

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA



UNIVERSIDADE
DE LISBOA

Information Search in Web Archives

Miguel Ângelo Leal da Costa

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE ENGENHARIA INFORMÁTICA

2014

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA



Information Search in Web Archives

Miguel Ângelo Leal da Costa

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE ENGENHARIA INFORMÁTICA

Tese orientada pelo Prof. Doutor Mário Jorge Costa Gaspar da Silva e pelo
Prof. Doutor Francisco José Moreira Couto

2014

Resumo

Os arquivos da web preservam informação que foi publicada na web ou digitalizada de publicações impressas. Muita dessa informação é única e historicamente valiosa. Contudo, os utilizadores não dispõem de ferramentas dedicadas para encontrar a informação desejada, o que limita a utilidade dos arquivos da web.

Esta dissertação investiga soluções para o avanço da recuperação de informação em arquivos da web (WAIR) e contribui para o aumento de conhecimento acerca da sua tecnologia e dos seus utilizadores. A tese subjacente a este trabalho é a de que os resultados de pesquisa podem ser melhorados através da exploração de informação temporal intrínseca aos arquivos da web. Esta informação temporal foi explorada de dois ângulos diferentes. Primeiro, a longa persistência dos documentos web foi analisada e modelada para melhor estimar a relevância destes em função da pesquisa. Segundo, foi concebido um enquadramento (framework) para ordenação de resultados dependente do tempo, que aprende e combina modelos específicos para cada período. Esta abordagem contrasta com a abordagem de um modelo único que ignora a variação das características da web ao longo do tempo.

A abordagem proposta foi validada empiricamente através de várias experiências controladas que demonstraram a sua superioridade em relação ao estado da arte em WAIR.

Palavras-Chave: Arquivamento da Web, Pesquisa de Informação, Aprendizagem Automática

Abstract

Web archives preserve information that was published on the web or digitized from printed publications. Many of that information is unique and historically valuable. However, users do not have dedicated tools to find the desired information, which hampers the usefulness of web archives.

This dissertation investigates solutions towards the advance of web archive information retrieval (WAIR) and contributes to the increase of knowledge about its technology and users. The thesis underlying this work is that the search results can be improved by exploiting temporal information intrinsic to web archives. This temporal information was leveraged from two different angles. First, the long-term persistence of web documents was analyzed and modeled to better estimate their relevance to a query. Second, a temporal-dependent ranking framework that learns and combines ranking models specific for each period was devised. This approach contrasts with a typical single-model approach that ignores the variance of web characteristics over time.

The proposed approach was empirically validated through various controlled experiments that demonstrated their superiority over the state-of-the-art in WAIR.

Keywords: Web Archiving, Information Search, Machine Learning

Resumo Estendido

A World Wide Web contém todo o tipo de informação, sendo muita dessa informação única e historicamente valiosa. Por exemplo, o discurso de um presidente depois de ganhar as eleições ou o anúncio de uma invasão iminente num país estrangeiro, podem-se tornar tão valiosos no futuro como os manuscritos antigos são hoje valiosos para compreender o passado. Contudo, o facto de a web ser constantemente atualizada, com milhões de documentos adicionados, modificados e apagados diariamente, faz com que a sua informação seja efémera. Estudos indicam que 80% das páginas web ficam indisponíveis ao fim de um ano. Ou seja, a grande maioria da informação que a humanidade está a criar hoje vai desaparecer dentro de poucos anos, originando uma lacuna de conhecimento para as gerações vindouras.

Para minorar o impacto deste problema, os arquivos da web preservam parte da informação publicada na web ou que foi digitalizada de publicações impressas. Identificaram-se arquivos distribuídos por 33 países em 5 continentes e, juntos, armazenam mais de 534 mil milhões de ficheiros (17 PB). Este número continua a crescer rapidamente à medida que novas iniciativas continuam a surgir. Contudo, para tornar estes dados acessíveis, os arquivos da web têm de evoluir de meros repositórios de documentos para arquivos de fácil acesso. Atualmente existe um grande desconhecimento sobre os utilizadores de arquivos da web, o que inevitavelmente leva a pressupostos errados quando se está a desenhar e otimizar tecnologia para eles. Para além disso, os arquivos da web são tendencialmente construídos usando tecnologia de motores de busca da web, ignorando a dimensão temporal dos dados e as necessidades de informação dos utilizadores. Em consequência, os utilizadores não conseguem encontrar a informação desejada, tornando os arquivos da web inúteis.

Esta dissertação investiga soluções para o avanço da recuperação de informação em arquivos da web (WAIR) e apresenta algumas contribuições visando o aumento de conhecimento acerca da sua tecnologia e dos seus utilizadores. Foram efetuados dois estudos sobre iniciativas de arquivos da web: (1) inquéritos; (2) recolha de dados de documentação técnica. Ambos os estudos foram seguidos de uma análise quantitativa e qualitativa dos dados, permitindo identificar os pontos fortes e fracos do estado da arte, as tendências e os problemas associados, e os desenvolvimentos necessários para satisfazer as necessidades de informação dos utilizadores. A compreensão destas necessidades, assim como o tipo de informação pesquisada e os padrões de pesquisa, foram obtidos através de três estudos sobre os utilizadores: (1) questionários online; (2) prospeção nos registos de pesquisa; (3) estudos em laboratório. O conhecimento obtido é fundamental para desenvolver tecnologia de pesquisa orientada para a satisfação dos utilizadores e apoiar decisões arquiteturais de um arquivo da web eficaz e eficiente. Por outro lado, o conhecimento obtido expôs falhas graves na tecnologia atual. Por exemplo, a tecnologia que suporta os utilizadores de arquivos da web foi desenvolvida para os utilizadores de motores de busca da web, que têm necessidades de informação diferentes.

Os estudos efetuados nesta dissertação mostram que a pesquisa textual é o método preferido dos utilizadores para achar informação em arquivos da web. Este tipo de pesquisa é semelhante à pesquisa típica de um motor de busca, em que o utilizador submete um conjunto de termos representativos da sua necessidade de informação e recebe uma lista de documentos ordenada por relevância para essa necessidade. Contudo, esta pesquisa textual é processada sobre a web de um período definido pelo utilizador, permitindo estudar o passado e suportar funcionalidades analíticas ao longo do tempo. Os arquivistas da web referem que este serviço de pesquisa é difícil de implementar e a eficácia dos serviços existentes não é satisfatória para os utilizadores. Com o rápido crescimento dos dados arquivados, este problema tende a agravar-se. Neste trabalho foi confirmada, pela primeira vez, a fraca

eficácia do estado da arte em sistemas WAIR, medida através de uma nova metodologia de avaliação. Esta metodologia foi proposta considerando as especificidades dos sistemas WAIR e seus utilizadores, ambos caracterizados nos vários estudos acima descritos.

A tese subjacente a este trabalho é a de que os resultados de pesquisa obtidos pelos sistemas de WAIR atuais podem ser melhorados através da exploração de informação temporal intrínseca aos arquivos da web. Esta informação temporal foi explorada de dois ângulos diferentes.

Primeiro, foram desenvolvidos modelos para ordenação de resultados em arquivos da web, baseados no pressuposto de que os documentos mais relevantes são mantidos acessíveis durante mais tempo na web. Por exemplo, se muitas pessoas lerem um jornal online com frequência, o autor desse jornal vai provavelmente garantir que a informação se mantém acessível e em alguns casos atualizada. A persistência dos documentos web foi analisada durante um intervalo de tempo de 14 anos e medida através do número de versões arquivadas e do tempo de vida dos documentos (diferença temporal entre a primeira e última versão arquivada). A modelação destas métricas de persistência permitiu estimar melhor a relevância dos documentos que satisfazem pesquisas navegacionais (pesquisas com o intuito de encontrar documentos específicos). Esta modelação é especialmente importante para arquivos da web, porque os modelos típicos para estimar a importância ou popularidade de documentos, baseiam-se em cliques nos resultados de pesquisa e hiperligações entre documentos que não estão disponíveis em quantidade suficiente neste contexto. Os arquivos da web recebem muito menos pesquisas e cliques que os motores de busca da web, e os grafos da web são muito mais esparsos, porque apenas uma pequena parte da web é usualmente arquivada.

Segundo, foi concebido um enquadramento (framework) para ordenação de resultados dependente do tempo. As características da web variam ao longo do tempo. Por exemplo, as páginas da década de

1990 compostas maioritariamente por texto e HTML eram mais simples do que as páginas da década de 2000, compostas por imensas tecnologias embutidas nas páginas, tais como JavaScript e CSS. As hiperligações entre documentos crescem segundo uma lei de potência. A linguagem evolui, com muitos termos que aparecem e desaparecem todos os anos. Por isso, este enquadramento aprende e combina múltiplos modelos, cada um específico de um período. A ideia subjacente é que um modelo treinado com dados de um período é provavelmente mais eficaz a ordenar resultados de pesquisa desse período do que de períodos diferentes. Para além disso, os dados de períodos mais próximos são provavelmente mais parecidos entre si do que aqueles de períodos mais afastados. Logo, a aprendizagem de um período deve ser maior quanto menor a distância temporal entre os dados. Esta abordagem que treina múltiplos modelos, contrasta com a abordagem de um modelo único que ignora a variação das características da web ao longo do tempo e ordena os documentos independentemente da sua data de criação e atualização.

As abordagens propostas foram validadas empiricamente através de várias experiências controladas. Foi usada uma coleção de testes representativa, que contém um corpus que abrange 14 anos de coleções web arquivadas. Os resultados das experiências demonstraram a significativa superioridade das abordagens, individualmente e em conjunto, em relação ao estado da arte em WAIR e validaram a hipótese apresentada. Por sua vez, a implementação das abordagens propostas num arquivo da web de larga escala, demonstrou a sua viabilidade e utilidade num sistema real. Os conjuntos de dados usados nas experiências e todo o código estão disponíveis em formato de acesso livre.

Acknowledgements

This thesis would not be possible without the help and encouragement of several people. I would first and foremost like to thank my parents, Dália and Armando Costa, who always taught me to chase my dreams and never give up. All I have and will accomplish are only possible due to their love and sacrifices.

I am deeply grateful to my advisor, Mário Silva, for his guidance, encouragement and continuous support throughout the course of this work. He has introduced me to the amazing research area of information retrieval in 2001 and has been teaching me a lot about it since then. I am also grateful to my co-advisor, Francisco Couto, for his guidance and assistance. Their knowledge has been a source of inspiration to me.

I have been fortunate in working as a member of the Portuguese Web Archive team. I thank to my present and past colleagues for their support and enlightening discussions, namely Daniel Gomes, João Miranda, Simão Fontes, David Cruz and André Nogueira. A special acknowledgement to my colleagues of the Portuguese Foundation for National Scientific Computing (FCCN) for their friendship and valuable feedback, especially to Fernando Ribeiro, Vitor Chixaro, Nelson Schaller and Hugo Mendes.

I would also have to thank my friends who are always reminding me that life is more than work and much more interesting outside a computer. Thank you João Pinto, Vanessa Martins, Rui Grilo, Ana Candeias, and all the others, for making my time worthwhile.

Finally, I would like to express my gratitude to Joana Pena for all her love, support and understanding. You gave me the strength to finish this journey. Thank you for everything my love.

This thesis is dedicated to my parents.

Contents

1	Introduction	1
1.1	Objectives	2
1.2	Research Methodology	4
1.3	Contributions	5
1.4	Publications	8
1.5	Overview	10
2	Background & State-of-the-Art	13
2.1	Web Archiving Workflow	14
2.1.1	Web Archive Architecture	17
2.2	Web Archiving Initiatives	19
2.2.1	Portuguese Web Archive	21
2.2.2	Access Types & Tools	22
2.3	User Studies	25
2.3.1	Web Archive Users	25
2.3.2	Information Needs	26
2.3.3	Search Patterns	27
2.4	Ranking	28
2.4.1	Conventional Ranking Models	28
2.4.2	Temporal Ranking Models	30
2.4.3	Learning to Rank	32
2.4.4	Query-Type-Dependent Learning to Rank	35
2.4.5	Datasets for Learning to Rank	36
2.5	IR Evaluations	37
2.5.1	Pooling	38

CONTENTS

2.5.2	Implicit Feedback	39
2.5.3	Crowdsourcing	40
2.5.4	Evaluation Measures	41
2.6	Summary	43
3	Characterizing Web Archives	45
3.1	Methodology	46
3.1.1	Comparison with other Surveys	48
3.2	Results	48
3.2.1	Initiatives	48
3.2.2	Archived Data	54
3.2.3	Access and Technologies	57
3.3	Summary	61
4	Characterizing Information Needs	63
4.1	The PWA User Interface	65
4.2	Methodology	67
4.2.1	Data Collecting Methods	67
4.2.2	Experiment #1: Search Logs	69
4.2.3	Experiment #2: Interactive Questionnaire	71
4.2.4	Experiment #3: Laboratory Study	73
4.3	Results	74
4.3.1	Experiment #1: Search Logs	75
4.3.2	Experiment #2: Interactive Questionnaire	77
4.3.3	Experiment #3: Laboratory Study	79
4.4	Summary	79
5	Characterizing Search Patterns	83
5.1	Logs Dataset	84
5.2	Methodology	84
5.2.1	Log Preparation	85
5.3	Results	86
5.3.1	Session Level Analysis	87
5.3.2	Query Level Analysis	88

