impact Incorporating time into relevance assessment algorithms is not a trivial task. Determining the value of time for a given user information need is challenging. A simple query like 'portuguese elections' can enclose contrasting temporal meanings. The user might be interested in the latest elections, in the upcoming ones or simply in a broad overview of Portuguese electoral acts. We think that there are many opportunities in the field of user interfaces, specifically for the interactive manipulation of temporal impact.

Bibliography

[1] TREC-BLOG - Information Retrieval Wiki. URL `http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG` (accessed July 22, 2010).

[2] ADAR, E., TEEVAN, J., DUMAIS, S. T., AND ELSAS, J. L. The Web Changes Everything: Understanding the Dynamics of Web Content. In *WSDM'09: Proceedings of the Second ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2009), ACM, pp. 282–291.

[3] ALLAN, J. *Topic Detection and Tracking: Event-based Information Organization*, vol. 12 of *The Information Retrieval Series*. Springer, 2002.

[4] ALONSO, O. *Temporal Information Retrieval*. PhD thesis, University of California, Davis, California, August 2008.

[5] ALONSO, O., GERTZ, M., AND BAEZA-YATES, R. On the Value of Temporal Information in Information Retrieval. *ACM SIGIR Forum 41*, 2 (December 2007), 35–41.

[6] AMITAY, E., CARMEL, D., HERSCOVICI, M., LEMPEL, R., AND SOFFER, A. Trend Detection Through Temporal Link Analysis. *J. Am. Soc. Inf. Sci. Technol. 55*, 14 (December 2004), 1270–1281.

[7] BAEZA-YATES, R., CASTILLO, C., AND SAINT-JEAN, F. Web Dynamics, Structure and Page Quality. In *Web Dynamics*, M. Levene and A. Poulovassilis, Eds. Springer Verlag, 2004, pp. 93–109.

[8] BAEZA-YATES, R., AND NETO, B. R. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[9] BALUJA, S., RAVICHANDRAN, D., AND SIVAKUMAR, D. Text Classification Through Time: Efficient Label Propagation in Time-Based Graphs. In *Proceeding of the International Conference on Knowledge Discovery and Information Retrieval (KDIR 2009)* (October 2009), INSTICC.

[10] BAR-ILAN, J. Comparing Rankings of Search Results on the Web. *Inf. Process. Manage. 41*, 6 (December 2005), 1511–1519.

[11] BAR-ILAN, J. Access to Query Logs - An Academic Researcher's Point of View. In *Query Log Analysis: Social And Technological Challenges. A Workshop at the 16th International World Wide Web Conference (WWW 2007)* (May 2007), E. Amitay, G. C. Murray, and J. Teevan, Eds.

[12] BAR-YOSSEF, Z., BRODER, A. Z., KUMAR, R., AND TOMKINS, A. Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. In *WWW '04: Proceedings of the 13th international conference on World Wide Web* (New York, NY, USA, 2004), ACM, pp. 328–337.

[13] BAR-YOSSEF, Z., AND GUREVICH, M. Random Sampling from a Search Engine's Index. In *WWW'06: Proceedings of the 15th International Conference on World Wide Web* (New York, NY, USA, 2006), ACM, pp. 367–376.

[14] BEITZEL, S. M., JENSEN, E. C., CHOWDHURY, A., GROSSMAN, D., AND FRIEDER, O. Hourly Analysis of a Very Large Topically Categorized Web Query Log. In *SIGIR'04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2004), ACM, pp. 321–328.

[15] BEITZEL, S. M., JENSEN, E. C., FRIEDER, O., LEWIS, D. D., CHOWDHURY, A., AND KOLCZ, A. Improving Automatic Query Classification via Semi-Supervised Learning. In *ICDM'05: Proceedings of the Fifth IEEE International Conference on Data Mining* (Houston, Texas, USA, November 2005), IEEE Computer Society, pp. 42–49.

[16] BENT, L., RABINOVICH, M., VOELKER, G. M., AND XIAO, Z. Characterization of a Large Web Site Population with Implications for Content Delivery. In *WWW'04: Proceedings of the 13th international conference on World Wide Web* (New York, NY, USA, 2004), ACM Press, pp. 522–533.

[17] BERBERICH, K. *Temporal Search in Web Archives*. PhD thesis, Max-Planck-Institut für Informatik, Saarbrücken, Germany, July 2010.

[18] BERBERICH, K., BEDATHUR, S., VAZIRGIANNIS, M., AND WEIKUM, G. BuzzRank ... and the Trend is Your Friend. In *WWW'06: Proceedings of the 15th international conference on World Wide Web* (New York, USA, May 2006), University of Southampton, United Kingdom, ACM Press.

[19] BERBERICH, K., VAZIRGIANNIS, M., AND WEIKUM, G. T-Rank: Time-Aware Authority Ranking. In *Algorithms and Models for the Web-graph : Third International Workshop, WAW 2004* (Berlin, Germany, January 2004), vol. 3243 of *Lecture Notes in Computer Science*, Springer, pp. 131–142.

[20] BREWINGTON, B. E., AND CYBENKO, G. How Dynamic is the Web? *Comput. Netw. 33*, 1-6 (June 2000), 257–276.

[21] BRODER, A. Z., GLASSMAN, S. C., MANASSE, M. S., AND ZWEIG, G. Syntactic Clustering of the Web. In *Selected Papers from the 6th International Conference on World Wide Web* (Essex, UK, 1997), Elsevier Science Publishers Ltd., pp. 1157–1166.

[22] BURIOL, L. S., CASTILLO, C., DONATO, D., LEONARDI, S., AND MILLOZZI, S. Temporal Analysis of the Wikigraph. In *WI'06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 45–51.

[23] BUSH, V. As We May Think. *The Atlantic Monthly 176*, 1 (1945), 101–108.

[24] CHEN, Y.-Y., GAN, Q., AND SUEL, T. Local Methods for Estimating PageRank Values. In *CIKM'04: Proceedings of the 13th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2004), ACM, pp. 381–389.

[25] CHI, Y., ZHU, S., SONG, X., TATEMURA, J., AND TSENG, B. L. Structural and Temporal Analysis of the Blogosphere Through Community Factorization. In *KDD'07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2007), ACM, pp. 163–172.

[26] CHO, J., AND GARCIA-MOLINA, H. The Evolution of the Web and Implications for an Incremental Crawler. In *VLDB'00: Proceedings of the 26th International Conference on Very Large Data Bases* (San Francisco, CA, USA, 2000), Morgan Kaufmann Publishers Inc., pp. 200–209.

[27] CLAUSEN, L. Concerning Etags and Datestamps. In *4th International Web Archiving Workshop (IWAW'04)* (2004), J. Masanès and A. Rauber, Eds.

[28] COBURN, A. Lingua::EN::Tagger - search.cpan.org. URL `http://search.cpan.org/perldoc?Lingua::EN::Tagger` (accessed October 27, 2009).

[29] CRASWELL, N., ROBERTSON, S., ZARAGOZA, H., AND TAYLOR, M. Relevance Weighting for Query Independent Evidence. In *SIGIR'05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2005), ACM Press, pp. 416–423.

[30] CROWDFLOWER. CrowdFlower - Docs - Gold. URL `http://crowdflower.com/docs/gold` (accessed July 22, 2010).

[31] DOUGLIS, F., FELDMANN, A., AND KRISHNAMURTHY, B. Rate of Change and other Metrics: a Live Study of the World Wide Web. In *USENIX Symposium on Internet Technologies and Systems* (December 1997).

[32] DUMAIS, S., CUTRELL, E., CADIZ, J. J., JANCKE, G., SARIN, R., AND ROBBINS, D. C. Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. In *SIGIR'03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (New York, NY, USA, 2003), ACM, pp. 72–79.

[33] ELSAS, J. L., AND DUMAIS, S. T. Leveraging Temporal Dynamics of Document Content in Relevance Ranking. In *WSDM'10: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining* (New York, NY, USA, February 2010), ACM, ACM, pp. 1–10.

[34] ERNSTING, B., WEERKAMP, W., AND DE RIJKE, M. Language Modeling Approaches to Blog Post and Feed Finding. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings* (2007).

[35] FERRO, L., MANI, I., SUNDHEIM, B., AND WILSON, G. TIDES Temporal Annotation Guidelines - Version 1.0.2. Tech. rep., The MITRE Corporation, McLean, Virginia, June 2001.

[36] FETTERLY, D., MANASSE, M., NAJORK, M., AND WIENER, J. L. A Large-Scale Study of the Evolution of Web Pages. *Softw. Pract. Exper. 34*, 2 (February 2004), 213–237.

[37] GIBSON, D., PUNERA, K., AND TOMKINS, A. The Volume and Evolution of Web Page Templates. In *WWW'05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web* (New York, NY, USA, 2005), ACM Press, pp. 830–839.

[38] GOMES, D., AND SILVA, M. J. Modelling Information Persistence on the Web. In *ICWE'06: Proceedings of the 6th International Conference on Web Engineering* (New York, NY, USA, 2006), ACM Press, pp. 193–200.