CONTENTS

5.3.3	Term Level Analysis	94
5.3.4	Temporal Level Analysis	95
5.4	Summary	98
6	Evaluating WAIR systems	101
6.1	Web Archive Characteristics	103
6.1.1	Corpus	103
6.1.2	Search Topics	104
6.1.3	Relevance Propagation	106
6.2	Evaluation Methodology	107
6.2.1	Evaluation Measures	109
6.3	Test Collection Construction	110
6.3.1	Corpus Selection	110
6.3.2	Search Topics Selection	111
6.3.3	Retrieval	113
6.3.4	Relevance Assessment	114
6.3.5	General Statistics	116
6.4	Results	117
6.4.1	Topic difficulty	118
6.4.2	Reusability	119
6.5	Summary	119
7	Improving WAIR systems	123
7.1	Web Documents Persistence	125
7.1.1	Collection Description	125
7.1.2	Document Persistence	125
7.1.3	Document Persistence & Relevance	127
7.1.4	Modeling Document Persistence	128
7.2	Temporal-Dependent Ranking	129
7.2.1	Ranking Problem	129
7.2.2	Temporal Intervals	130
7.2.3	Temporal-Dependent Models	130
7.2.4	Multi-task Learning	132
7.2.5	L2R Algorithm	133

CONTENTS

7.3	Experimental Setup	134
7.3.1	L2R Dataset	135
7.3.2	Ranking Features	137
7.3.3	Ranking Algorithms	138
7.3.4	Ranking Models Compared	139
7.3.5	Evaluation Methodology and Metrics	140
7.4	Results	140
7.4.1	Results Analysis	145
7.5	Summary	146
8	Conclusions	149
8.1	Caveats	151
8.2	Outlook	153
8.3	Resources	155
A	List of Web Archives Surveyed	157
B	Ranking Features	161
	References	165

List of Figures

2.1	Web archiving workflow.	14
2.2	Web archive architecture.	18
2.3	Document archived on October 13, 1996: <i>homepage of Portugal</i> . .	23
2.4	A general paradigm of L2R.	34
3.1	Countries hosting web archiving initiatives.	52
3.2	Cumulative number of initiatives created per year.	53
3.3	Size of archived collections.	55
3.4	Usage of file formats to store web contents.	57
3.5	Access type provided by web archives.	58
3.6	Technologies used by web archives.	60
4.1	Search interface after a full-text search.	65
4.2	Advanced search interface.	66
4.3	Search interface after a URL search.	67
4.4	Data collecting methods used.	69
4.5	Survey about the search of the Portuguese Web Archive.	71
4.6	Tag clouds of search queries.	78
5.1	Distribution of ranks clicked on SERPs.	93
5.2	Cumulative distributions of queries.	94
5.3	Cumulative distribution of full-text query terms.	95
5.4	Years included in queries restricted by date.	96
5.5	Clicks on years with archived versions (from oldest to newest). .	97
6.1	Methodology for building a WAIR test collection.	108

LIST OF FIGURES

6.2	Form used to assess navigational topics.	115
6.3	Navigational topics sorted by the average of the 9 tested ranking models.	119
7.1	Distribution of the lifespan of documents in years.	126
7.2	Distribution of the number of versions of documents over 14 years.	126
7.3	Fraction of documents with a lifespan longer than 1 year in each relevance level.	127
7.4	Fraction of documents with more than 10 versions in each relevance level.	128
7.5	Weights of training instances when learning ranking models.	131
7.6	NDCG results of the temporal-dependent ranking framework using regular features.	143
7.7	NDCG results of the temporal-dependent ranking framework using regular and temporal features.	144

List of Tables

2.1	Contingency table of the variables that form IR evaluation measures.	41
3.1	General statistics of web archiving initiatives.	49
4.1	Distribution of information needs in the three experiments.	74
4.2	Distribution of sessions searched between dates per information need.	75
4.3	Topics searched per navigational needs.	76
4.4	Topics searched per informational needs.	76
4.5	Distribution of information needs on several studies.	80
5.1	General statistics of user interactions.	86
5.2	Session duration (minutes).	87
5.3	Number of queries per session.	88
5.4	General statistics of modified queries and terms.	89
5.5	Number of terms changed per modified full-text query.	90
5.6	Advanced operators per full-text query.	91
5.7	Number of terms per query.	92
5.8	SERPs viewed per query.	92
5.9	Queries restricted by date.	96
5.10	Comparison between users of web search engines and web archives.	98
6.1	Web crawls that compose the corpus of the test collection.	111
6.2	Relevance judgments in the WAIR test collection per relevance grade.	116
6.3	Test collection statistics.	117
6.4	Results of the tested ranking models.	118

LIST OF TABLES

7.1	Relevance judgments in the L2R dataset per relevance grade.	135
7.2	Data partitioning for 5-fold cross validation.	137
7.3	Results of the tested ranking models.	141
7.4	Top 6 most important ranking features for the temporal-dependent ranking framework.	146
A.1	List of web archives surveyed.	158
A.2	Characteristics of web archives surveyed.	159
B.1	List of ranking features of the L2R dataset for WAIR research. . .	163

Chapter 1

Introduction

The World Wide Web has a democratic nature, where everyone can publish all kinds of information using different types of media. News, blogs, wikis, encyclopedias, photos, interviews and public opinions are just a few examples of this enormous list. Part of this information is unique and historically valuable. For instance, the speech of a president after winning an election or the announcement of an imminent invasion of a foreign country, might become as valuable as ancient manuscripts are today. However, since the web is too dynamic, a large amount of information is lost everyday. 80% of web pages are not available after one year ([Ntoulas et al., 2004](#)). 13% of web references in scholarly articles disappear after 27 months ([Dellavalle et al., 2003](#)). 11% of social media resources, such as the ones posted in Twitter, are lost after one year ([SalahEldeen & Nelson, 2012](#)). All this information will likely vanish in a few years, creating a knowledge gap for future generations. The UNESCO recognized the importance of digital preservation in 2003, by stating that the disappearance of digital information constitutes an impoverishment of the heritage of all nations ([UNESCO, 2003](#)). It is therefore important to preserve these data, not only for historical and social research ([Ackland, 2005](#); [Arms et al., 2006a,b](#); [Foot & Schneider, 2006](#); [Franklin, 2004](#); [Kitsuregawa et al., 2008](#)), but also to support current technology, such as assessing the trustworthiness of statements ([Yamamoto et al., 2007](#)), detecting web spam ([Chung et al., 2009](#)), improving web information retrieval ([Elsas & Dumais, 2010](#)) or forecasting events ([Radinsky & Horvitz, 2013](#)).

1. INTRODUCTION

At least 68 web archiving initiatives¹ undertaken by national libraries, national archives and consortia of organizations are acquiring and preserving parts of the web. Together, they hold more than 534 billion files (17 PB) and this number continues to grow as new initiatives continue to arise. Some country code top-level domains and thematic collections are being archived regularly², while other collections related to important events, such as September 11th, are created at particular points in time³. Web archives also contribute to preserve contents born in non-digital formats that were afterwards digitized and published online, such as The Times Archive⁴ with news since 1785. As result, web archives contain often millions or billions of archived documents and cover decades or even centuries in the case of digitized publications. The historic interest over these documents is also growing as they age, becoming a unique source of past information for widely diverse areas, such as sociology, history, anthropology, politics, journalism, linguistics or marketing. However, for making historical analysis possible, web archives must turn from mere document repositories into accessible archives.

Much attention has been given to preserving the past content of the web, but little in finding efficient and effective ways to search and explore the archived data. Web archives are built on top of web search engine technology and are accessed through indexing a series of web snapshots accumulated over the years as a single collection. This ignores the temporal dimension of the collected data and inevitably creates unsatisfied users, even more because the technology is not designed and optimized for their information needs. The huge volume and fast growing of web archive data only increases the challenge of finding information. In a nutshell, web archives provide poor access services to their users and without access, web archives are useless.

1.1 Objectives

This dissertation investigates solutions towards the advance of web archive information retrieval (WAIR). It intends to overcome the challenges that hamper users

¹ http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

² e.g. Internet Archive available at <http://www.archive.org>

³ e.g. Library of Congress Web Archives available at <http://www.loc.gov/minerva>

⁴ <http://www.thetimes.co.uk/tto/archive/>

1.1 Objectives

from using web archives, namely the lack of knowledge about their technology and users, and the poor search effectiveness (i.e. quality of search results) of web archives. These challenges are evidenced in the research literature. For instance, a survey on web archiving initiatives in the USA conducted by the [NDSA Content Working Group \(2012\)](#) stated that "the lack of knowledge about web archive usage and users is clearly a topic that merits further investigation". [Dougherty et al. \(2010\)](#) wrote that "to date, there is still no reliable full text search tool for web archives and, although several groups are currently working on the problem, it remains one of the greatest obstacles to providing archives usable for a wide variety of researchers". A survey on European web archives conducted by the [Internet Memory Foundation \(2010\)](#) reported that 82% of these archives considered "the improvement of access tools a high priority".

To address the above challenges, we need to answer three essential research questions:

Q1: Does the state-of-the-art in WAIR meet the users' information needs?

Q2: Why, what and how do web archive users search?

Q3: How to improve WAIR?

The improvement of IR technology, regarding its effectiveness, is typically achieved by creating novel ranking features and models to better estimate the relevance of documents to a query. Both depend on the characteristics of data, which in web archives are primarily many years of collected web snapshots. Previous research, such as the analysis of the evolution of the web ([Miranda & Gomes, 2009a](#)) and the language of its content ([Tahmasebi et al., 2012](#)), showed that many information can be extracted from these data. In recent works, temporal information has been leveraged to improve the search effectiveness of IR systems ([Elsas & Dumais, 2010](#)). This leads me to posit that the time dimension present in the data of web archives likely conceals temporal information that can be exploited to extract more discriminative ranking features and design more effective ranking models. Currently, WAIR systems do not take into account the time dimension of archived data. For instance, the variance of web characteristics over long periods of time is completely ignored and hence documents that were created

1. INTRODUCTION

many years apart are searched exactly the same way. Therefore, I propose the following

hypothesis: the search results achieved by state-of-the-art WAIR systems can be improved by exploiting temporal information intrinsic to web archives.

1.2 Research Methodology

Besides the analysis of existing approaches, the validation of the hypothesis of this dissertation entails:

1. Surveying the status of current web archiving technology to understand its trends, strengths and limitations. There is a lack of knowledge in the research community about the state-of-the-art in web archiving that this dissertation tries to fulfill. This knowledge is essential to identify the developments that are still missing and which ones need improvement towards the satisfaction of the user information needs.

[Related to Q1]

2. Studying via data collecting methods, such as online questionnaires, search log mining and laboratory studies, the information needs, expectations and search patterns of web archive users. A clear understanding of users is fundamental for the development of useful search functionalities and the architectural design decisions for a state-of-the-art web archive. This knowledge also gives new insights in web archiving.

[Related to Q1 and Q2]

3. Developing novel information retrieval (IR) and machine learning (ML) approaches to support time-travel queries, i.e. full-text search on the state of the web within a user-specified time interval. This is considered a *killer application* for web archives, making historical analysis possible and supporting analytical functionalities over time (Weikum *et al.*, 2011). The temporal characteristics of successive web snapshots are exploited to create discriminative ranking features and learned by temporal-dependent ranking

models that take into account the variance of web characteristics over time.

[Related to Q3]

An evaluation methodology and a test collection, addressing the specificities of real WAIR systems, were created to evaluate the proposed approaches and support the validation of the thesis statement through various controlled experiments. Experimental results showed a significant gain in search effectiveness, when compared against the state-of-the-art in WAIR and even against stronger baselines using state-of-the-art learning to rank (L2R) algorithms, which validates my thesis.

The research presented in this dissertation was made in the context of the Portuguese Web Archive (PWA) project, which resulted in the development of the PWA system that integrates the developed techniques. The PWA system enables users to access past information published on the web and ensure its long-term preservation. The system is available at <http://archive.pt>.

1.3 Contributions

This dissertation concerns providing better WAIR functionalities for users and makes several contributions towards that goal. The list of contributions is presented below with references to the corresponding research questions and the chapters where the contributions are discussed:

An updated and the most comprehensive characterization of the state-of-the-art in web archiving, addressing the volume of archived data, used formats, number of people engaged, access type and the employed technology ([Gomes et al., 2011](#)). A Wikipedia page with information about web archiving initiatives was created to complement the presented work and has been collaboratively kept up-to-date by the community.

[Related to Q1 and addressed in Chapter 3]

A deeper knowledge of web archive users about why, what and how do they search. The answers obtained for the first time are essential to point out directions for developing technology that can better satisfy the users ([Costa](#)

1. INTRODUCTION

& Silva, 2010a,b, 2011). I resort to three instruments to collect quantitative and qualitative data, namely search log mining, an online questionnaire and a laboratory study.

[Related to Q1 and Q2, and addressed in Chapters 4 and 5]

A proposal of an evaluation methodology for WAIR systems based on a list of requirements compiled from previous characterizations of web archives and their users (Costa & Silva, 2009, 2012). The methodology, along with a test collection created to support it, enabled for the first time to measure the effectiveness of state-of-the-art WAIR technology. The test collection was made available to the research community.

[Related to Q1 and Q3, and addressed in Chapter 6]

The engineering of novel ranking features optimized for web archives, using the test collection as a fundamental piece in this process (Costa & Silva, 2012; Costa *et al.*, 2014). The features exploit temporal information intrinsic to web archives, along with the regular topical information used in web search engines. Results confirm that these features are good at discriminating relevant from not-relevant documents for the user queries.

[Related to Q3 and addressed in Chapters 6 and 7]

The demonstration of the usefulness of the learning to rank (L2R) framework in WAIR. I applied, for the first time, the state-of-the-art L2R framework and L2R algorithms to improve the search effectiveness of web archives (Costa *et al.*, 2014; Gomes *et al.*, 2013). A specific dataset for this task was developed and made available to the research community to support research on L2R in WAIR.

[Related to Q3 and addressed in Chapter 7]

A proposal of a temporal-dependent ranking framework that addresses the fact that the characteristics of web documents vary over time influencing ranking models (Costa *et al.*, 2014). By simultaneously learning ranking models from disjoint temporal intervals of web snapshots, I outperformed the search effectiveness of web archives over single-model approaches that fit all data

independently of when documents are created or updated.

[Related to Q3 and addressed in Chapter 7]

The empirical validation of the novel ranking features and the proposed framework, which in turn validates the thesis. I conducted experiments on a large-scale real-world web archive corpus that covers a timespan of 14 years. I demonstrated the superiority of the proposed features and methods over the existing state-of-the-art by achieving up to four times better results. This has a large impact on user satisfaction.

[Related to Q1, Q2 and Q3, and addressed in Chapter 7]

Web search engines face many challenges related to scalability and information overload (Baeza-Yates *et al.*, 2007b). Web archives face a greater challenge, because they accumulate previous documents and indexes, unlike web search engines that tend to drop the old versions when new ones are discovered. Even so, web archives have a much smaller budget, which leads them to find solutions that provide satisfactory results in *Google time* with much less resources. Despite not being the main research topic of this thesis, I also contribute with the lessons learned while researching and developing an efficient and effective WAIR system for the PWA (Gomes *et al.*, 2008, 2013), which includes the design of a distributed and scalable WAIR architecture according to the temporal dimension where indexes are partitioned by time (Costa *et al.*, 2013a) and used for query suggestion (Costa *et al.*, 2013b). The PWA is now the largest full-text searchable web archive publicly available and I believe that sharing my experience obtained while developing and operating a running service will enable other organizations to start or improve their web archives. Moreover, the integration of this research in the PWA contributes directly to real users having a better experience in finding and exploring past information.

The PWA serves other purposes beyond the preservation of historical and cultural aspects, such as the characterization of the Portuguese web (Miranda & Gomes, 2009a) and the aggregation of special contents for research communities (Garzó *et al.*, 2013; Lopes *et al.*, 2010). Another important aspect is the contribution to the dissemination of the Portuguese language on the web, which is used by 254 million people and considered the fifth most popular language on the

1. INTRODUCTION

Internet⁵. The PWA also provides access to web contents of interest to scientists working in different fields, such as History, Sociology or Linguistics ([Gomes & Costa, 2014](#)). Finally, it reduces national dependence on foreign services regarding web data processing and searching, and supplies evidence in court cases that require information published on the web that is no longer available online.

Despite this work being focused in web archives, results may have interest to other research domains, such as web IR and digital libraries. For instance, the temporal features extracted from web archives or the temporal-dependent ranking framework can be used to improve the results of web search engines or other IR systems containing versioned documents.

The developed software is publicly available under the LGPL license and can be accessed at <http://pwa-technologies.googlecode.com>. The datasets for research are available at the same URL.

1.4 Publications

The research presented in this dissertation was originally published in several peer-reviewed international conferences and workshops. Next, a list of publications and the chapters where they are included are presented.

The following publications are about characterizations of the state-of-the-art in WAIR and web archive users:

1. Daniel Gomes, João Miranda and Miguel Costa, *A Survey on Web Archiving Initiatives*. In the 1st International Conference on Theory and Practice of Digital Libraries, Berlin, Germany. September 2011.
[This publication is included in Chapter 3]

2. Miguel Costa and Mário J. Silva, *Understanding the Information Needs of Web Archive Users*. In the IPRES2010 10th International Web Archiving Workshop, Vienna, Austria. September 2010.
[This publication is included in Chapter 4]

⁵ <http://www.internetworldstats.com/stats7.htm>

1.4 Publications

3. Miguel Costa and Mário J. Silva, *Characterizing Search Behavior in Web Archives*. In the WWW2011 1st Temporal Web Analytics Workshop, Hyderabad, India. March 2011.

[This publication is included in Chapter 5]

4. Miguel Costa and Mário J. Silva, *A Search Log Analysis of a Portuguese Web Search Engine*. In the INForum - Simpósio de Informática, Braga, Portugal. September, 2010.

[This publication is included in Chapter 5]

The next publications are about improving the state-of-the-art in WAIR towards the user information needs:

5. Miguel Costa and Mário J. Silva, *Evaluating Web Archive Search Systems*. In the 13th International Conference on Web Information System Engineering, Paphos, Cyprus. November 2012.

[This publication is included in Chapter 6]

6. Miguel Costa and Mário J. Silva, *Towards Information Retrieval Evaluation over Web Archives* (poster). In the SIGIR 2009 Workshop on the Future of IR Evaluation, Boston, U.S. July 2009.

[This publication is included in Chapter 6]

7. Miguel Costa and Francisco M. Couto and Mário J. Silva, *Learning Temporal-Dependent Ranking Models*. Accepted for publication in the 37th Annual ACM SIGIR Conference, Gold Coast, Australia. July 2014.

[This publication is included in Chapter 7]

8. Daniel Gomes, Miguel Costa, David Cruz, João Miranda and Simão Fontes, *Creating a Billion-Scale Searchable Web Archive*. In the WWW2013 3rd Temporal Web Analytics Workshop, Rio de Janeiro, Brazil. May 2013.

[This publication is included in Chapter 7]

Other works were developed during the research of this thesis, which resulted in several other publications:

1. INTRODUCTION

9. Daniel Gomes and Miguel Costa, *The Importance of Web Archives for Humanities*. In the International Journal of Humanities and Arts Computing. April 2014.
10. Miguel Costa, João Miranda, David Cruz and Daniel Gomes, *Query Suggestion for Web Archive Search*. In the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal. September 2013.
11. Daniel Gomes, David Cruz, João Miranda, Miguel Costa and Simão Fontes, *Acquiring and providing access to historical web collections* (demo). In the Demos Track of the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal. September 2013.
12. Miguel Costa, Daniel Gomes, Francisco M. Couto and Mário J. Silva, *A Survey of Web Archive Search Architectures*. In the WWW2013 3rd Temporal Web Analytics Workshop, Rio de Janeiro, Brazil. May 2013.
13. Daniel Gomes, David Cruz, João Miranda, Miguel Costa and Simão Fontes, *Search the Past with the Portuguese Web Archive* (demo). In the Demos Track of the 22nd International World Wide Web Conference, Rio de Janeiro, Brazil. May 2013.
14. Daniel Gomes, André Nogueira, João Miranda, Miguel Costa, *Introducing the Portuguese web archive initiative*. In the ECDL2008 8th International Web Archiving Workshop, Aarhus, Denmark. September 2008.

1.5 Overview

The remaining of this thesis is organized as follows. Chapter 2 provides background to this work and the necessary overview of the state-of-the-art in information retrieval and web archiving to understand the following chapters.

Chapters 3 to 5 give characterizations of the state-of-the-art in WAIR and web archive users. In Chapter 3, two surveys on web archiving initiatives are presented, covering several aspects of web archiving, such as the volume of archived data, used formats, number of people engaged and the underlying technologies.

1.5 Overview

Chapter 4 studies the information needs of web archive users. I used three methods to collect quantitative and qualitative data from users, namely, search log mining, an online questionnaire answered by users while searching, and a laboratory study. In Chapter 5, search patterns and behaviors of users are researched. I conducted a quantitative analysis of the PWA search logs and compared it against the results obtained with users of web search engines.

Chapters 6 and 7 discuss how to improve the state-of-the-art in WAIR. Chapter 6 proposes an evaluation methodology to measure the effectiveness of WAIR systems and describes the construction of a test collection to empirically validate the methodology and support experiments. Chapter 7 introduces novel ranking features that exploit temporal information intrinsic to web archives and studies how to adapt ranking models to the evolution of web data throughout time. I built a specific dataset for this task that was made available to the research community to foster research in WAIR.

Chapter 8 concludes with an overall summary of the thesis and a discussion of some directions for future work.

Chapter 2

Background & State-of-the-Art

Information retrieval (IR) is a broad interdisciplinary research field that draws on many other disciplines, such as computer science, mathematics, cognitive psychology, linguistics and library science. It studies the computational search of information within collections of data with little or no structure ([Baeza-Yates & Ribeiro-Neto, 2011](#); [Manning *et al.*, 2008](#)). Often, IR deals with the matching of natural language text documents against users' queries, but it also studies other forms of content, such as the web and its search engines. The latter have become the dominant form of information access.

Web archiving is a research field concerned with the preservation of the information published on the web for future generations ([Masanès, 2006](#)). The dynamic and ephemeral nature of the web means that web sites are continually evolving or disappearing. Web archiving mitigates this problem by studying strategies to select, acquire, store and manage portions of the web. These strategies must handle the rapid obsolescence of technologies for contents to remain accessible and usable for as long as they are needed. The effective use of these archived contents is also object of research, including IR and analytical tools to extract knowledge from them.

This chapter presents a brief technical background and overview of the state-of-the-art in IR and web archiving, which are useful for understanding subsequent chapters. It starts by addressing the web archiving workflow in Section 2.1, with the different data transformation phases. Section 2.2 gives a glimpse of web archive initiatives around the world that strive to preserve information available

2. BACKGROUND & STATE-OF-THE-ART



Figure 2.1: Web archiving workflow.

on the web before it vanishes and the mechanisms developed to provide access to this information. The Portuguese Web Archive (PWA) is one of such initiatives that is showcased in this dissertation.

A clear understanding of the users is fundamental to support technical design decisions. However, studies of web archive users are rare in the research literature. Section 2.3 reviews the few existing user studies and surveys the users' information needs and search patterns on web archives and most similar IR systems.

Effective and efficient full-text search is still one of the greatest barriers to make web archives accessible to users. Novel ranking methods are proposed in this thesis to tackle this challenge. Section 2.4 shows how the ranking of search results is processed in search engines, and how learning to rank (L2R) technology and temporal information can improve it. Finally, Section 2.5 describes how to measure the effectiveness of the ranking methods and Section 2.6 presents a summary of the chapter.

2.1 Web Archiving Workflow

In a web archive, the data passes through several phases where they are transformed in a pipeline until presented to the user. Figure 2.1 illustrates the typical web archiving workflow with the following phases:

Acquisition: the web data can be acquired by several paths, such as from an entity that archived it previously or from the digitization of print publications (e.g. The Times Archive⁶). However, the most usual path is to crawl portions of the web. Crawling is the process of seeking and collecting data.

⁶<http://www.thetimes.co.uk/tto/archive/>

2.1 Web Archiving Workflow

It starts with the downloading of a set of URLs, which are then parsed to extract the URLs they link to. This process is continuously repeated for the extracted URLs that have not been downloaded yet, until a stop condition is met. The decision of what to archive is complex, since there is not enough storage space to preserve everything and the web is permanently growing. Thus, some web archives prefer a more granular selection to exhaustively crawl a limited number of web sites, such as the ones related to elections ([Paynter *et al.*, 2008](#)). Others prefer a wider selection of the web, but shallower, such as a top-level domain ([Gomes *et al.*, 2008](#)). The selection criteria of what to archive also depends on legal issues, such as copyright, data protection and libel ([Shiozaki & Eisenschitz, 2009](#)).

Storage: the web data from different sources are persistently stored on secondary memory. If the data sources are too heterogeneous, their data may be combined to provide users with a unified view (e.g. using ETL processes or specific wrappers). Usually, web archives concatenate sequences of compressed web documents into long files of size close to 100MB, where each document is preceded by a small header. This format, called ARC, was originally developed by the Internet Archive ([Burner & Kahle, 1996](#)). It offers an easier way to manage and speed up access to documents, since file systems have difficulty to handle billions of files. Recently, ARC was extended to the new WARC format that supports relations between contents ([ISO 28500:2009, 2009](#)). The web documents and their sites can undergo several processes during or after storage. For instance, they can be enriched with descriptive meta-data or their quality can be ensured by checking if all necessary files have been captured and will render. The requirements for authenticity and integrity depend on the purpose of the collection. Some cases require preserving only intellectual content, while others such as in legal evidence, may need the context of resources that include their provenance.

Indexing: the stored web data is read, uncompressed, broke up into words (tokenized) and syntactically analyzed (parsed) by the indexing system. Parsing

2. BACKGROUND & STATE-OF-THE-ART

is necessary to separate text from meta-data and identify the structural element to which each segment of the text belongs to (e.g. title, headings). It is challenging because there are hundreds of file formats that must be handled and continue to evolve, such as HTML, PDF or new formats. Other processes can be applied, such as the link extraction for link analysis algorithms and enhancing, with anchor text, the content of documents that the links point to. Then, index structures over the words and the meta-data are created for efficient search. Usually, the word occurrences in documents, fonts and positions are recorded in the index for better estimate document relevance. The inverted index (a.k.a. inverted file) is the index structure usually chosen, because it is the most efficient for textual search ([Zobel & Moffat, 2006](#)). Still, the efficiency of this structure can be further improved in web archives if time is considered as a criterion to partition and distribute it among several computers ([Costa *et al.*, 2013a](#)).

Searching: the index structures are used to lookup the documents that match a received query. This match depends of the implemented retrieval model. Usually, for large-scale collections such as the web, a retrieval model is chosen where all query terms (or related terms) must occur on the matching documents. Even so, and despite query optimizations to only select the best candidates from the billions of archived documents, millions of documents can match a query. This order of magnitude is too large for the users to efficiently explore and find information. Hence, the matching documents are ranked by their relevance scores that measure how well they satisfy a user's information need, formally represented by a query. This relevance is computed with a set of heuristics on data features, such as the query terms proximity on a document content or the number of links a document receives. The accessibility of web archives also depends on the laws of the country where they are hosted. For instance, the web archive of the National Library of France and the Finnish Web Archive are "dark archives" that are only accessible on-site ([Niu, 2012b](#)).