[39] GRIMES, C. Microscale Evolution of Web Pages. In *WWW'08: Proceeding of the 17th International Conference on World Wide Web* (New York, NY, USA, 2008), ACM, pp. 1149–1150.

[40] HANANI, U., SHAPIRA, B., AND SHOVAL, P. Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction 11*, 3 (August 2001), 203–259.

[41] HE, B., MACDONALD, C., AND OUNIS, I. Retrieval Sensitivity Under Training Using Different Measures. In *SIGIR'08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2008), ACM, pp. 67–74.

[42] HEARST, M. A., HURST, M., AND DUMAIS, S. T. What Should Blog Search Look Like? In *Proceeding of the 2008 ACM workshop on Search in Social Media (SSM'08)* (New York, NY, USA, 2008), I. Soboroff, E. Agichtein, R. Kumar, I. Soboroff, E. Agichtein, and R. Kumar, Eds., ACM, ACM, pp. 95–98.

[43] HENZINGER, M. Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms. In *SIGIR'06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2006), ACM, pp. 284–291.

[44] HOWE, J. The Rise of Crowdsourcing. *Wired magazine 14*, 6 (June 2006).

[45] INTERNET ARCHIVE. Internet Archive. URL `http://www.archive.org` (accessed October 14, 2009).

[46] JATOWT, A., AND ISHIZUKA, M. Temporal Web Page Summarization. In *WISE 2004* (2004), Lecture Notes in Computer Science, Springer-Verlag, pp. 303–312.

[47] JATOWT, A., AND ISHIZUKA, M. Temporal Multi-Page Summarization. *Web Intelli. and Agent Sys. 4*, 2 (April 2006), 163–180.

[48] JONES, K. S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation 28*, 1 (1972), 11–21.

[49] KENDALL, M. G. A New Measure of Rank Correlation. *Biometrika 30*, 1/2 (June 1938), 81–93.

[50] KIM, S. J., AND LEE, S. H. An Empirical Study on the Change of Web Pages. In *Web Technologies Research and Development - APWeb 2005*. Springer Berlin / Heidelberg, 2005, pp. 632–642.

[51] KITTUR, A., CHI, E. H., AND SUH, B. Crowdsourcing User Studies with Mechanical Turk. In *CHI'08: Proceeding of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2008), ACM, pp. 453–456.

[52] KLEINBERG, J. M. K. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM 46*, 5 (September 1999), 604–632.

[53] KOEHLER, W. A Longitudinal Study of Web Pages Continued: A Report After Six Years. *Information Research 9*, 2 (2004).

[54] LEMPEL, R., AND MORAN, S. SALSA: The Stochastic Approach for Link-Structure Analysis. *ACM Trans. Inf. Syst. 19*, 2 (April 2001), 131–160.

[55] LIEBSCHER, R., AND BELEW, R. Lexical Dynamics and Conceptual Change: Analyses and Implications for Information Retrieval. *Cognitive Science Online 1* (2003).

[56] LIEBSCHER, R., AND BELEW, R. K. Temporal Feature Modification for Retrospective Categorization. In *FeatureEng'05: Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing* (Morristown, NJ, USA, 2005), Association for Computational Linguistics, pp. 17–23.

[57] LIN, Y.-R., SUNDARAM, H., CHI, Y., TATEMURA, J., AND TSENG, B. L. Splog Detection Using Self-similarity Analysis on Blog Temporal Dynamics. In *AIRWeb'07: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web* (New York, NY, USA, 2007), ACM Press, pp. 1–8.

[58] MACDONALD, C., AND OUNIS, I. The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection. Tech. rep., Department of Computing Science, University of Glasgow, Scotland, United Kingdom, 2006.

[59] MANI, I., PUSTEJOVSKY, J., AND SUNDHEIM, B. Introduction to the Special Issue on Temporal Information Processing. *ACM Transactions on Asian Language Information Processing (TALIP) 3*, 1 (March 2004), 1–10.

[60] MANI, I., AND WILSON, G. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)* (Hong Kong, 2000), pp. 69–76.

[61] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, 2007.

[62] MARKTEST. Netscope - Sistema site-centric de mediação de tráfico internet. URL `http://www.netscope.marktest.pt` (accessed January 18, 2010).

[63] MCCOWN, F., AND NELSON, M. L. Agreeing to Disagree: Search Engines and Their Public Interfaces. In *JCDL'07: Proceedings of the 2007 Conference on Digital libraries* (New York, NY, USA, 2007), ACM Press, pp. 309–318.

[64] MCNEMAR, Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika 12*, 2 (June 1947), 153–157.

[65] MEADOW, C. T., BOYCE, B. R., KRAFT, D. H., AND BARRY, C. L. *Text Information Retrieval Systems*, third ed. Academic Press, March 2007.

[66] NAJORK, M. A., ZARAGOZA, H., AND TAYLOR, M. J. HITS on the Web: How does it Compare? In *SIGIR'07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2007), ACM, pp. 471–478.

[67] NTOULAS, A., CHO, J., AND OLSTON, C. What's New on the Web?: The Evolution of the Web from a Search Engine Perspective. In *WWW'04: Proceedings of the 13th International Conference on World Wide Web* (New York, NY, USA, 2004), ACM Press, pp. 1–12.

[68] NUNES, S. Exploring Temporal Evidence in Web Information Retrieval. In *BCS IRSG Symposium Future Directions in Information Access (FDIA 2007)* (August 2007), A. MacFarlane, L. Azzopardi, and I. Ounis, Eds., BCS IRSG, BCS IRSG, pp. 44–50.

[69] NUNES, S., RIBEIRO, C., AND DAVID, G. FEUP at TREC 2008 Blog Track: Using Temporal Evidence for Ranking and Feed Distillation. In *17th Text REtrieval Conference (TREC 2008)* (November 2008), E. M. Voorhees and L. P. Buckland, Eds., NIST.

[70] OUNIS, I., AMATI, G., PLACHOURAS, V., HE, B., MACDONALD, C., AND LIOMA, C. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)* (2006).

[71] PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford InfoLab, November 1999.

[72] PASS, G., CHOWDHURY, A., AND TORGESON, C. A Picture of Search. In *InfoScale'06: Proceedings of the 1st International Conference on Scalable Information Systems* (New York, NY, USA, 2006), ACM, pp. 1+.

[73] ROBERTSON, S., WALKER, S., BEAULIEU, M. M., GATFORD, M., AND PAYNE, A. Okapi at TREC-4. In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)* (1995), pp. 73–96.

[74] SILVERSTEIN, C., MARAIS, H., HENZINGER, M., AND MORICZ, M. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum 33*, 1 (1999), 6–12.

[75] SUGIYAMA, K., HATANO, K., YOSHIKAWA, M., AND UEMURA, S. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In *HYPERTEXT'03: Proceedings of the 14th ACM Conference on Hypertext and hypermedia* (New York, NY, USA, 2003), ACM Press, pp. 198–207.

[76] THOMAS, C., AND SHETH, A. P. Semantic Convergence of Wikipedia Articles. In *WI'07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 600–606.

[77] TUFTE, E. R. *Beautiful Evidence*, first edition ed. Graphics Press, 2006.

[78] WIKIPEDIA CONTRIBUTORS. July 2006 Java earthquake - Wikipedia, the free encyclopedia. URL `http://en.wikipedia.org/wiki/July_2006_Java_earthquake` (accessed October 7, 2009).

[79] WIKIPEDIA CONTRIBUTORS. Wikipedia - Wikipedia, the free encyclopedia. URL `http://en.wikipedia.org/wiki/Wikipedia` (accessed May 5, 2008).

[80] WIKIPEDIA CONTRIBUTORS. Wikipedia:Manual of Style - Wikipedia, the free encyclopedia. URL `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style` (accessed July 5, 2010).

[81] WONG, K.-F., XIA, Y., LI, W., AND YUAN, C. An Overview of Temporal Information Extraction. *International Journal of Computer Processing of Oriental Languages 18*, 2 (2005), 137–152.

[82] YAHOO! Random Link - Yahoo! URL `http://random.yahoo.com/bin/ryl` (accessed October 7, 2009).

[83] YAHOO! Yahoo! Developer Network Home. URL `http://developer.yahoo.com` (accessed October 7, 2009).

[84] YAHOO! Yahoo! Directory. URL `http://dir.yahoo.com` (accessed October 7, 2009).

[85] YAMAMOTO, Y., TEZUKA, T., JATOWT, A., AND TANAKA, K. Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis. In *9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management* (2007), Lecture Notes in Computer Science, Springer Verlag.

[86] YANG, L., QI, L., ZHAO, Y. P., GAO, B., AND LIU, T. Y. Link Analysis using Time Series of Web Graphs. In *CIKM'07: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (New York, NY, USA, 2007), ACM, pp. 1011–1014.