Presentation: the search results are formated and displayed in ranked lists for end user consumption. Usually, each result is augmented with meta-data,

2.1 Web Archiving Workflow

such as the title, URL and timestamp of when it was archived (Cruz & Gomes, 2013). A view with all the archived versions of a URL is also provided. Results can also be clustered by time for an easier perception of their temporal distribution or displayed along a timeline to support exploration tasks (Alonso *et al.*, 2009b). The search user interface of web archives contains some temporal controls, especially to narrow results by date range. When an archived document is shown, all of its hyperlinks are changed so that the references will point to the web archive instead of the live web. This enables users to interactively browse the web as it was in the past. There are also some visualization tools for mined archived content. For instance, the visualization tools of the UK web archive⁷ produce N-gram charts of the occurrence of terms or phrases over time and tag clouds of content written on web sites.

Preservation: this is a parallel process in this workflow, to guarantee that the web documents are accessible for long-term. Data must be replicated within the data center and between data centers spread across different geographic locations. Data must also be stored in a tamper-proof manner to prevent someone from rewriting history. Malicious people could try to take advantage of this fact for their own benefit. The monitoring of potential obsolescence in file formats and technology must be constant for a timely migration of the data before it is no longer accessible or usable. The preservation of a digital content must also include the preservation of the technology that supports the reproduction of the original content or the necessary steps for this technology be emulated in the future.

The acquisition, storage, indexing and preservation phases are conducted offline, while the searching and presentation phases are executed online.

2.1.1 Web Archive Architecture

Figure 2.2 presents the logical architecture of a web archive. Its main software components are overlapped over the web archiving workflow showing how they

⁷<http://www.webarchive.org.uk/ukwa/visualisation>

2. BACKGROUND & STATE-OF-THE-ART

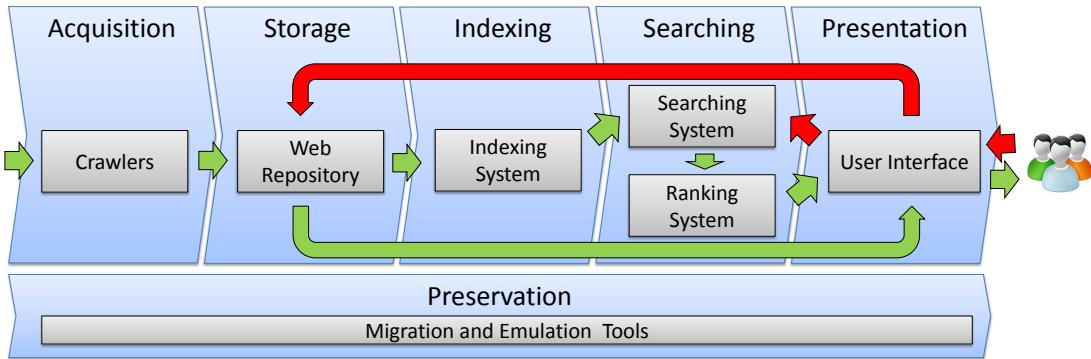


Figure 2.2: Web archive architecture.

execute the tasks of the phases of the workflow where they are displayed. Different configurations of this architecture can be set according to the requirements of a web archive. For instance, the web archives that operate over large-scale datasets use distributed architectures with many machines running parallel tasks.

The crawlers, the web repository and the indexing system, perform the crawling, storage and indexing, respectively. The searching phase is executed by the searching and ranking systems working in tandem. The searching system matches the documents against the queries and in some cases it may refine or expand the query using semantically similar terms, since the terms used for a given concept in the content may be different from the ones used in the query (e.g. *plane* vs. *aircraft*). In turn, the ranking system estimates the relevance of the matching documents for the queries. These documents are then sorted in descending order by their relevance score, which enables users to find information effectively and efficiently. The user interface enables the interaction between the users and the system. It receives the user requests and redirects the queries to the searching system and the requests of document versions to the web repository. It then performs the presentation of search results or document versions accordingly. The migration and emulation tools are used for the digital preservation. The green arrows (from left to right) in Figure 2.2 represent the data flow between the components, while the red arrows (from right to left) represent the user requests.

This thesis focus mainly on the searching process, despite the influence of all the other processes of the workflow in the final outcome. I described in a previous

work, the searching system architecture of the PWA and compared it with other existing searching architectures, in terms of performance, scalability and ease of management (Costa *et al.*, 2013a). I also compared the strategies to partition and distribute the indexes by time. However, in this thesis, I focus especially in the improvement of the ranking system, which is a core problem for information retrieval and web archiving.

2.2 Web Archiving Initiatives

Cultural heritage institutions, such as museums, libraries and archives, have been preserving the intangible culture of our society (e.g. folklore, traditions, language) and the legacy of physical artifacts (e.g. monuments, books, works of art). Web archives are a novel form of cultural heritage institutions mandated to preserve similar artifacts. However, the artifacts of web archives are digital-born and digitized contents.

Web archives are a special type of digital libraries. Both share the responsibility to preserve information for future generations. This includes all types of multimedia, such as images and videos, besides the digital counterparts of printed documents. The main difference is that web archives usually grow to a data size that exceeds traditional organization and management of typical digital libraries. Digital libraries are based on meta-data describing manually curated artifacts and catalogs of these artifacts, which are usually used to explore and search digital collections, for instance, through faceted search. However, the experience from the Pandora (National Library of Australia)⁸ and the Minerva (Library of Congress)⁹ projects showed that this is not a viable option for web archives. The size of the web makes traditional methods for cataloging too time-consuming and expensive, beyond the capability of libraries staff. One of the conclusions from the final report of the Minerva project is that automatic indexing should be the primary strategy for information discovery (Masanès, 2006).

The Internet Archive, a USA-based non-profit foundation, was one of the first web archives and has been broadly archiving the web since 1996. It leads the

⁸<http://pandora.nla.gov.au>

⁹<http://www.loc.gov/minerva>

2. BACKGROUND & STATE-OF-THE-ART

most ambitious initiative. In 2013, the Internet Archive was already preserving 240 billion archived documents with a total of about 5 PB of data (Kahle, 2013). The Pandora and Tasmanian web archives from Australia, and the Kulturarw3 web archive from Sweden, were also created in 1996. Many other initiatives followed since then and a significant effort has been employed by the research community to the web archiving domain. Many of these initiatives are members of the International Internet Preservation Consortium (IIPC), which leads the development of several open source tools, standards and best practices of web archiving (Grotke, 2008). In this thesis, I conducted two surveys to identify the web archiving initiatives across the world and collect comprehensive information about them. A timeline of some of these initiatives can be obtained online¹⁰.

Several research projects have been initiated for improving web archiving technologies. The Living Web Archives (LiWA) aimed to provide contributions to make archived information accessible and addressed IR challenges, such as web spam detection, terminology evolution, capture of stream video, and assuring temporal coherence of archived content (Masanès, 2011). LiWA was followed by the Longitudinal Analytics of Web Archive data (LAWA), which aims to build an experimental testbed for large-scale data analytics (Weikum *et al.*, 2011). Particular emphasis is given to developing tools for aggregating, querying and analyzing web archive data that has been crawled over extended time periods. The Web Archive Retrieval Tools (WebART) project focus on the development of web archive access tools especially tailored to facilitate research in humanities and social sciences (Huurdeeman *et al.*, 2013). The Collect-all ARchives to COnmunity MEMories (ARCOMEM) project was about developing innovative tools and methods to help preserving and exploiting the social web (Risse & Peters, 2012). The Memento project adds a temporal dimension to the HTTP protocol so that archived versions of a document can be served by the web server holding that document or by existent web archives if the web server do not contain the requested versions (Van de Sompel *et al.*, 2009). Users only have to install a browser plug-in, which makes this an easy solution for them to adopt.

¹⁰ <http://timeline.webarchivists.org>

2.2.1 Portuguese Web Archive

The Portuguese Web Archive (PWA) is one of the ongoing web archiving initiatives. The main scientific questions of this thesis are inseparably connected to this project. The main objectives of the PWA are to provide public access mechanisms to the archived information and ensure its long-term preservation. The PWA follows two projects in which I participated, one concerned with searching the Portuguese web and another with its preservation. The Portuguese web is broadly considered the part of the web of interest to the Portuguese.

Tumba!¹¹ was a web search engine optimized for the Portuguese web, which was available as a public service from 2002 to 2006 (Costa, 2004; Costa & Silva, 2010a). Several experiments were conducted on the different data processing phases of this project, spanning from the crawling of documents to the presentation of results.

Tomba was a web archive prototype for the Portuguese web operated between 2006 and 2007 (Gomes *et al.*, 2006). The main difference from the Tumba! web search engine was that Tomba provided support for the storage and access to several versions of documents from consecutive snapshots of the web. These snapshots came from Tumba! and included only the textual part of the crawled documents. The prototype was publicly available with 57 million documents searchable by URL.

The PWA is Tomba's successor since 2008 (Gomes *et al.*, 2008, 2013). It continues to archive the Portuguese web and has been extended to also archive the webs of some Portuguese speaking countries. The PWA archiving policy currently includes the set of documents satisfying one of the following rules:

1. hosted on a site under the Portuguese (.PT), Angola (.AO), Cape Verde (.CV) or Mozambique (.MZ) domains;
2. hosted on a site under other domain, but embedded in a document under the .PT, .AO, .CV or .MZ domains;
3. suggested by users and manually validated by the PWA team.

¹¹<http://xldb.fc.ul.pt/wiki/Tumba!>

2. BACKGROUND & STATE-OF-THE-ART

On average, 78 million files are downloaded in each crawl and 764 thousand files are downloaded each day. The PWA team has also integrated web collections from several other sources, such as the Internet Archive and the National Library of Portugal.

In January 2010, a beta version of a search service over the PWA was released and has since then been available at <http://archive.pt>. In December 2012, the service was providing public access to 1.2 billion (10^9) files, ranging from 1996 to 2011, and searchable both by full-text and URL. As far as I know, this is the largest web archive collection searchable by full-text and over such a large time span. The documents can then be accessed and navigated as they were in the past. Figure 2.3 depicts one of the historical documents from the beginning of the web in Portugal. It was the *homepage of Portugal* in 1996 with the country map and former Portuguese colonies, Macau and Timor. It is interesting to see hyperlinks to a homepage of Europe and another of the World, suggesting that the topology of the web was very different at that time. The PWA is also being used as a source of information for research and engineering projects through its OpenSearch API¹².

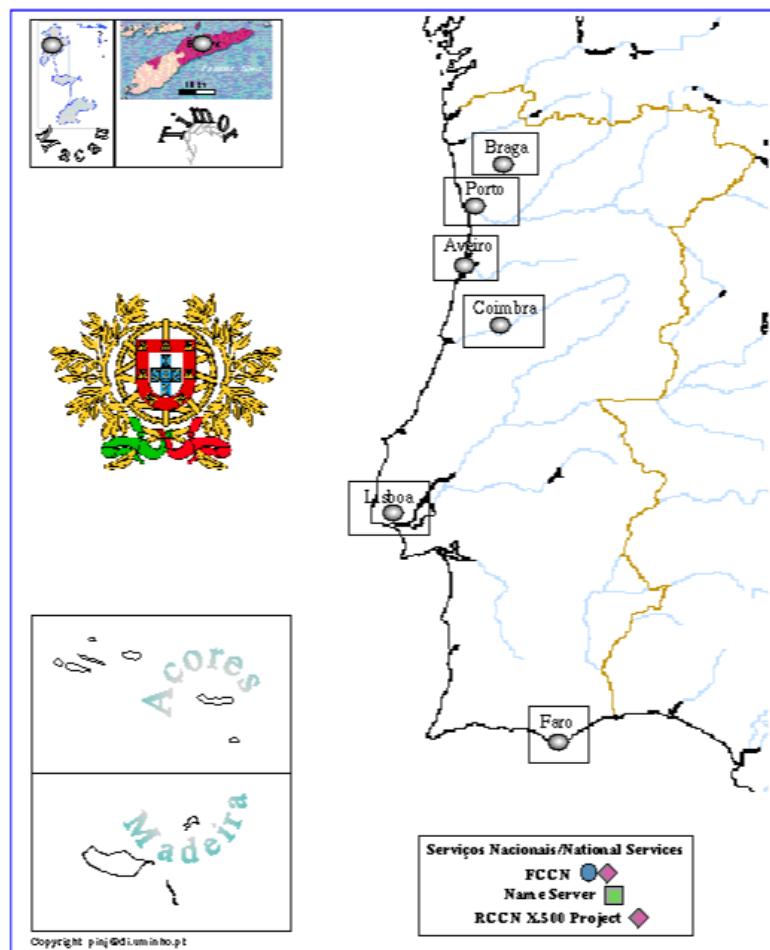
2.2.2 Access Types & Tools

Much of the effort on web archive development focuses on acquiring, storing, managing and preserving data (Masanès, 2006). However, the data must also be accessible to users who need to exploit and analyze them. Due to the challenge of indexing all the collected data, the prevalent access method in web archives is based on URL search, which returns a list of chronologically ordered versions of that URL, such as in the Internet Archive's Wayback Machine (Jaffe & Kirkpatrick, 2009; Tofel, 2007). The Internet Memory Foundation (2010) survey on European web archives reported that 68% of web archives support this type of access. However, URL search is limited, as it forces the users to remember the URLs, some of which refer to content that ceased to exist many years ago.

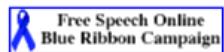
Another type of access is meta-data search, i.e. the search by meta-data attributes, such as category or theme. According to the Internet Memory Foun-

¹²<http://code.google.com/p/pwa-technologies/wiki/OpenSearch>

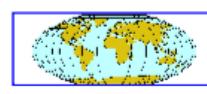
«Home Page» de Portugal / Portugal Home Page



 SAPO		Servidor de Apontadores Portugueses	
Acerca de Portugal	Portugal Cultural	Conferências e outros eventos	Ajuda
About Portugal	Cultural events		
Conferences and other events	Help		



Se não possuir suporte gráfico, prima a [aqui](#) / If you don't have graphics support, click [here](#)



[Europe Home Page](#) [EC Home Page](#) [World Home Page](#)

Figure 2.3: Document archived on October 13, 1996: *homepage of Portugal*.
Original URL: <http://s700.uminho.pt/homepage-pt.html>.

2. BACKGROUND & STATE-OF-THE-ART

dation (2010) survey, meta-data search is provided by 65% of European web archives. For instance, the Library of Congress Web Archives¹³ support search on bibliographic records. Some web archives support filtering results by domain and media type, while others organize collections by subject or genre to provide browsing functionality, such as the Pandora Australia’s web archive (Niu, 2012a). Nevertheless, most web archives support narrowing the search results by date range.

Full-text search has become the dominant form of information access, especially in web search systems, such as Google. These systems have a strong influence on the way users search in other settings. This explains why full-text search was reported as the most desired web archive functionality (Ras & van Bussel, 2007). Despite the high computational resources required for this purpose, 70% of the European web archives surveyed support full-text search for at least a part of their collections. Other results obtained, to be detailed ahead in this thesis, also show a strong preference of users for full-text search. As a result, I focus in this thesis on full-text search and in its challenges. These challenges have been previously addressed in some studies. For instance, in 2009, the Internet Archive indexed the first five years of their archive (1996-2000) and made them available for full-text search, but the results were poorly ranked and were full of spam (Dougherty *et al.*, 2010). In this thesis, I will show that the large majority of web archives that support full-text search presently use technology based on the Lucene search engine¹⁴ or extensions of Lucene to handle the data formats of web archives, such as NutchWAX¹⁵. The search services provided by these web archives are visibly poor and frequently deemed unsatisfactory. Cohen *et al.* (2007) showed that the out-of-the-box Lucene produces low quality results, with a MAP (Mean Average Precision) of 0.154, remarking that is less than half the MAP of the best systems participating in the TREC Terabyte track. Despite not evaluating the search effectiveness of web archives, these MAP results suggest that their effectiveness is poor.

¹³<http://www.loc.gov/webarchiving>

¹⁴<http://lucene.apache.org>

¹⁵<http://archive-access.sourceforge.net/projects/nutch>

There are several tools created for web archiving. The site¹⁶ of the International Internet Preservation Consortium (IIPC) has a list with many of these tools for acquisition, curation, storage and access. Thomas *et al.* (2010) present a comprehensive list of available tools and services that can be used in web archives. However, no one has identified which is the state-of-the-art technology to access web archive data. This thesis tries to change this reality and presents in Chapter 3 the first study providing a world-wide overview about the types of access and technologies used in web archives.

2.3 User Studies

2.3.1 Web Archive Users

Previous sections showed that there are several web archiving initiatives currently harvesting and preserving the web heritage. Still, very few studies about web archive users were made. The IIPC Access Working Group (2006) reported a number of possible user scenarios over a web archive. The scenarios are related to professional scopes, such as a journalist investigating a story or a lawyer looking for evidence, and have associated the technical requirements necessary to fulfill them. These requirements include a wide variety of search and data mining applications that have not been developed yet, but could one day play an important role. However, the hypothetical scenarios did not come directly from web archive users. Reynolds (2013) published a report with use cases of web archives related with data mining and visualization on archived contents. The report includes examples of tools and works performed with these tools.

The National Library of the Netherlands conducted a usability test on the searching functionalities of its web archive (Ras & van Bussel, 2007). Fifteen users participated in the test. One of the results was a compiled list of the top ten functionalities that users would like to see implemented. Full-text search was the first one, followed by URL search. Strangely, functionalities related with the time dimension were not mentioned on the top ten functionalities, despite this dimension being present in all the processes of a web archive. The users'

¹⁶<http://www.netpreserve.org/web-archiving/tools-and-software>

2. BACKGROUND & STATE-OF-THE-ART

choices can be explained by web archives being mostly based on web search engine technology. As a result, web archives offer the same search functionalities. This inevitably constrains user behavior. Another explanation is that Google became the norm, influencing the way users search in other settings.

The above studies provide limited information about web archive users. This thesis provides a deeper understanding of these users and addresses unanswered questions related to user information needs in Chapter 4 and search patterns in Chapter 5.

2.3.2 Information Needs

User information needs have been investigated in different IR systems, web search engines being the most studied. There exists a consensus among researchers about the taxonomy proposed by Broder (2002) and refined by Rose & Levinson (2004). Broder classified web search engine queries into three broad classes according to the user goal:

navigational to reach a web page or site in mind;

informational to collect information about a topic, usually from multiple pages without a specific one in mind;

transactional to perform a web-mediated activity (e.g. shopping, downloading a file, finding a map).

Broder used two methods to determine the percentages of queries in each of these classes. The first, was a pop-up window with a questionnaire presented to random users. It achieved a response ratio of about 10%. The second, involved the manual classification of 400 queries. Both methods were applied on the Altavista web search engine and the results drawn from them presented a good correlation. Rose and Levinson extended Broder's taxonomy of web search, creating subclasses for the informational and transactional categories. They analyzed not only the queries, but also the clicks on results and the subsequent queries made by the users. They manually classified three sets of approximately 500 queries randomly selected from the Altavista search logs. There are other taxonomies

for web search proposed in the literature. [Jansen *et al.* \(2008a\)](#) presented an integrated view of them.

Different IR systems and environments have users with different information needs. For instance, [Church & Smyth \(2009\)](#) used diary studies to explore information needs of mobile users. Three needs were identified. The first is the same informational need that web search engine users have. The second is a geographical need, similar to an informational need, but dependent on location. The third is a personal information management need, focused on finding private information of the user.

2.3.3 Search Patterns

Web usage mining focuses on using data mining to analyze search logs or other activity logs to discover interesting patterns. [Srivastava *et al.* \(2000\)](#) pointed five applications for web usage mining: personalization, for adjusting the results according to the users' profile; system improvement, for a fast and efficient use of resources; site modification, for providing feedback on how the site is being used; business intelligence, for knowledge discovery aimed to increase customer sales; and usage characterization to predict users' behavior. I focus on usage characterization.

There are several user study methods that can be used for search pattern analysis ([Kelly, 2009](#)). Qualitative studies, such as surveys ([Aula *et al.*, 2005](#); [Teevan *et al.*, 2004](#)) and laboratory studies ([Aula *et al.*, 2010](#); [Kellar *et al.*, 2007](#)), provide rich information that can explain some of the patterns found, especially when using quantitative studies, such as log analysis ([Fox *et al.*, 2005](#); [Jansen & Spink, 2006](#)).

An analysis on the access logs of the Internet Archive's Wayback Machine showed that most users request a single URL, while only a few users see several versions of the same URL published throughout time ([AlNoamany *et al.*, 2013](#)). Another study showed that users of the Wayback Machine requested mostly pages written in English, followed by pages written in European languages. Most users searched for or linked to pages archived in the Wayback Machine, likely because the requested pages no longer existed on the live web. Most pages link to versions

2. BACKGROUND & STATE-OF-THE-ART

of 2008 and then there is a sharp decline as the years diminish. It seems that the referrer wants to redirect to the most recent archived version.

The above are the only known studies related with search patterns in web archives, but they only address URL search. Several other studies scan logs from web search engines with the goal of understanding how these systems were used. A common observation across these studies is that most users conduct short sessions with only one or two queries, composed by one or two terms each ([Jansen & Spink, 2006](#)). This discovery implies that the use of web search engines is different from traditional IR systems, which receive queries three to seven times longer ([Jansen *et al.*, 2000](#)). Queries for special topics (e.g. sex), special types (e.g. question-format) and multimedia formats (e.g. images) are also longer ([Markey, 2007](#)). This shows that the search patterns vary not only among IR systems, such as search engines, online catalogs and digital libraries, but also depend on the type of information that users search. According to [Weber & Castillo \(2010\)](#), another aspect that differentiates search patterns is users' demographics (i.e. age, gender, ethnicity, income, educational level).

2.4 Ranking

Large-scale IR systems usually retrieve millions of documents matching a full-text query, which makes it extremely hard for a user to find relevant information. To overcome this problem, ranking models estimate document relevance based on how well documents match user queries ([Baeza-Yates & Ribeiro-Neto, 2011](#); [Manning *et al.*, 2008](#)). Documents are then sorted in descending order by their relevance score as a mean for users to find information effectively and efficiently. Next, I present some of the ideas about the existing models and how to create them, which is a central problem in IR, web archives in particular.

2.4.1 Conventional Ranking Models

Early IR systems used the Boolean model, based on set theory and Boolean algebra, which consider a document relevant only if it contains all query terms. Later models, such as the Vector Space Model (VSM), allowed partial matches

2.4 Ranking

of query terms and document ranking through the computing of relevance degrees for each document. In the VSM, both documents and queries' terms are represented as vectors in an Euclidean space, where the dimensionality of the vectors is the number of distinct terms in the collection. The inner product between the vectors measures the query and document similarity. The TF-IDF is one of the well-known functions that computes term weights for a document vector (Salton & Buckley, 1988). The term weight increases proportionally to the number of times a term occurs in the document and decreases with the frequency of the term in the collection. There are other weighting functions that provide good results, such as the formula BM25, which normalizes the weight by document length (Robertson *et al.*, 1995). Some variants, such as BM25F, take the document structure into account (Zaragoza *et al.*, 2004). BM25F scores a term differently if it occurs on the title, URL or anchor texts of other documents linking to the document. BM25 and its variants are based on the probabilistic relevance framework introduced by Robertson & Jones (1976).

All the above ranking models assume that terms are independent (bag-of-words model). For example, a document would have the same relevance for the query *European Union*, whether the query terms occurred together or far apart. Some models overcome this by considering the terms' proximity (Tao & Zhai, 2007). Language models estimate the probability of a document generating the terms in the query (Song & Croft, 1999). They handle the dependency between query terms by taking into consideration the fact that the probability of a term depends on the probability of previous adjacent terms.

Other type of data can be exploited to create ranking models besides the document content. For instance, social annotations from sites, such as *delicious.com*, provide a good summary of the key aspects of the document (Bao *et al.*, 2007). Logs of search engines are an exceptional source to analyze where the users clicked after submitting a query (Joachims, 2002; Radlinski & Joachims, 2005).

All previous models estimate the documents' relevance according to a given query and that is why they are denoted query-dependent models. On the other hand, query-independent models rank documents according to an importance, quality or popularity measure computed independently of the query. One of the most used sources to compute importance values is the hyperlink structure of

2. BACKGROUND & STATE-OF-THE-ART

the web. In turn, one of the most well-known algorithms that uses this source is PageRank, because it is partially responsible for the Google’s initial success (Page *et al.*, 1998). PageRank relies on the assumption that the importance of a document depends on the number and the importance of the documents linking to it. There are many other algorithms taking use of the web link structure, such as HITS (Kleinberg, 1999) or HostRank (Xue *et al.*, 2005). There are also algorithms considering different sources for basing documents’ importance. For instance, Kraaij *et al.* (2002) considered the URL depth and Richardson *et al.* (2006) the number of times a document was visited, the document length or the degree of conformance to W3C standards.

2.4.2 Temporal Ranking Models

Some works leveraged temporal information to improve ranking models. One of the most common ideas is incorporating in language models the heuristic that the prior probability of a document being relevant is higher in the most recent documents (Li & Croft, 2003). Boosting the most recent documents is desirable for queries where the user intends to find recent events or breaking news, such as in news search engines. Another idea, by Elzas & Dumais (2010), is to favor more dynamic documents, since documents with higher relevance are more likely to change or change to a greater degree. According to Adar *et al.* (2009), more popular and revisited documents are also more likely to change. On the other hand, the most persistent terms are descriptive of the main topic and likely added early in the life of a document (Adar *et al.*, 2009; Aji *et al.*, 2010). These persistent terms are especially useful for matching navigational queries, because the relevance of documents for such query terms is not expected to change over time.

The distribution of the documents’ dates reveals time intervals that are likely to be of interest to the query. For instance, when searching for *tsunami*, the peaks in the distribution may indicate when tsunamis occurred. Thus, some studies exploited the distribution of the publication dates of the top-k query matches to boost documents published within relevant intervals (Dakka *et al.*, 2010; Jones & Diaz, 2007). However, identifying the dates of web documents is

not straightforward. The meta-data from the document’s HTTP header fields, such as Date, Last-Modified and Expires are not always available, nor reliable. For instance, servers often send an invalid Last-Modified date of when the content was changed (Clausen, 2004). Studies estimate that from 35% up to 64% of web documents have valid last-modified dates (Amitay *et al.*, 2004; Gomes & Silva, 2006). However, these percentages can be significantly improved by using the dates of the web document’s neighbors, especially of the web resources embedded in the selected document, such as images, CSS and JavaScript (Nunes *et al.*, 2007).

The content itself is a valuable source of temporal information, but is likely to be the most difficult to handle. Temporal expressions can be extracted from text with the help of NLP and information extraction technology (Alonso *et al.*, 2007). Their inherent semantic is then mapped into the corresponding time intervals, which are used to measure the temporal distance to the search period of interest. Thus, instead of treating temporal expressions as common terms, they can be integrated in the language model to estimate the probability of a document generating the temporal part of the query (Berberich *et al.*, 2010; Irem Arikān & Berberich, 2009). Notice however, that these expressions may refer to a time completely different from the publication date of the document. For instance, they can refer to an event occurred in the past or future.