[87] YU, P. S., LI, X., AND LIU, B. On the Temporal Dimension of Search. In *WWW Alt.'04: Proceedings of the 13th International World Wide WebCconference on Alternate Track Papers & Posters* (New York, NY, USA, 2004), ACM Press, pp. 448–449.

[88] ZHANG, G.-Q., ZHANG, G.-Q., YANG, Q.-F., CHENG, S.-Q., AND ZHOU, T. Evolution of the Internet and Its Cores. *New Journal of Physics 10* (December 2008).

[89] ZUBIAGA, A. Enhancing Navigation on Wikipedia with Social Tags. In *Wikimania 2009* (August 2009).

Appendix A

WIKICHANGES

In this appendix, we present WIKICHANGES, a prototype web application that we have developed to explore, compare and analyze revision timelines. We developed this web-based software system to produce update profiles for Wikipedia articles. Perl was the primary programming language used and the graphics are based on amCharts [1], a dynamic charting library in Flash.

When a query is presented to the system, the first step is the association of the query with a single article. This is done by searching a local file containing the names of all articles in Wikipedia. Then, the system checks if the local cache already has a copy of the updates to the article. If not, a background process is started to access the Wikipedia API and download the revision history for the given article. Since this is usually a long process, the user is informed and redirected to a temporary web page. If, on the other hand, a local copy is found, the system only downloads the latest updates while the user waits for the web page to finish loading. Before redirecting the user to the final web page, aggregated update values by month and day are calculated and written to the cache. Since WIKICHANGES is used in a multi-user environment, it was necessary to implement a simple concurrency control feature to avoid simultaneously downloading the same files. The system's homepage is presented in Figure A.1, some additional examples of side-by-side plots are presented in figures A.2 and A.3.

The amCharts library is very flexible and has a number of useful built-in features. It's possible to navigate through the graph and have access to the individual values in each dataset as dynamic tooltips. The user is allowed to zoom in a portion of the data by

1. http://www.amcharts.com/

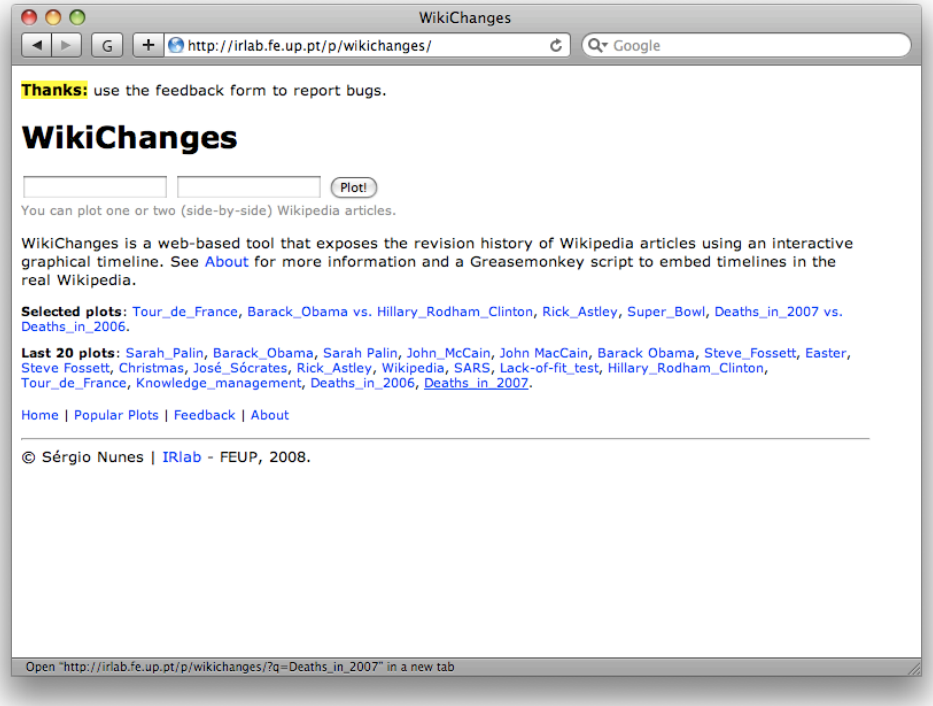

Figure A.1: WikiChanges homepage screenshot.

selecting a specific range using the mouse. Also, it is possible to associate an URL with each single data point. We've used this feature to allow the user to view the automatic summary for a specific period. If the user clicks on a data point, he is redirected to a page where the month's summary is shown. Figure A.4 shows the Wikipedia article on the late USA adventurer *Steve Fossett* as viewed in WIKICHANGES. The summary shown for February 2006 is presented as a *tag cloud* in the bottom of the screen. This summary is automatically generated online using the algorithm described in the previous section. Due to the simplicity of the algorithm, this is feasible in real-time without caching the results. The size of each term is plotted in a logarithmic scale based on the term score. In the example shown, the terms highlighted are related to the record for the longest flight 'without landing' by any aircraft in history (at the time). It is important to note that, since terms are both unigrams and bigrams, some words appear multiple times within the tag cloud. For instance, the word *"world"* occurs isolated but also grouped with *"record"*. This tag cloud is a good example of an automatic summary produced by our algorithm. However, it is important to note that most tests conducted returned ambiguous or generic summaries.

To increase the visibility of an article's update profile we developed a Gresemonkey extension. Greasemonkey [2] is a plug-in for the Firefox web browser that allows the user to "*customize the way webpages look and function*". Our Greasemonkey extension adds a sparklines to Wikipedia article pages as shown in Figure A.5. Sparklines are "*small, high resolution graphics embedded in a context of words, numbers, images*" proposed by Tufte [77] to succinctly present data where it is discussed. We opted to show the sparklines next to the article's title for higher visibility and context. The sparkline was built using WIKICHANGES and the Google Chart API [3] for improved response time. The interactive WIKICHANGES tool and the Greasemonkey script are available online at `http://irlab.fe.up.pt/p/wikichanges`.

2. `http://www.greasespot.net/`
3. `http://code.google.com/apis/chart`

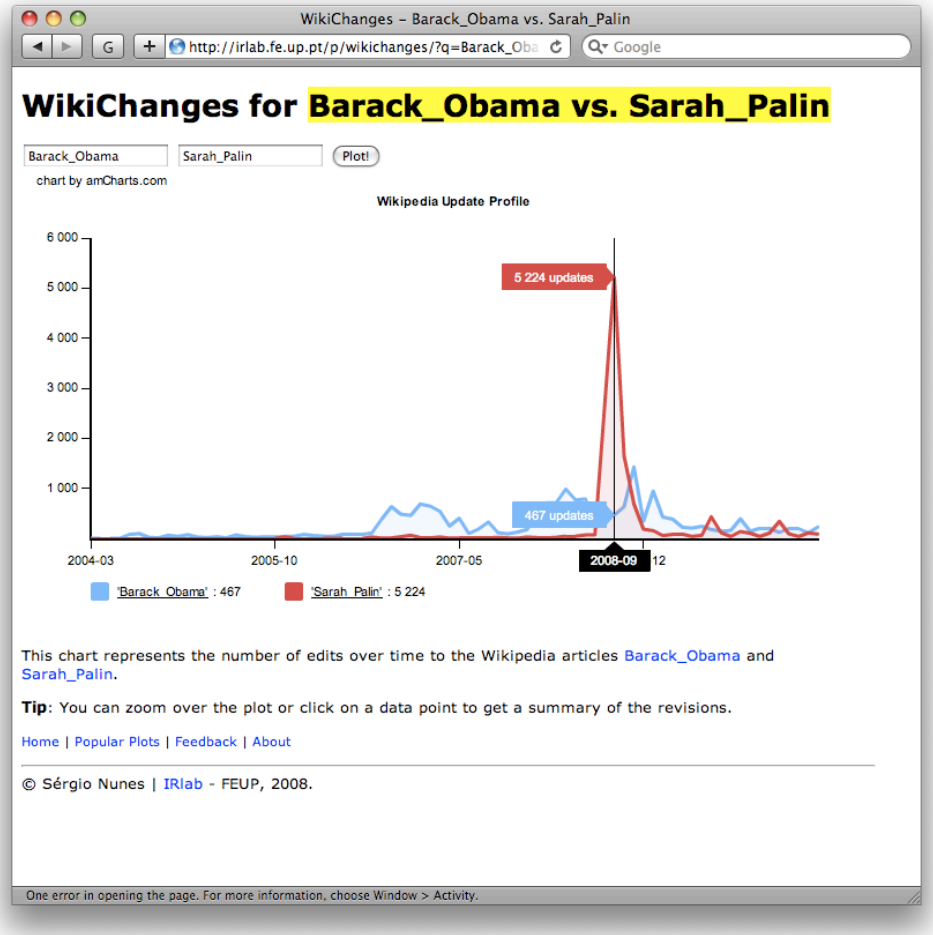

Figure A.2: WikiChanges screenshot for the articles on *Barack Obama* and *Sarah Palin*.