Query logs are another source that can be exploited, for instance, to detect temporal implicit intents in queries (Metzler *et al.*, 2009). If a query is likely to contain calendar years then, it may have a temporal intent. In this case, the documents having those years in their content should be boosted. Micro-blogging sources, such as Twitter, can also be used to improve the ranking of web search engines when the users expect information that is both topically relevant and fresh (Dong *et al.*, 2010b).

Temporal information can also improve link-based ranking algorithms. A known problem in these algorithms is that they underrate recent documents, because the indegree used to compute the popularity of web documents, such as in PageRank algorithm (Page *et al.*, 1998), favors older documents that have already accumulated a significant number of references over time. This problem can be overcome by weighing higher the inlinks of the sources updated more

2. BACKGROUND & STATE-OF-THE-ART

recently ([Amitay et al., 2004](#); [Dai & Davison, 2010](#); [Yu et al., 2004](#)). The idea is to reflect the freshness of source documents on the importance of the document they link to. Additionally, the update rates of the sources can also be considered ([Berberich et al., 2005](#)) or the obsolete links that point to documents that are no longer accessible ([Bar-Yossef et al., 2004](#)). This gives a clear indication that the documents have not been maintained and contain outdated information.

In Chapters [6](#) and [7](#) of this thesis, I study and present novel temporal ranking models to improve WAIR.

2.4.3 Learning to Rank

The conventional and temporal ranking models presented above are just a few examples of the large number of proposals over the years. They exploit different features to determine whether a page is relevant for a query. The question now is which ones are better suited for web archives?

Previous IR evaluations showed that combinations of ranking models tend to provide better results than any single model ([Brin & Page, 1998](#); [Craswell et al., 2005](#); [Liu et al., 2007](#)). An individual model is also more susceptible to influences caused by the lack or excess of data (e.g. spam). Therefore, it is advantageous to use different aspects of the data to build a more precise and robust ranking model. By robust, I mean a model capable of coping well with variations in data. For instance, a document can receive a low relevance score due to a small query term frequency, but a high number of inlinks can identify the document as important. All these factors must be properly balanced by the model.

The generation of a ranking model can be decomposed in a four step pipeline:

1. extraction of low-level ranking features, such as the term frequency or document length;
2. assembling of the latter in high-level ranking features (a.k.a. ranking functions), such as BM25 ([Robertson et al., 1995](#));
3. selection of the most suitable features for a retrieval task;
4. combination of the features in a way to maximize the results' relevance.

For simplicity, this combination can be linear, i.e. for a query-document pair with a vector of ranking features associated, \vec{d} , the values produced by the n selected ranking features are added after each feature f_i is weighted by a coefficient λ_i and adjusted with a value b_i :

$$\text{rankingModel}(\vec{d}) = \sum_{i=1}^n \lambda_i f_i(\vec{d}) + b_i$$

However, the best combination between features can be non-linear. There are several ways to combine them non-linearly. One solution is to map features from its original space into a high-dimensional space, $\vec{d} \mapsto \Phi(\vec{d})$. Then by the means of the *kernel trick* it is possible to apply linear methods to non-linear data ([Schölkopf & Smola, 2002](#)). Another solution is to combine features in a non-linear way, such as in the case of genetic programming through the use of crossover and mutation operations ([Yeh et al., 2007](#)).

The first two steps of the model generation are well studied and some ranking features, such as BM25, are good ranking models by themselves ([Manning et al., 2008](#)). However, combining them manually is not trivial. There are search engines using hundreds of features. Manual tuning can lead to overfitting, i.e. it fits training data closely, but fails to generalize to unseen test data. Hence, in the last few years the fourth step has been concentrating attention from the machine learning and information retrieval communities. Supervised learning algorithms have been employed to tune the weights between combined ranking features, resulting in significant improvements ([Liu, 2009](#)). As illustrated in Figure 2.4, these so-called L2R algorithms that compose a learning system receive as input a training set composed by n queries, where each query q has associated a set of p feature vectors \vec{d} and a set of relevance judgments y . The L2R algorithms then learn ranking models by minimizing the difference between their prediction and the relevance judgments y . Finally, each model is tested with a test set similar to the training set. The predictions of the model are compared with the known relevance judgments (*ground truth*) to measure its effectiveness.

The way L2R algorithms learn can be categorized into three approaches:

pointwise approach estimates the relevance score of each document with respect to a query, by using each document feature vector as a training instance

2. BACKGROUND & STATE-OF-THE-ART

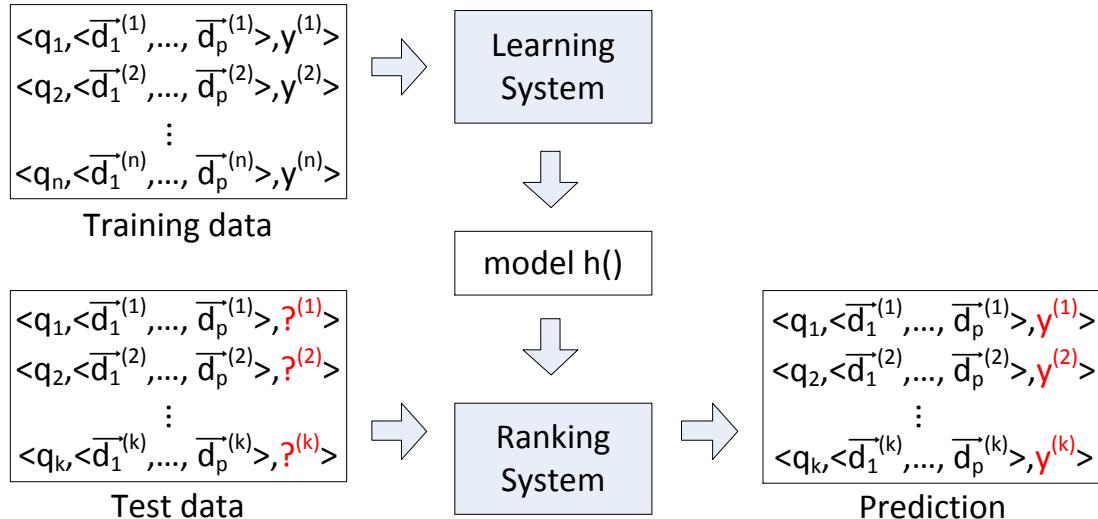


Figure 2.4: A general paradigm of L2R (copy from ([Liu, 2009](#))).

([Breiman, 2001](#)). This approach treats the L2R problem as a standard classification or regression task. An example is PRank that learns through ordinal regression ([Crammer & Singer, 2002](#));

pairwise approach uses feature vectors of pairs of documents as instances to learn a model that minimizes the pairs ranked in the wrong relative order (e.g. d_1 is more relevant than d_2) ([Freund *et al.*, 2003](#)). This approach enables to use clickthrough data from search engines as relevance judgments ([Joachims, 2002](#));

listwise approach trains with feature vectors of a ranked list of documents associated with a query. The learned model minimizes the permutations between pairs of documents ([Cao *et al.*, 2007](#)) or minimizes IR measures used in evaluation ([Xu & Li, 2007](#)). This approach, contrary to others, considers the rank position of documents.

While the pointwise approach only focuses on one document at a time, the pairwise considers the dependency between documents. Even so, there is a gap between IR evaluation measures used to evaluate the model and measures used to learn the model. To overcome this, the listwise approach considers the position of the documents in the ranked list and their association with a query.

2.4.4 Query-Type-Dependent Learning to Rank

The L2R framework learns ranking models that fit all training data. However, a generic model is not always the best solution and may be overcome by a query-type-dependent model. Kang & Kim (2003) showed this by automatically classifying queries and then creating a ranking model for each query class. However, it is often hard to classify a given web search query due to its small number of terms, which makes this technique unfeasible in some cases or imprecise when the wrong model is chosen.

To avoid the misclassification problem, Geng *et al.* (2008) proposed a K-Nearest Neighbor algorithm for query-type-dependent ranking. They created a ranking model for each query q by using the k-nearest training queries of q measured by the similarity of their feature values. The query feature values were computed as the mean of the feature values of the top search results ranked by a reference model (BM25). However, the training time required to create all these models is quite large and each model is learned with just the training data associated to the k-nearest queries.

Bian *et al.* (2010b) proposed a method that learns a ranking model and a soft query classifier simultaneously. They used a loss function per class that was weighted by the soft query classifier as the probability of a query belonging to the class. The sum of the weighted loss functions forms a global loss function. The proposed method has the advantage of learning with the whole training data instead of using just a part, as in previous works. However, the required query taxonomies for classification must be available, precise and fine-grained enough. Bian *et al.* (2010a) developed a different method that does not require a pre-defined query taxonomy. They employed a clustering method to identify a set of query topics based on features extracted from the top search results. Then a set of models, one per query topic, were simultaneously learned, by minimizing a global loss function that combines the ranking risks of all query topics. Each query contributed to learn each model according to the similarity between the query and the respective topic. Thus, each model was learned using the entire training data.

2. BACKGROUND & STATE-OF-THE-ART

Dai *et al.* (2011) followed this work, but integrated the criteria of freshness with topical relevance to simultaneously optimize both. Freshness and topical relevance labels were combined with a weighted harmonic mean to form a single optimization function. Dong *et al.* (2010a) also optimized both freshness and topical relevance using a single optimization function, but their labels were combined by demoting the topical relevance if the results were somehow outdated. However, they developed a classifier to identify breaking-news queries and explored three approaches to learn a ranking model for such queries. Their approaches, which aimed to solve the problem of the insufficient freshness training data, include: a *compositional model* learned with freshness training data and the output of another model learned with topical training data; a *over-weighting model* learned using a loss function with different weights for the topical and freshness training data; and an *adaptation model* composed by regression trees learned with topical training data and appended with other trees learned with freshness training data.

My work is inspired by the work of Bian *et al.* (2010a), but I innovated by learning ranking models taking into account the specificities of each time period. It will be described in detail in Chapter 7.

2.4.5 Datasets for Learning to Rank

In the last years, large web search engine companies such as Microsoft, Yahoo! and Yandex have made benchmark datasets available to research L2R. These datasets aggregate IR test collections, including their corpora, query sets, relevance judgments, evaluation metrics and evaluation tools. In addition to that, the datasets have different feature values extracted for each $\langle query, document \rangle$ pair, eliminating the usual parsing and indexing difficulties. The results of some state-of-the-art L2R algorithms are also provided for a direct comparison.

LETOR was released by Microsoft Research Asia in 2007 and was the first dataset publicly available (Qin *et al.*, 2010). It was constructed based on multiple data corpora and query sets available from TREC competitions that are widely used in the IR community. Many researchers have been using LETOR, but some noticed that conclusions drawn from experiments on LETOR and from large real datasets were different. LETOR was too small to draw reliable conclusions. For

instance, [Taylor *et al.* \(2008\)](#) reported that SoftRank achieved an improvement of 15.8% on TREC data, but on their internal web search data the improvement was negligible. Since then, larger datasets have been released.

In 2009, the Russian web search engine Yandex released an internal L2R dataset for a competition called Internet Mathematics¹⁷. The dataset includes 9 124 queries and 245 features for each *<query, document>* pair. In 2010, Yahoo! Labs released two datasets used internally and organized a L2R challenge to promote the datasets and foster the development of state-of-the-art L2R algorithms ([Chapelle & Chang, 2011](#)). The released datasets comprise 36 thousand queries, 883 thousand documents and 700 features. The datasets of Yandex and Yahoo! do not contain the original queries or the URLs of original documents, neither reveal the semantics of the features. Web search engines rarely disclose this information to avoid the reverse engineering of the ranking features.

In 2010, two other datasets were released. Microsoft released a dataset with more than 30 thousand queries and 136 features extracted from Microsoft Bing¹⁸. [Alcântara *et al.* \(2010\)](#) released a dataset with 29 clickthrough features extracted from the search logs of the TodoCL search engine.

None of the existing datasets contain temporal features or any features created from web archives. To complement the above datasets, I created and released a dataset to foster research in L2R for WAIR, which will be described in Chapter [7](#).

2.5 IR Evaluations

IR evaluations straddle two opposite, but complementary views: a user-centered and a system-centered ([Kelly, 2009](#)). The goal of user-centered evaluations is to measure how people can use a system to retrieve relevant documents. These evaluations provide rich qualitative data about user interactions with the system, for instance, from experiments with users in a laboratory ([Aula *et al.*, 2010](#)) or in their natural environment (*in situ*) ([Kellar *et al.*, 2007](#)). The goal of system-centered evaluations is to quantify the extent to which a system retrieves relevant documents, independently of how well users interact with it. The most popular

¹⁷<http://imat2009.yandex.ru/en/datasets>

¹⁸<http://research.microsoft.com/en-us/projects/mslr>

2. BACKGROUND & STATE-OF-THE-ART

example is the Cranfield paradigm established in the 1960s by [Cleverdon \(1967\)](#). This paradigm defines the creation of test collections for evaluating retrieval results composed by three parts, namely:

- a corpus** representative of the items (often documents) that will be encountered in a real search environment;
- a set of topics** describing user information needs;
- relevance judgments** (a.k.a. *qrels*) indicating the degree of relevance of each document retrieved for each topic.

The effectiveness of an IR system is then measured by comparing its results against the known relevant documents for each topic.

Assessing all documents for each topic with degrees of relevance is impractical due to the size of web collections. Hence, assessment paradigms were designed to diminish the human effort, while maintaining a sufficient assessment coverage to guarantee reliable evaluations. Next, I present the three most used assessment paradigms that can be applied in WAIR.