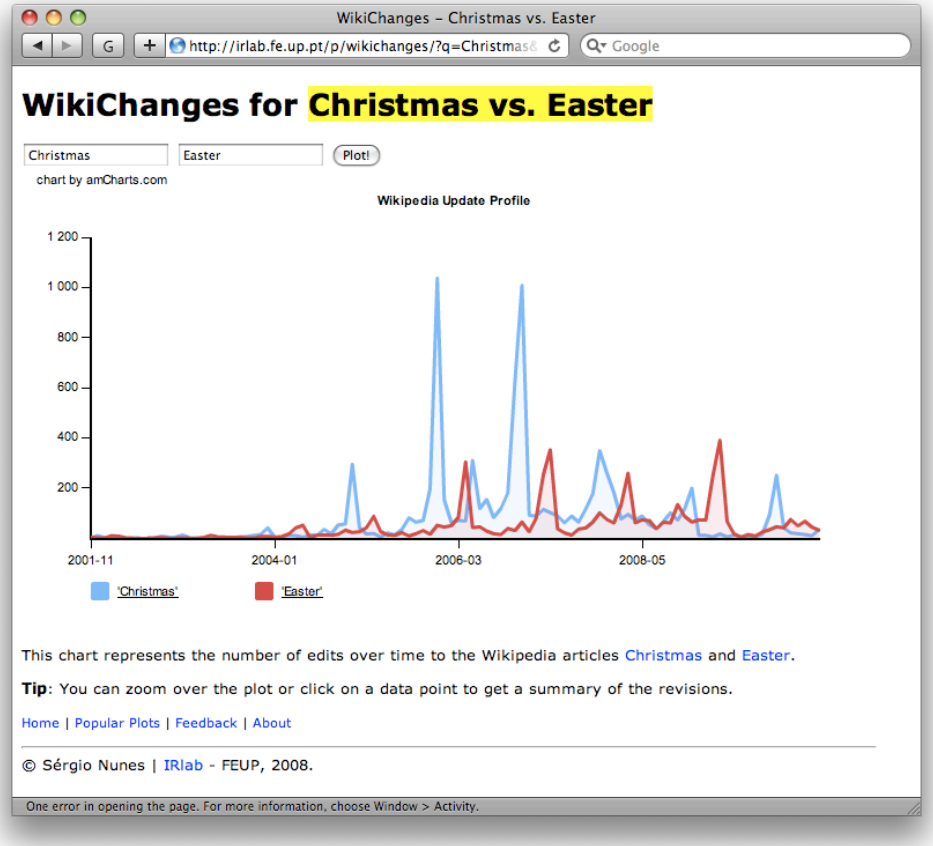

Figure A.3: WikiChanges screenshot for the articles on *Christmas* and *Easter*.

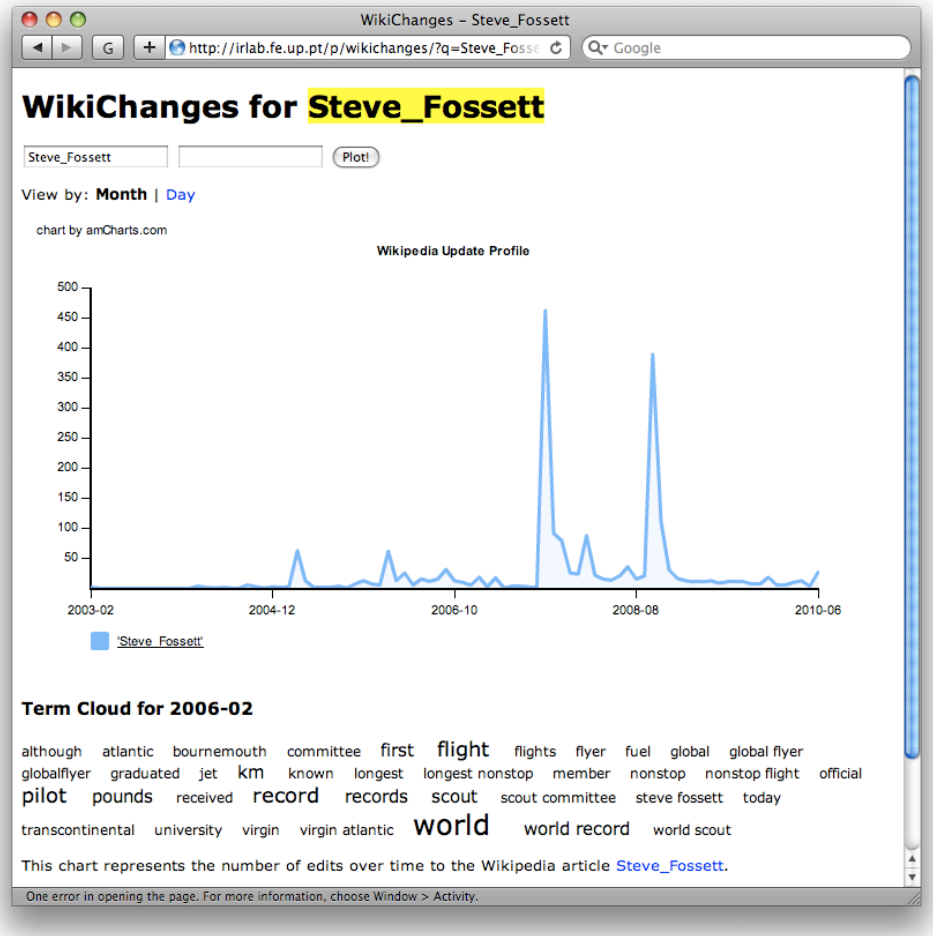

Figure A.4: WikiChanges screenshot for the article on *Steve Fossett*, including an automatic summary.

Figure A.5: Screenshot of WikiChanges sparkline extension embedded in Wikipedia.

Appendix B

TREC Blog Track Participations

B.1 Introduction

In this appendix we include the details about our participation at TREC's Blog Track. The Blog Track was first introduced in TREC 2006 with the main goal of exploring "the information seeking behaviour in the blogosphere" [1]. We have participated in two editions of this track, each one using a different dataset.

B.2 TREC 2008 Blog Track

This section presents the participation of FEUP in the TREC 2008 Blog Track. FEUP participated in two tasks, the baseline adhoc retrieval task and the blog finding distillation task. Our approach was focused on the use of the temporal information available in the TREC Blog06 collection. For the baseline adhoc retrieval task a simple temporal sort was evaluated. In the blog finding distillation task we tested three alternative scoring functions based on temporal evidence. All features were combined with a BM25 baseline run using a standard rank aggregation approach. We observed small, but statistically significant, improvements in several evaluation measures when temporal information is used.

The Blog Track uses the TREC Blog06 [58] collection, a large sample of documents crawled from the blogosphere. The collection contains more than 100,000 feeds of blogs and over 3.2 million permalink documents, both crawled over an eleven week period.

The TREC 2008 Blog Track edition was structured in four distinct tasks: baseline adhoc retrieval, opinion finding, polarized opinion finding and blog finding distillation. We did not participate in any of the opinion retrieval tasks. Thus, considering the adhoc and distillation tasks and the number of submissions allowed, we defined the goals for our participation as follows.

For the adhoc retrieval task the unit of retrieval is the post, while for the blog distillation it is the feed. In the adhoc retrieval task we simply judge if the ordering of posts by publication date is a relevant criteria. For the blog distillation task we evaluated if the temporal dispersion of posts within a given feed is a positive criteria for retrieval. These ideas can be summarized in the following research questions:

– *Is the temporal order of documents a relevant criterion for adhoc retrieval?*
– *Is the temporal dispersion of relevant posts within a single feed a relevant criterion for feed distillation?*

B.2.1 System Overview

We used the Terrier information retrieval platform [70] for all our experiments. Terrier is a state-of-the-art system with a very simple installation procedure and advanced customization features. Also, it supports the indexing of standard TREC collections natively, which greatly reduced the initial setup overhead. The TREC Blog06 collection is structured in feeds (∼39GB), permalinks (∼89GB) and homepages (∼29GB). We used Terrier to build an index based only on the permalink documents, excluding the following TREC tags: `DOCHDR`, `FEEDNO`, `BLOGHPNO`, `BLOGHPURL`, `PERMALINK`.

For the topics defined in each task, we retrieved a set of documents using Terrier's implementation of the BM25 model [73]. We used the default parameters defined in Terrier: $k_1 = 1.2d$, $k_3 = 8d$ and $b = 0.75d$. This set of results was then used in the adhoc and distillation tasks as described in sections B.2.3 and B.2.4. All our retrieval experiments were made using this setup. It is worth noting that we did not made any effort to remove SPAM from the collection. The TREC Blog06 has been intentionally injected with SPAM to better reflect a realistic setting [58]. However, we opted to ignore SPAM since we believed that the presence of SPAM would not prevent us from addressing our initial research questions.

B.2.2 Relevance Over Time

In the Blog06 collection, from the 3.2 million permalink documents available, almost 2 million have a date within the period of the collection (~60%). This is an encouraging figure for research based on the temporal properties of these documents. The temporal information associated with each permalink document is included in the collection. Dates were derived directly from the feeds and validated using the crawl date. We ignored high granularity information like hours, minutes and seconds, and only consider days as a unit of time.

Using the qrels from the TREC Blog Track, we first observed how the absence of date information was reflected in the distribution of both relevant and non-relevant posts. We discarded polarity information available on the qrels (i.e. positive, negative opinion) since we are only interested in the relevance of documents. In Figure B.1 we present the distribution of missing temporal information over relevance in all editions of the track. For instance, in the 2006 qrels, 23% of all relevant documents have no date information attached to them, in contrast 40% of all non-relevant documents have no date information. Non-relevant blog posts tend to have missing or invalid date information.

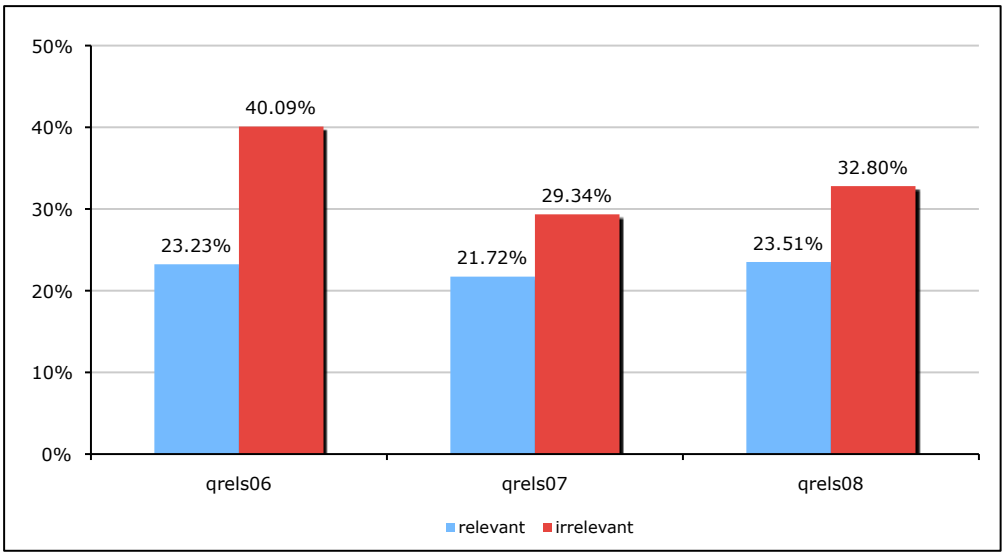

Figure B.1: Documents without temporal information.

Considering the judged documents containing temporal information, we extracted the percentage of both relevant and non-relevant documents for each day of the collection. For instance, in the 57th day of the collection (January 31st 2006) we found 2.79% of all

relevant documents from the 2006 qrels. In contrast, we found 1.28% of all non-relevant documents in that same day. This results in a difference of 1.51%. Figure B.2 shows this difference for each one of the 77 days of the collection. A trend line is also included in the plot showing the growth over time. We found similar trends using qrels from 2007 and 2008. This observation supports the idea that the recency of documents is a positive factor for document relevance.

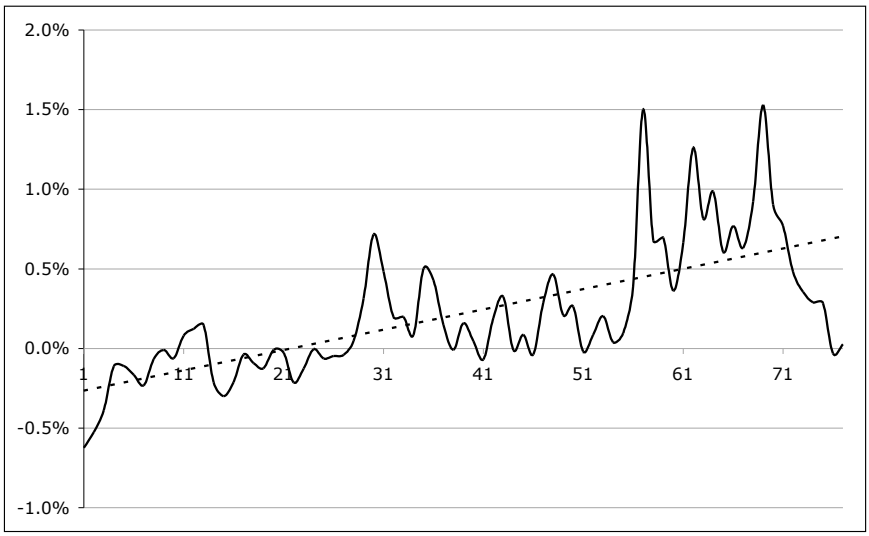

Figure B.2: Difference between relevant and non-relevant posts over time. Oldest days are on the left side of the plot.

To further support this claim, we created runs using topics from all editions of the TREC Blog Track and computed MAP values based on the official qrels. These runs are exclusively based on temporal information. For each track edition we computed two runs, one ordered by oldest posts first and another ordered by newest posts first. Results from this experiment are listed in Table B.1. All improvements observed in runs ordered by newest first are statistically significant at a level of 0.05 using a paired sample t-test.

Table B.1: MAP values for temporally-ordered ranks.

	2006	2007	2008
newest first	**0.1321**	**0.1145**	**0.0985**
oldest first	0.1162	0.0841	0.0823

B.2.3 Baseline Adhoc Retrieval

The baseline adhoc retrieval task is a classic retrieval task, where the goal is to find all relevant information (blog posts) about a given topic. No opinion-finding techniques should be used in this task. We submitted two runs for this task. The first run is a standard automatic title-only run using Terrier with the BM25 model (see previous section). The second run is a combination of the BM25 run and a temporal score based on the post's publish date.

Typically, each blog post has a timestamp representing the date when the text was published. We defined the temporal score as a value between 0 and 1 computed by a linear transformation of each timestamp. To determine if older or newer posts are more relevant for adhoc retrieval, we built two different temporal ranks (ranked by newest and ranked by oldest). For instance, given that the collection spans from December 6th 2005 to February 21st 2006, a post published on the 1st of January 2006 would have a temporal score of $\frac{26days}{77days} = 0.34$ (ranked by newest) or a score of $\frac{51days}{77days} = 0.66$ (ranked by oldest).

As mentioned previously, in the TREC Blog06 collection approximately 60% of the posts have a valid date. We have discarded from the temporal ranks all posts not containing a valid timestamp (either missing or out of bounds).

To test if the timestamp of posts is a valid feature for adhoc ranking, we combined the initial BM25 rank with the two temporal ranks using a standard rank aggregation formula (Equation B.1). We discarded all score information in this approach.

$$\alpha \times rank_{bm25} + (1 - \alpha) \times rank_{temporal} \tag{B.1}$$

The α parameter was determined using data from the 2007 edition of the Blog Track. With $\alpha = 0.99$ and a temporal rank ordered by newest first, both P@20 and R-prec exhibited small improvement with 2007 topics and qrels. We submitted both the BM25

Table B.2: Adhoc Retrieval Task runs.

Run	MAP	R-prec	b-Pref	P@20
BM25	0.2482	0.3214	0.3454	0.5243
BM25T	**0.2483**	**0.3225**	**0.3455**	**0.5280**

baseline rank and this combined rank to the baseline task.

Unfortunately, after submission, we detected a flaw in the combined run. Thus, this approach was not directly evaluated in TREC's assessment procedures. Nevertheless, after correcting the initial bug, we used the results produced by TREC assessors to conduct local evaluations and recreate the flawed run (named **BM25T**). The results are summarized in Table B.2.

In MAP and b-Pref we observed only a very small improvement (0.03% in both) not statistically significant. The improvement observed in P@20 (0.7%) is statistically significant at a level of 0.1 ($p = 0.093$). Finally, the improvement verified in R-prec (0.34%) is statistically significant at a level of 0.05 ($p = 0.046$).

Despite the small improvements, these results confirm our initial expectations that basic temporal information is a positive criterion for adhoc retrieval. In Section B.2.6 we discuss these results in more detail and propose ideas for future research.

B.2.4 Blog Finding Distillation

The blog distillation (or feed search) task consists in identifying blogs with a principle, recurring interest in a given topic. We defined a baseline run for this task by combining each post score into a feed score. Using the initial BM25 score for each blog post and topic, we calculated a feed score by adding all post scores (from the feed) and dividing by the number of posts available in the collection for that same feed. This approach was submitted as a run named **feupbase**.