2.5.1 Pooling

Pooling is based on the assumption that the top-ranked documents of many and diversified IR systems aggregate most of the relevant documents ([Voorhees & Harman, 2005](#)). For that, each participant (group or individual) submits several runs, where each run corresponds to a list of the top-ranked documents (usually 1 000) for each query derived from a topic. The pool aggregates the top-ranked documents (usually 50 or 100) for each of the selected runs, which are then judged with relevancy degrees (usually binary or ternary) by several expert assessors following strict guidelines. All unpooled documents are considered not-relevant. The pool is considered the ground-truth and is used to evaluate all the submitted runs. Results show that a total of 50 topics and 100 top-ranked documents assessed per topic is sufficient to fairly compare the IR systems, i.e. the ranking between the evaluated systems is stable even if their performance scores vary after changing the pool ([Buckley & Voorhees, 2000](#); [Sanderson, 2005](#); [Voorhees, 2000](#);

Zobel, 1998). The diversity of the runs is also important to find new techniques that present good results.

Researchers contribute to this process in order to benefit from the collected data. The problems of pooling is that first, it is necessary to motivate a significant part of the research community to resolve an IR problem. Second, it requires a great deal of effort by everyone to assess the documents, even with modifications in pooling to reduce this effort (Aslam *et al.*, 2006). For instance, the average number of documents assessed per topic on the first eight years of the TREC's ad-hoc tracks was 1 464 (Voorhees & Harman, 1999). Third, human judges have a relatively low agreement due to the inherent subjectivity of the task (Bailey *et al.*, 2008; Voorhees, 2000). Still, the ranking between the evaluated systems is resilient to the judge variation detected.

2.5.2 Implicit Feedback

Logs of search engines can be analyzed to improve their ranking quality (Joachims, 2002; Radlinski & Joachims, 2005) and model user interactions (Jansen & Spink, 2006; Markey, 2007). Many studies follow this approach, because it makes it possible to record and extract a large amount of implicit feedback at low cost. Top commercial web search engines receive hundreds of millions of queries per day. Logs also have the advantage of being a non-intrusive mean of collecting user data about the searching process. Most users are not aware that their interactions are being logged, which leads them to behave as if they were not under observation. Another use of this feedback is that it can be used to produce relevance judgments over the specific collections being served, in contrast to more general collections made available for testing.

On the other hand, search logs are limited to what can be registered. In public search engines, there is often no contextual information about the users, such as their demographic characteristics, the motivations that lead them to start searching, and their degree of satisfaction with the system. The major disadvantage of collecting implicit feedback is that the gathered data is noisy, thus being hard to interpret. For instance, Fox *et al.* (2005) discovered that the viewing time of a document is an indicator of relevance. However, the amount of

2. BACKGROUND & STATE-OF-THE-ART

time the document is open after selected, does not necessary correspond to the reading time by the user.

Clicks on the result lists provide important feedback about the users choices. However, this data is also problematic because it contains many false positives (clicks on not-relevant documents due to misleading ranking or snippets) or false negatives (relevant documents that are not clicked because they are placed too low in the ranking or have poor snippets). The noise can be mitigated by considering a large number of replicated feedback. According to [Joachims *et al.* \(2005\)](#), the clicks are reasonably accurate if they are used as relative judgments between documents on ranked lists of results. For instance, if the second result is clicked and the first is not, then we can conclude that the second tends to be more relevant.

2.5.3 Crowdsourcing

Crowdsourcing emerged as an alternative to conduct relevance evaluations ([Alonso *et al.*, 2008](#)) and user studies ([Kittur *et al.*, 2008](#)) by taking advantage of the power of millions of people connected through the Internet. The idea is to post tasks on the web in the form of an open call, which are outsourced by a large group of online users in exchange of a small payment. These are easy tasks for people, but hard for computers. For instance, assessing the relevance of documents.

There are several online labour markets for crowdsourcing. Amazon Mechanical Turk¹⁹ has been adopted in many of the crowdsourcing relevance evaluations ([Alonso & Mizzaro, 2009](#); [Alonso *et al.*, 2008](#); [Kittur *et al.*, 2008](#)). It accepts just about anyone possessing basic literacy. Its use requires splitting large tasks into smaller parts for people willing to complete small amounts of work for a minimal amount of money.

There are other applications exploiting the power of crowdsourcing by presenting the tasks as a game to motivate participants ([von Ahn & Dabbish, 2008](#)). A successful example is the Google's Image Labeler, where players label images for free while they play ([von Ahn & Dabbish, 2004](#)). Besides entertainment, user participation can be pursued by promoting their social status, for instance using

¹⁹ <http://www.mturk.com>

	Relevant	Not-relevant
Retrieved	a	b
Not retrieved	c	d

Table 2.1: Contingency table of the variables that form IR evaluation measures.

leader boards. Hybrid approaches may consider aspects of entertainment and social status, but also monetary rewards to winners.

This paradigm substitutes expert judges by non-experts, which creates doubts about the assessments' reliability. [Bailey et al. \(2008\)](#) concluded that there is a low level of agreement between both groups. As consequence, this produces small variations in performance that can affect the relative order between the assessed systems. [Snow et al. \(2008\)](#) showed the opposite. A few non-experts can produce just as good or even better judgments than one expert. [Alonso & Mizzaro \(2009\)](#) used the TREC data to demonstrate in a small scale (for 29 documents of a topic) that Mechanical Turk users were accurate in assessing relevance and in some cases were more precise than the original experts.

2.5.4 Evaluation Measures

Many IR evaluation measures exist to quantify the system performance or different aspects of user satisfaction. When using test collections for evaluation, many measures are created from a combination of variables exhibited in the contingency Table 2.1. Precision (P) and recall (R) are two of the most known IR evaluation measures that use these variables:

$$P = \frac{a}{a+b}$$

$$R = \frac{a}{a+c}$$

Precision measures the fraction of retrieved documents that are relevant and recall measures the fraction of relevant documents that are retrieved. There is an inverse relation between both measures. If a query is broadened to increase recall by finding more relevant documents, more not-relevant documents will also be inadvertently added and the precision will drop. On the other hand, if a query is

2. BACKGROUND & STATE-OF-THE-ART

narrowed to target more relevant documents and avoid not-relevant, precision will likely increase, but some relevant documents will be inevitably discarded leading recall to decrease.

In some IR systems, such as in web search engines, precision is more important than recall. For these systems, previous studies show that users tend to see only the first page of search results ([Jansen & Spink, 2006](#)). Thus, some evaluations use precision at fixed cutoffs n , for instance precision at 10 (P@10), where only the top 10 search results are examined to check the number of relevant documents, $r(n)$ ([Manning *et al.*, 2008](#)). Precision at cutoff n is defined as:

$$P@n = \frac{r(n)}{n}$$

Previous measures ignore the ranking position of relevant documents. However, users consider the ranking and expect that IR systems will retrieve the relevant documents ranked at the higher positions as possible. Hence, ideal measures should reflect this behavior. One of the most used measures considering the ranking position is Average Precision (AP). It gives in a single value the average of precisions across various levels of recall ([Manning *et al.*, 2008](#)):

$$AP = \frac{\sum_{i=1}^n P@i * rel(i)}{r}$$

where n is the number of documents retrieved, i the rank position, $rel(i)$ a function that returns 1 if the document at position i is relevant or 0 otherwise, and r the total number of relevant documents for the searched topic. In a nutshell, AP calculates the average of the precision at the rank position of each relevant document. Mean Average Precision (MAP) calculates the mean of AP for all topics. Both measures have shown to be stable and discriminate well among retrieval strategies ([Buckley & Voorhees, 2000, 2004](#)).

Success at rank k (S@ k) is a simple measure that calculates the proportion of queries for which one or more relevant documents are in the top k search results ([Craswell & Hawking, 2005](#)). For instance, S@10 indicates how often an IR system finds at least one relevant document in the top 10, which typically is the first page of search results.

The mentioned measures until now can only be used with binary judgments (relevant or not-relevant). When multiple grades of relevance are available, the

Normalized Discount Cumulative Gain (NDCG) is one of the most used measures ([Järvelin & Kekäläinen, 2002](#)). It is the extension of the Cumulative Gain (CG) that is the sum of the relevance values, $rel(i)$, from the top n retrieved documents:

$$CG@n = \sum_{i=1}^n rel(i)$$

$CG@n$ ignores the rank of the documents at the top n . The Discounted Cumulative Gain (DCG) overcomes this by discounting progressively the relevance values as the ranking moves down. The discount is a log-based function. DCG at cutoff n is defined as:

$$DCG@n = rel(1) + \sum_{i=2}^n \frac{rel(i)}{\log_2(i)}$$

NDCG normalizes DCG over an ideal ordering of the relevant document, IDCG, to get a value between 0 and 1, with 1 representing the ideal ranking:

$$NDCG@n = \frac{DCG@n}{IDCG@n}$$

2.6 Summary

This chapter presented the fundamental concepts of information retrieval and web archiving that serve as introductory content for the rest of this work. In particular, it outlined the typical web archiving workflow and showcased the Portuguese Web Archive (PWA), which provided most of the data used in this research work. It also surveyed important web archiving initiatives across the globe. I can conclude that the research community has been dedicated significant effort to improving web archiving technologies. However, the information about the state-of-the-art in WAIR technology is scarce.

One main idea from the research literature is that web archives usually grow to a data size that exceeds the capacity of traditional digital library management methods, based on human generated meta-data. Automatic indexing should be the main strategy for information search. The studies related to web archive users showed that full-text is the most desired web archive functionality. However, there is no evaluation of the technology used by current web archives to support full-text search. There is also a lack of information about the users which inhibits

2. BACKGROUND & STATE-OF-THE-ART

the development of effective and useful technology. Several unanswered questions related to user information needs and search patterns require research.

This chapter finalizes with a description of how to compute the ranking of full-text search results and how rankings are evaluated. It also introduced ranking models and methods for automatically creating such models using the L2R framework. Another important topic discussed was how to leverage temporal information to improve IR. The L2R framework and temporal information were never used to improve WAIR despite their good results in other IR areas.

Chapter 3

Characterizing Web Archives

The previous chapter provided an overview of information retrieval and web archiving. Web archiving initiatives incorporate both types of technology. However, despite the existence of web archives since 1996 and of their joint efforts to preserve the web, the information about web archiving initiatives and the services they provide is scarce. Without knowing the status of the current web archiving technology it is impossible to understand its limitations and what developments are still needed for their users.

Motivated by the lack of knowledge in the research community about the state-of-the-art in web archiving, I have conducted two surveys that provide the most comprehensive picture of world-wide initiatives aimed at preserving information published on the web. The two surveys gathered results about existing web archiving initiatives and analyzed characteristics, such as the location, creation year, selection policy, used formats, number of people engaged, volume of archived data, access type and employed technology. I also analyzed the evolution of web archiving initiatives from 2010 to 2014.

The main contributions reported in this chapter are:

1. a comprehensive characterization of the status of web archiving and an analysis of its evolution;
2. a characterization of the state-of-the-art in WAIR technology and the identification of its limitations;

3. CHARACTERIZING WEB ARCHIVES

3. a Wikipedia page created with information about web archiving initiatives that has been collaboratively kept up-to-date by the community.

The remainder of this chapter is organized as follows. Section 3.1 describes the methodology employed on the surveys on web archiving initiatives conducted in 2010 and 2014. Section 3.2 presents the results obtained from the surveys and an analysis of the advancements made in web archiving in that period. Section 3.3 finalizes with a summary of the chapter.

3.1 Methodology

Initially, this research aimed to obtain answers to the following questions about web archiving initiatives across the globe:

1. What is the name of your web archiving initiative (please state if you want to remain anonymous)?
2. How many people work at your web archive (in person-month)?
3. Which is the amount of data that you have archived (number of files, disk space occupied)?

During October 2010, together with my colleagues at the PWA, I have attempted to gather this information from the official sites of known web archives and published documentation, but had little success because the published information was frequently insufficient or obsolete. Plus, many official sites were exclusively available on the native language of the hosting country (e.g. Chinese) and automatic translation tools were insufficient to obtain the required information. Thus, we decided to contact directly the community to complete the survey. The questions were sent to a web archive discussion list, published on the site of the Portuguese Web Archive (PWA) and disseminated through its communication channels (Twitter, Facebook, RSS). We obtained 27 answers. Then, we sent direct e-mails to the remaining web archives referenced by the International Internet Preservation Consortium ([Grotke, 2008](#)), National Library of Australia

3.1 Methodology

in its PADI (Preserving Access to Digital Information) page²⁰ and International Web Archiving Workshops²¹. We were able to establish contact and obtain direct answers from 33 web archiving initiatives. Finally, we distributed the collected data among the respondents for validation.

The methodology used in this research enabled web archivists to openly present information about their initiatives. For some situations, we had to actively interact with the respondents to obtain the desired information. We observed that terminology and language barriers led to different interpretations of the questions by the respondents, who involuntarily provided inaccurate answers. For instance, we assumed in the third question that each archived file was the result of a successful HTTP download (e.g. page, image or video), but some respondents interpreted it as the number of files created to store web contents in bulk, such as ARC files (Burner & Kahle, 1996). The posterior statistical analysis of the results enabled the detection of abnormal values and correction of these errors through interaction with the respondents. I believe that the adopted methodology enabled the extraction of more accurate information and valuable insights about web archiving initiatives world-wide, than a typical one-shot online survey with closed answers. However, the cost of processing the results for statistical analysis was significantly higher.

This survey uncovered that the publicly available information about web archives is frequently obsolete or nonexistent. However, the data collected and validated later enabled the creation of a Wikipedia page named *List of Web Archiving Initiatives*²², so that the published information could be collaboratively kept up-to-date. Since then, the web archiving community has been updating this information, making it a useful resource. In order to observe how web archiving changed since the first survey, in 2014 I conducted the same analysis on the data published in the Wikipedia page and compared it against the 2010 results. In case of doubt or lack of information, I consulted the official sites of the initiatives. Nevertheless, the data collection methodologies used in 2010 and 2014 were a little different, which could bias the comparison of results.