Given the limit of four runs by team for this task, we opted to test two different strategies using temporal features, both individually and combined. In the following sections we describe these approaches, presenting the ideas for the sources of temporal evidence and how they have been combined. In this section, we present both ideas that make use of temporal evidence and show how we combined these two approaches in a mixed setting. Finally, in Section B.2.4 a brief overview of the results obtained in this task is

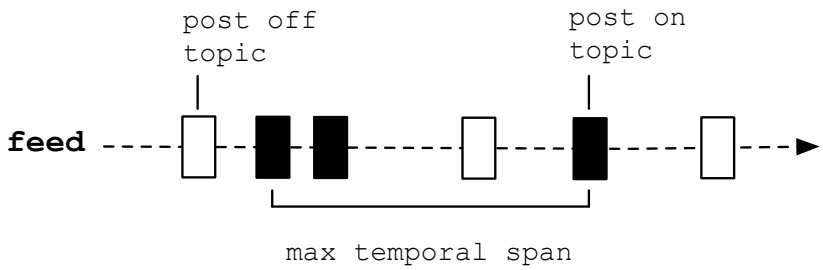

Figure B.3: Temporal Span.

presented.

Using Temporal Span

A blog (or feed) can be seen as a sequence of temporally ordered texts. Some texts will be relevant for a given topic, while others won't. One of our first ideas was to evaluate if the maximum temporal span covered by the relevant posts is a positive criteria in the task of blog distillation.

In Figure B.3 we illustrate this simple idea. The feed depicted in the figure has six posts, three relevant to the topic (in black) and three not relevant (in white). The **temporal span** of a topic in a feed corresponds to the period between the newest relevant post and the oldest relevant post. As noted, this proposed feature is topic-dependent.

Having a list of feeds ranked by temporal span for each topic, we combine this rank with the baseline rank based on BM25. We used a standard rank aggregation approach as defined by Equation B.1. To choose α we used data from the TREC 2007 Blog Track edition (topics and qrels) and optimized for b-Pref. The run combining the BM25 baseline and the temporal span was submitted with reference **feupfs**.

Using Temporal Dispersion

Still focused on the temporal properties of a feed, we investigated how the dispersion of relevant posts in a feed would impact the feed distillation task. Consider the two feeds depicted in Figure B.4, where only relevant posts are included. In the top feed (*feed a*) the relevant posts are less dispersed than the posts in *feed b*. Which pattern is

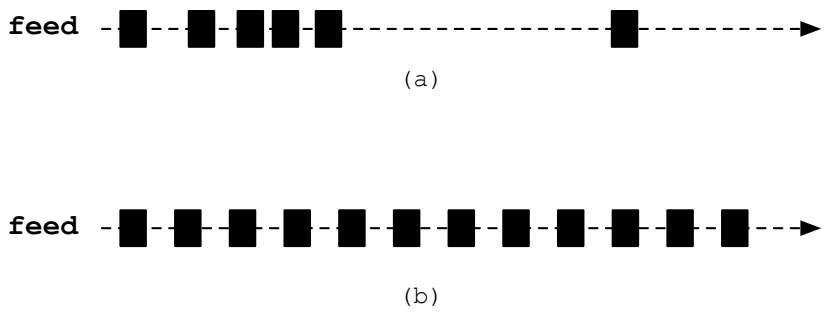

Figure B.4: Examples of Temporal Dispersion.

more relevant for the feed distillation task? Is the dispersion of relevant posts over time a property of relevant blogs?

We opted to use **negentropy** (or negative entropy) as a feature to represent the temporal dispersion of relevant posts in a feed. Negentropy is a measure used in information theory to represent the distance to normality. Negentropy is always positive and reaches its minimum for a gaussian random variable. We tested several alternative measures using data from previous editions of this track (e.g. Kurtosis, Skewness). The best results were obtained using the negentropy-based measure.

First, each post's publish date was converted to a relative global scale between 0 and 1. Each date was converted to the number of days since the beginning of the collection (6th of December 2005) and then divided by the total number of days in the collection (77 days). For each feed we obtained an ordered set of values between 0 and 1, corresponding to the publish dates of the relevant posts. Consider $p(i)$ to be the intervals between the subsequent values in this set. The negentropy of the relevant posts of a feed is given by Equation B.2, where N is the total number of intervals between relevant posts (equal to the total number of relevant posts minus one).

$$Negen = -1 \times \frac{\sum_{i=1}^{N} p(i) \times ln(p(i))}{ln(N)} \tag{B.2}$$

As an example, consider a feed containing four relevant posts published at the following (normalized) times: 0.3, 0.4, 0.5, 0.5. The intervals between the values are, respectively: $p(1) = 0.1$, $p(2) = 0.1$, $p(3) = 0$. The negentropy value for this set of posts is shown below [1].

1. Note that we have considered that $0 \times ln(0) = 0$.

$$-1 \times \frac{0.1 \times ln(0.1) + 0.1 \times ln(0.1) + 0 \times ln(0)}{ln(3)} = 0.42$$

In the example shown in Figure B.4, the negentropy (i.e. dispersion) of *feed b* would be greater than the negentropy of *feed a*.

This feature was combined with the base BM25 rank using the same approach defined previously (see Equation B.1). The α parameter was tuned using data from the previous track edition, and optimized for b-Pref. This run was submitted with reference **feupne**.

Mixed Approach

Our last run submitted for this task combined both the temporal span score and the temporal dispersion score, with the base BM25 feed scores. This run was derived by combining the two best runs including temporal span and temporal dispersion. These two runs were combined using Equation B.1. The label for this run is **feupfsne**.

Results

The FEUP team submitted four runs to the blog distillation task. Our goal was to evaluate the impact of temporal features in a specific information retrieval task. We submitted one baseline run (feupbase) that completely discards all temporal information. Then, we tested two time-dependent features in runs feupfs (temporal span) and feupne (temporal dispersion). Finally, we mixed these two features in run feupfsne.

Unfortunately, an implementation bug was found in one of our basic functions. This resulted in erroneous values in all submitted runs. After identifying the bug we reconstructed the runs and results using the official qrels. All results presented in Table B.3 were reconstructed offline.

Both proposed features improve over the temporally agnostic baseline in most metrics. The best results were achieved with the run combining the BM25 baseline and feed dispersion. The improvements for this run in b-Pref (+4.55%) and P@10 (+4.27%) are statistically significant at a level of 0.1. All statistically significant improvements are highlighted in bold ($p < 0.1$). Again, these results support our initial hypotheses about

Table B.3: Results for reconstructed runs for the Distillation Task.

Run	α	MAP	b-Pref	P@10
feupbase		0.1993	0.2436	0.3280
feupfs	0.90	0.1999	0.2530	**0.3400**
feupne	0.85	0.2007	**0.2547**	**0.3420**
feupfsne	0.10	0.2008	**0.2547**	**0.3420**

the value of temporal information for the distillation task. We briefly discuss these results in Section B.2.6.

B.2.5 Related Work

The temporal information contained in the TREC Blog06 collection has been explored previously in the task of SPAM detection. Lin et al. [57] have shown that SPAM blogs (splogs) have a very distinct temporal dynamics pattern, typically due to the use of automated publishing mechanisms (e.g. bulk submissions at a given time). Their approach has proven to be very successful in splog detection. Our work also investigates the temporal features of the same collection but to address different tasks.

In a related work, Ernsting et al. [34] used a language modeling approach in the tasks of blog post and feed finding. In this work, a time-based probability of the document being considered was defined. More recent documents were considered to be more relevant (i.e. better reflect the current interests of a blogger). Results showed that, using this time-dependent prior, only a slight statistically significant improvement in MAP was observed. The authors also note an improvement in P@30. Our work is distinct since we propose and test different temporal features. Also, we used a probabilistic model as a framework for experimentation.

B.2.6 Conclusions

The general research question that we are tackling can be summarized as follows: *"Is temporal information a valuable evidence for web information retrieval tasks?"*. Addressing this problem has always been problematic due to the lack of standard test collections containing temporal information [68]. The TREC Blog06 collection emerged as an exception in this landscape of static (snapshot-like) corpora.

Although an important portion of the blog posts in the collection do not contain temporal information (∼40%), we tried to make use of this evidence in two of the Blog Track tasks: adhoc retrieval and blog distillation. We tested three time-dependent features: temporal order, temporal span and temporal dispersion. In each task we combined these with a BM25-based baseline.

We used a standard rank aggregation approach to combine the features. The weights used in the aggregation equation were tuned using data from last year's Blog Track edition and optimizing for b-Pref [41]. It is important to note that the rank aggregation approach is based solely on the rank of each result, thus discarding all the information contained in the scores.

Overall, results were positive and support our initial hypothesis — temporal information can be used as a source of valuable features for standard information retrieval tasks. We found statistically significant improvements in standard IR measures using simple time-dependent features like temporal order of posts.

This was a first approach to the use of temporal features for traditional information retrieval tasks based on a real web collection. Results were positive and encourage further research. We identify two main directions for future research: the definition of further time-dependent features, and the development of better feature combination formulas. Since we only used rank order in the aggregation formula, it should be possible to improve these results given that scores contain more information [29].

B.3 TREC 2009 Blog Track

This section describes the participation of FEUP in the TREC 2009 Blog Track. FEUP participated in the faceted blog distillation task with work focused on the use of temporal features available in the new TREC Blogs08 collection. The approach presented in this section uses the temporal information available in most individual posts to amplify (or reduce) each post's score. Blog scores, and subsequent ranks, are obtained by combining individual posts' scores. While preparing the runs, no endeavors were made to identify a priori any temporal differences between the three distinct facets.

In the 2009 edition of the TREC Blog Track, two significant changes were introduced. First, several new tasks were initiated, most notably a faceted blog distillation task. Second, a new base collection was used. This new collection is significantly larger than

the previous Blogs06 collection and covers a much broader period of time. Given our interest in temporal properties, this is particularly relevant.

Our previous participation in the Blog Track [69] has shown that temporal information can have a positive impact in blog search. We continue this line of research by considering new approaches for incorporating time-sensitive features in the new faceted distillation task.

B.3.1 Blogs08 Collection Overview

The Blogs08 test collection was released in early April 2009 and is the official collection for the 2009 edition of the TREC Blog Track. For preparing this collection, a total of 1,303,520 feeds were polled once a week from January 14th, 2008 to February, 10th 2009 (394 days). The polled feeds, associated permalinks and homepage documents were stored, resulting in collection with a total compressed size of 453 GB.

Figure B.5 presents an overview of the total number of posts per day found in the collection. The date information is obtained directly from the `DATE_XML` field available in the collection. As reported in the Blogs08 specification, *"DATE_XML is the date of issue of the permalink, as stated in the RSS or Atom feed. As such tags are optional in the feeds, this information is not always present. Should you choose to use this information, you should make your own decision on how to supplement it when it is not present for a document."*

From the total number of permalink documents (posts) available in the collection (28,488,766), 97.5% had date information, while only 707,991 (2.5%) had an empty `DATE_XML` field. When considering only the posts with date information, 60.4% reported a date within the official crawling period. These posts are considered to have *valid dates*, while posts that report dates outside the crawling period are considered to have *invalid dates*. In Figure B.5, valid and invalid posts were identified by using different colors. Table B.4 summarizes the distribution of posts over a selected number of periods. There is a significant amount of documents published before the crawling period. This result is similar to that observed in the Blogs06 collection [58].

B.3.2 System Overview

The Terrier information retrieval platform [70] was used to index the permalink documents included in the Blogs08 collection, with the following TREC tags excluded:

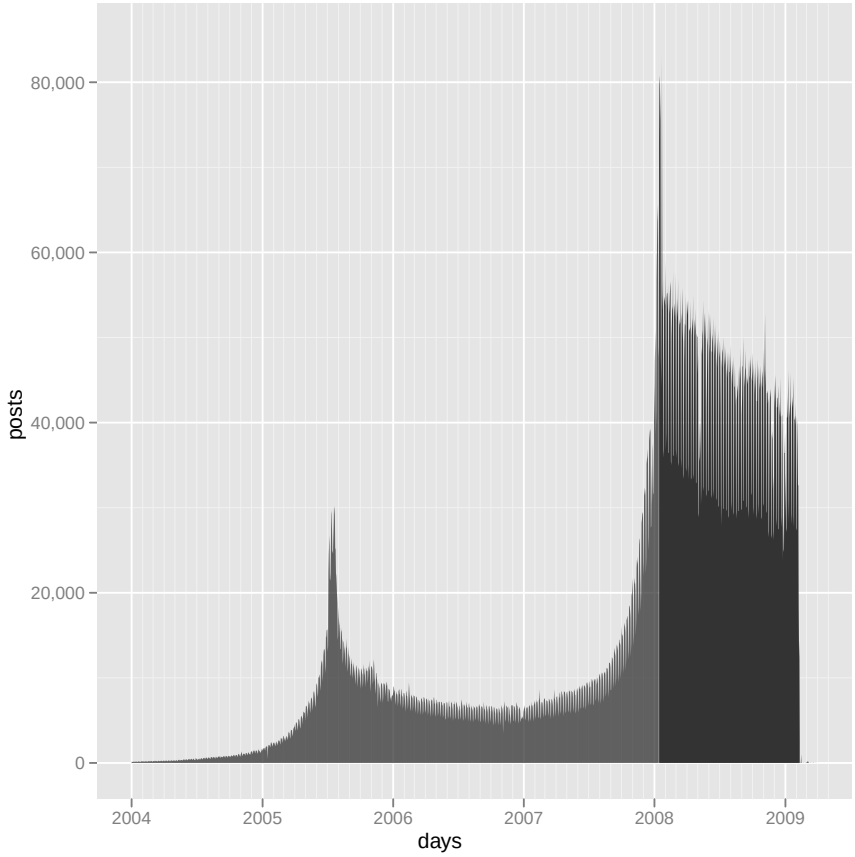

Figure B.5: Blogs08 collection overview.

Table B.4: Summary of distribution of post's dates.

Period	**Total Posts**
Crawling Period	16,787,445 (58.9%)
After Crawling Period	4,987 (0.02%)
Before Crawling Period	10,988,343 (38.6%)
in 2007	4,406,209 (15.5%)
in 2006	2,386,807 (8.4%)
in 2005	3,283,107 (11.5%)
in 2004	205,590 (0.7%)
Without Date	707,991 (2.5%)

DOCHDR, DATE_XML, FEEDNO, BLOGHPNO, BLOGHPURL, PERMALINK. Documents were retrieved using Terrier's implementation of the BM25 model [73], maintaining the default parameters: $k_1 = 1.2d$, $k_3 = 8d$ and $b = 0.75d$.

Document retrieval was done in two steps. First, a *phrase query* was used to retrieve documents, i.e. all terms needed to appear in the same phrase (e.g.: "term1 term2"). However, for some topics, this approach returned zero results. Thus, for these few topics, a more relax query was used — terms only needed to appear in the same document (e.g. "+term1 + term2"). It is worth noting that no effort was made to identify or remove SPAM from the collection.

B.3.3 Faceted Blog Distillation

The Faceted Blog Distillation is a new task introduced in the 2009 edition of the TREC Blog Track. This task is a refinement of the previous *blog distillation task*, where quality aspects of the retrieved blogs were not evaluated. These quality aspects were introduced by considering *facets* as proposed by Hearst et al. [42]. Three (binary) facets were considered in this first edition: *opinionated/factual*, *personal/official* and *in-depth/shallow*.

At FEUP we are focused on the study of the temporal properties available in blogs and their value for ranking tasks. We have prepared and submitted several runs that use temporal information to rank blogs. We have not tried to identify a priori how temporal properties could influence each facet property. Instead, we have adopted an exploratory attitude by submitting the same run for each topic and all facet options. In other words, each submitted run has the same ranking for the three facet options: *no facet applied*, *facet on for 1st value* and *facet on for 2nd value*. In a nutshell, for a given topic, the same rank of blogs was submitted for all *facet options*. This permits a detailed analysis of the impact of our approach in each topic and with each facet option.

Baseline

Given the BM25 post-based ranks (see Section B.3.2), we prepared a baseline blog-based run by simply adding each feed's posts' scores and then dividing by the total number of posts available in the collection for that same feed. This run was submitted with the reference **FEUPirlab1**.

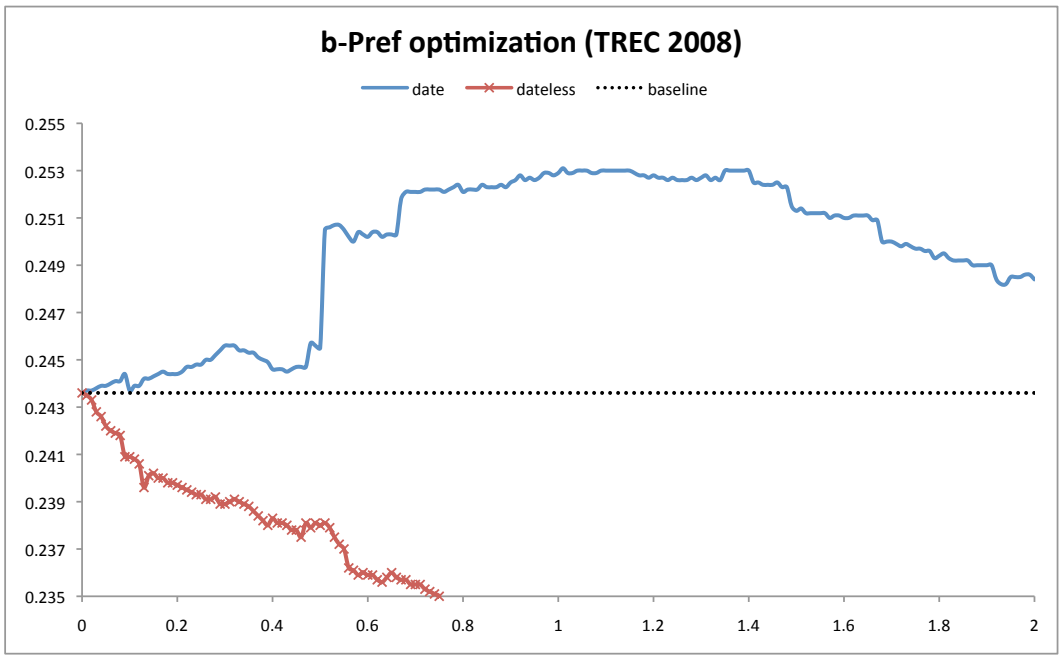

Figure B.6: b-Pref values using TREC 2008 data for posts with valid and invalid dates.

Due to hardware limitations, the original post-based runs used for preparing the blog-based runs were limited to 1,000 items. Permalink results with a rank higher than 1,000 were discarded. For topics that generate a lot of results from the same blogs, this might have a negative impact in the final blog distillation task rank.

Boost Invalid Dates

In a first approach, temporal information was introduced in the ranking formula by distinguishing between posts with valid dates (i.e. dates within the crawling period) and posts with invalid dates. The initial idea was to revise each post's original score based on its publish date being valid or not.

We implemented this approach using a simple formula: $score_{new} = score_{original} \times (1 + \alpha)$. The value of α was determined using data from the 2008 edition of the Blog Track, and conducting a linear search with $\alpha \in [0, 2]$, using 0.01 increments as shown in Figure B.6. Boosting posts with invalid dates results in consistent improvements in b-Pref. The highest boosting observed was of 3.8%. We prepared a run using $\alpha = 1$, i.e. posts with invalid dates had their scores doubled (a 100% boost) before computing the ag-

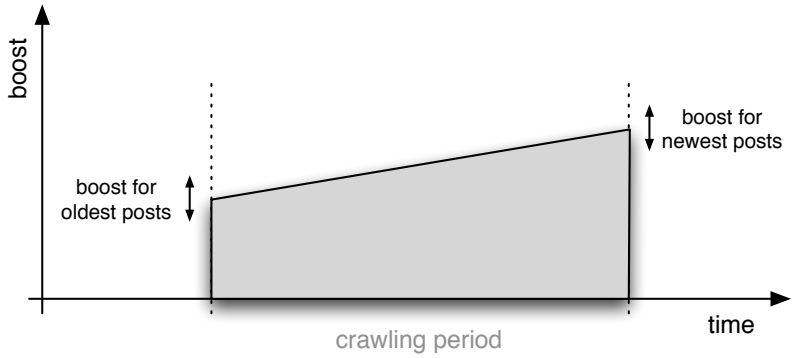

Figure B.7: Boosting valid dates.

gregated feed scores. This run was submitted with reference **FEUPirlab2**.

Boost Valid Dates

For posts with valid dates, two different scenarios were considered — boost newer posts and boost older posts. A simple linear scale was used as illustrated in Figure B.7. Given two boost values for the limits of the crawling period (start and end), the posts in between were augmented following a linear scale. For example, if the starting boost parameter is higher than the ending boost parameter, older post are valued more than newer posts. In the figure, newer posts' scores have a greater boost than older posts.

We did not perform an exhaustive search over all possible values to identify the optimum parameters due to time limitations. Instead, we conducted a manual exploratory search testing values between 0 and 2, with 0.01 increments. The best b-Pref improvement was achieved using a starting boost of 0.1 and an ending boost of 0.43. A run named **FEUPirlab3** was prepared with the previously defined boost applied to posts with invalid dates and a linear boost between 0.1 and 0.43 applied to posts with valid dates, i.e. newer posts have an higher boost. Additionally, a final run named **FEUPirlab4** was prepared with an inverse boost applied to posts with valid dates, from 0.43 to 0.1 (i.e. boost older posts). This last run was submitted to evaluate the impact of older posts in each facet. It is important to note that these runs were prepared using the previous **FEUPirlab2** run as a starting point.

Table B.5: Results of the faceted blog distillation task with facets off.

Run	MAP	b-Pref	R-prec	P@10
FEUPirlab1	0.1694	0.1911	0.2294	0.3179
FEUPirlab2	0.1752	0.1986	0.2447	0.3282
FEUPirlab3	0.1691	0.1950	0.2388	0.3179
FEUPirlab4	0.1662	0.1881	0.2248	0.3103

B.3.4 Results

Our team at FEUP submitted four runs to the faceted blog distillation task as described in the previous section. The first run is *temporally agnostic* (i.e. all temporal information was discarded), and is used as a baseline to observe the impact of temporal features. Table B.5 presents a summary of the official results for each run when facets are off and considering the official 39 topics. All statistically significant improvements are underlined ($p < 0.1$). Boosting posts with invalid dates resulted in an improvement of 3.42% in MAP and 3.24% in P@10. On the other hand, the refinements applied to posts with valid dates were inconclusive. Although some isolated improvements are observed, we cannot state that boosting newer posts or older posts produces better results.

A detailed analysis for each facet option is presented in Table B.6. Again, all statistically significant improvements are underlined ($p < 0.1$). The most consistent improvements are observed in the *opinionated* and *official* facet options. For the opinionated facet, boosting the posts without dates combined with boosting the newer posts, resulted in an improvement of 12.21% in MAP. A similar result was observed in the official facet, with an improvement of 6% in MAP. Boosting older posts had a positive (although sporadic) impact in the *in-depth* and *personal* facets. Overall, the worst results were found in the *shallow* facet, with very low MAP values. As pointed in the results tables, few improvements are statistically significant. This can be partially explained by the small number of cases for each facet option. For instance, in many facet options, we have fewer than 10 paired cases.

Table B.6: Results of the faceted blog distillation task for each facet option.

Run	MAP	R-prec	P@10
	in-depth (N=18)		
FEUPirlab1	0.1490	0.1441	0.2167
FEUPirlab2	0.1489	0.1625	0.2111
FEUPirlab3	0.1412	0.1523	0.1889
FEUPirlab4	0.1494	0.1385	0.2111
	opinionated (N=13)		
FEUPirlab1	0.0999	0.1360	0.1462
FEUPirlab2	0.1068	0.1458	0.1692
FEUPirlab3	0.1121	0.1466	0.1846
FEUPirlab4	0.0934	0.1360	0.1538
	personal (N=8)		
FEUPirlab1	0.1764	0.1975	0.1750
FEUPirlab2	0.1791	0.2464	0.2000
FEUPirlab3	0.1203	0.1749	0.1625
FEUPirlab4	0.1749	0.2168	0.1625
	shallow (N=18)		
FEUPirlab1	0.0506	0.0731	0.0667
FEUPirlab2	0.0491	0.0564	0.0611
FEUPirlab3	0.0497	0.0638	0.0722
FEUPirlab4	0.0500	0.0759	0.0611
	factual (N=13)		
FEUPirlab1	0.1369	0.1184	0.1308
FEUPirlab2	0.1339	0.1107	0.1308
FEUPirlab3	0.1370	0.1258	0.1231
FEUPirlab4	0.1347	0.1143	0.1308
	official (N=8)		
FEUPirlab1	0.1499	0.1078	0.1250
FEUPirlab2	0.1523	0.1126	0.1250
FEUPirlab3	0.1589	0.1126	0.1500
FEUPirlab4	0.1470	0.0989	0.1125

B.3.5 Related Work

In the context of the TREC Blog Track, previous work has shown that temporal information available in posts can have a positive effect in both document ranking [34, 69] and SPAM detection [57]. This work differs from our previous approach [69] in two distinctive aspects. First, it uses a new collection spanning a significantly larger time period (slightly more than a year). This aspect is particularly relevant for our work, given that it depends directly on features derived from dates. A broader collection has a higher number of potential *temporal bins* to discriminate results. Second, in our previous approach we combined a starting BM25-based rank with temporally biased ranks. Rank combination discards the finer-grained score values. In this work we have added a temporal bias to the original scores. Also, given the faceted nature of the task, we have made an initial investigation about the impact of time in these three facets.

B.3.6 Conclusions

The TREC Blogs08 collection is a new resource released in early 2009. In comparison with the Blogs06 collection, a larger number of blogs was crawled over a significantly larger time period (394 *vs* 77 days). Given our interest in temporal properties, this new collection is specially valuable. Our main goal was to conduct a first exploration of the Blogs08 collection focused on the temporal properties of blog posts. In contrast with our previous approach, we integrated the temporal features in the document scores before producing the final ranks. Our previous strategy was based on combining two ranks, discarding finer-grained score information.

From these experiments, we can draw some preliminary conclusions. Favoring posts with invalid dates produces consistent improvements. To a certain extent, this can be explained by the fact that the large majority of posts with invalid dates are prior to the crawling period. Thus, by enhancing posts with invalid dates we are giving a higher score to older posts, which tend to be associated with greater authority. In some facets, such as opinionated and official, the positive impact of this strategy was clear. However, from this data, it is not possible to draw a straightforward conclusion relatively to the impact of older and newer posts in the blog distillation task. It is worth highlighting that the calibration of the scoring parameters was done using data and assessments from a different collection, thus it is expected to have limitations.

It is our conviction that time is a complex dimension that cannot be treated in a linear,

one-sided fashion. For instance, on one hand, recent posts tend to be valued because they are more up-to-date and focused on the current subjects. On the other hand, old posts have an intrinsic value derived from their longevity and established nature.