²⁰<http://www.nla.gov.au/padi>

²¹<http://iwaw.europarchive.org>

²²http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives

3. CHARACTERIZING WEB ARCHIVES

3.1.1 Comparison with other Surveys

After the 2010 survey, I learned of two other surveys on web archiving which had obtained related information, such as the access type provided by the initiatives and the technology used to support them. The first survey was conducted by the [Internet Memory Foundation \(2010\)](#) over European web archives in 2010, from now on referred to as the IMF survey. The second survey was published by the [NDSA Content Working Group \(2012\)](#) in 2012 and covered organizations of the USA involved or planning to archive content from the web. This survey is referred to from now on as the NDSA survey. In this chapter I analyze and compare the results of the surveys whenever possible, despite my surveys having covered world-wide web archiving initiatives, while the IMF survey focused just on initiatives from Europe and the NDSA survey on initiatives from the USA. Still, these two last surveys and my first survey took place between 2010 and 2012, which makes their results comparable in time.

3.2 Results

3.2.1 Initiatives

Table [3.1](#) shows general statistics of web archiving initiatives surveyed in 2010 and 2014. Tables [A.1](#) and [A.2](#) in Appendix A present the 42 web archiving initiatives identified across the world in 2010, ordered alphabetically by their hosting country (question 1). Web archiving initiatives are very heterogeneous in size and scope. For instance, the web archive (WA) of Čačak aims to preserve sites related to this Serbian city, while the Internet Archive has the objective of archiving the global web. The obtained results of 2010 show that 80% of the archives exclusively hold content related to their hosting country, region or institution. However, initiatives hosted in the USA, such as the Latin American WA, Internet Archive or the WA Pacific Islands, also preserve information related to foreign countries. The creation and operation of a web archive is complex and costly. The Internet Archive, Internet Memory Foundation and California Digital Library provide web archiving services (WAS) that can be independently operated by

3.2 Results

characteristics	2010	2014
total initiatives	42	68
countries hosting initiatives	26	33
total people (full-time)	112	108
total people (part-time)	166	197
total people	278	305
median people (full-time)	2.5	2
median people (part-time)	2	2
average people (full-time)	3.5	2.2
average people (part-time)	5	4

Table 3.1: General statistics of web archiving initiatives.

third-party archivists. The WAS are named Archive-It²³, ArchiveTheNet²⁴ and Web Archiving Service²⁵, respectively. These services enable focused archiving of web contents by organizations, such as universities or libraries, that otherwise could not manage their own archives. For instance, the Archive-It service is used by the North Carolina WA, the ArchiveTheNet is used by the UK Government WA and the Web Archiving Service by the University of Michigan WA.

I detected an increase in the number of web archiving initiatives, from 42 in 2010 to 68 in 2014. There are now 11 initiatives (16%) providing WAS that can be independently operated by third-party archivists to easily capture and preserve web content, against the previous 3 WAS offered in 2010. Of the 11 WAS, 6 operate in the USA, where most of them offer electronic discovery (ediscovery) services for enterprises, which are required by law since 2006 for the discovery of information in civil litigation or government investigations. At least 13 initiatives (19%) are contracting WAS. In 2010, this percentage was 16%.

Human Resources

The measurement of human resources engaged in web archiving activities was not straightforward (question 2). Most respondents could not provide an effort measurement in person-month. The presented reasons were that the teams were

²³<http://www.archive-it.org>

²⁴<http://archivethe.net>

²⁵<http://webarchives.cdlib.org>

3. CHARACTERIZING WEB ARCHIVES

too variable and some services were hired to third-party organizations out of their control. Instead, most of the respondents described their staff and hiring conditions. The obtained results of 2010 show that web archiving engaged at least 112 people in full-time and 166 in part-time. The total of 278 people that preserved and provided access to the past web since its inception contrasts with the resources invested to provide access to a snapshot of the current web. For instance, Google by itself had 24 400 full-time employees in 2010, from which 9 508 worked in research and development, and 2 768 in operations ([United States Securities and Exchange Commission, 2010](#)). The web archive teams are typically small, presenting a median staff of 2.5 people in full-time (average of 3.5) and 2 people in part-time (average of 5). The staff is mostly composed by librarians and information technology engineers. The results show that 11 initiatives (26%) did not have any person dedicated full-time. The effort of part-time workers is variable, for instance, at the Library of Congress they spent only a few hours a month. Most of the human resources were invested on data acquisition and quality control. The IMF survey corroborates that web archive teams are small, but the number of staff depends on the phase of the project. Its results show that 38% of fully operational initiatives count more than 5 full-time employees, while 67% that started a project count between 2 and 5 employees.

In 2014, the size of the teams continue to be highly variable, where initiatives have teams without any person working in full-time, such as the University of Texas at San Antonio WA, while other teams have 12 people working in full-time, such as the Internet Archive, or 80 people working in part-time, such as the Library of Congress. As shown in Table [3.1](#), in 2014 the web archiving initiatives have in total 108 people working in full-time and 197 in part-time. There was an increase from 278 to 305 people working in this area. The teams continue to be mostly small, having a median staff of 2 people in full-time (average of 2.2) and 2 people in part-time (average of 4). There are 3 initiatives that do not have any person dedicated full-time, against the 11 of 2010. Despite the large increase of the number of initiatives, the total number of people working on them increased only slightly, which led to a decrease in the median and average team size.

Location

Figure 3.1(a) presents the countries that hosted web archiving initiatives in 2010. The 42 initiatives were spread across 26 countries. There were 23 initiatives hosted in Europe, 10 in North America, 6 in Asia and 3 in Oceania. Half of the initiatives were hosted in countries belonging to the Organisation for Economic Co-operation and Development (OECD). From the 34 countries that belong to the OECD, 21 (62%) hosted at least one web archiving initiative, which is an indicator of the importance of web archiving in developed countries. Most of the countries hosted one (74%) or two initiatives (22%). The only country that hosted more than two was the USA with a total of 9 initiatives. Although being part of a country, initiatives like the Tasmanian WA (Australia), North Carolina WA (USA) or Digital Heritage Catalonia (Spain) were hosted at autonomous states and aimed at preserving regional content. When comparing the number and location of initiatives with other surveys, I detected that many were missing. The IMF survey found 41 European initiatives fully operational, while I found 23. The NDSA survey found 49 initiatives in the USA, but I found only 9.

Figure 3.1(b) presents the location of all countries hosting web archiving initiatives in 2014. The 68 web archiving initiatives are spread by 33 countries from which 21 countries only have one initiative and 3 countries have 2 initiatives. In 2010 there were only 26 countries hosting web archiving initiatives, which shows a growing awareness of the importance of web archiving all over the world. The USA continues to be the country with the most initiatives, increasing from 9 in 2010 to 19 in 2014. The second country with most initiatives is France, with 5 initiatives. Germany and Switzerland share the third place with 4 initiatives each. The distribution of the initiatives over the world is 38 in Europe (previously 23), 22 in North America (previously 10), 8 in Asia (previously 6), 3 in Oceania (equal) and 1 in Africa (previously 0). There were increases in almost all continents, especially in Europe and in North America. Africa received its first initiative hosted in Egypt, while South America does not have any yet.

3. CHARACTERIZING WEB ARCHIVES

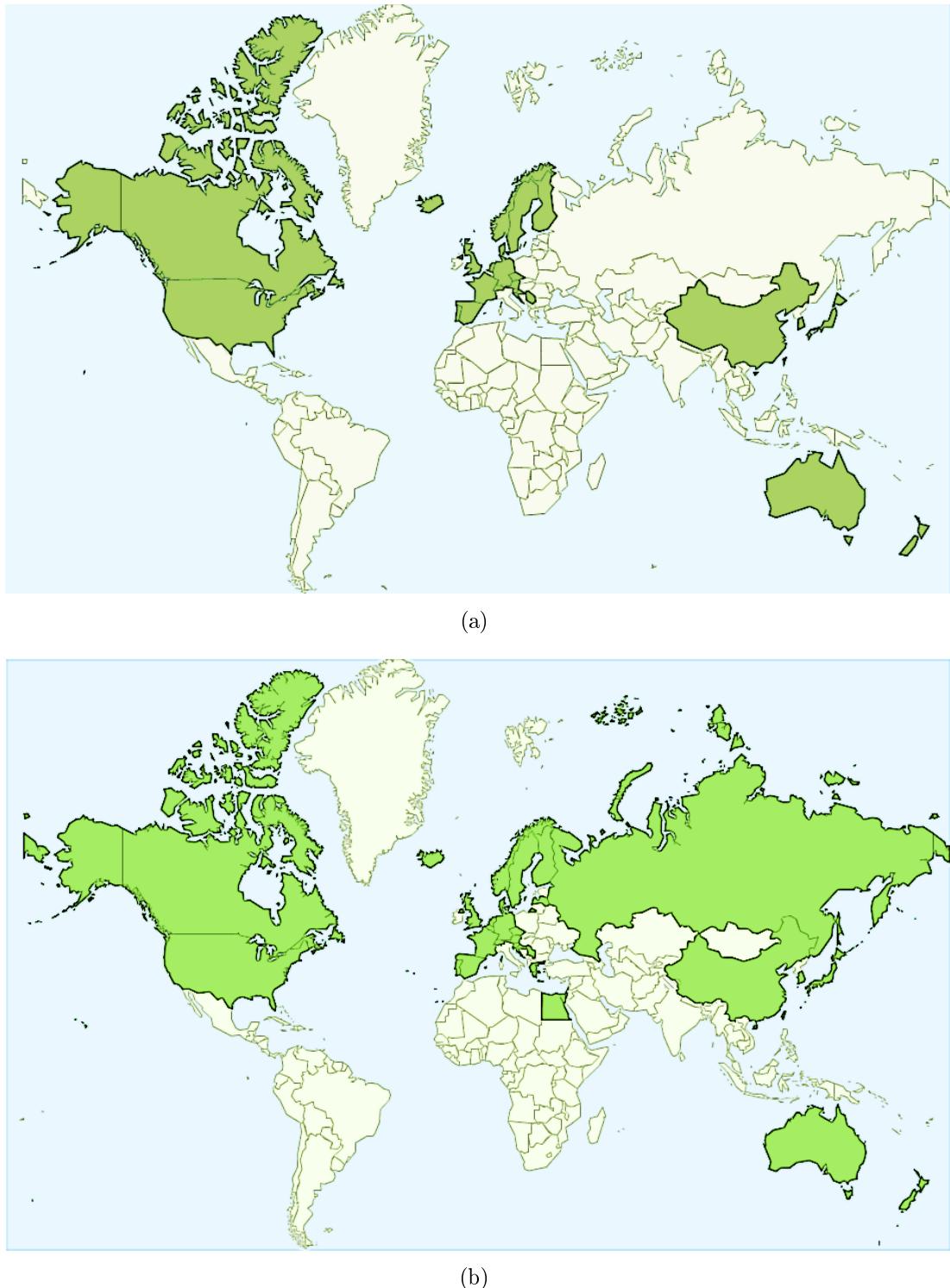


Figure 3.1: Countries hosting web archiving initiatives in (a) 2010 and (b) 2014 (in green).

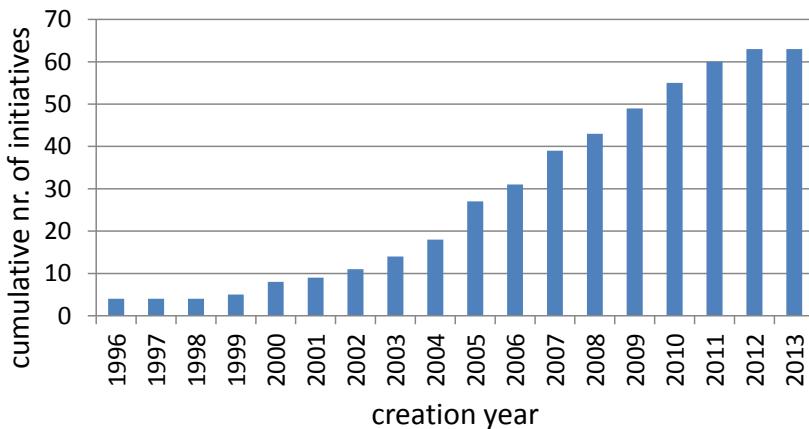


Figure 3.2: Cumulative number of initiatives created per year.

Growth

Figure 3.2 displays the evolution of the number of web archiving initiatives created per year, including the new initiatives recorded on the Wikipedia page. The first initiatives were created in 1996: the Internet Archive founded by Brewster Kahle, the Australia WA (Pandora) and Tasmanian WA from Australia, and the Kulturarw3 from Sweden. There was a small growth from 4 initiatives in 1996 to 14 initiatives in 2003, which represents an average of 1.8 new initiatives per year. After 2003, many new initiatives appeared to solve the web ephemerality problem. For instance, in 2005 and 2007, 9 and 8 initiatives were created, respectively. There was an average growth of 5.4 initiatives per year from 2004 to 2012. There is no information of new initiatives created in 2013. One possible explanation for the significant and constant growth since 2003 was the concern raised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarding the preservation of the digital heritage (UNESCO, 2003). The NDSA survey also shows a constant growth, especially between 2007 and 2011, when there was a great increase of initiatives mainly due to universities starting their web archiving programs. Universities created 29 (out of 49) initiatives in these 5 years, which indicates an emergent awareness in the academic community of the importance of preserving web content.

3. CHARACTERIZING WEB ARCHIVES

3.2.2 Archived Data

Selection Policy

Since the resources are scarce and not all the web can be preserved, the selection policy of most web archiving initiatives is to preserve the most relevant parts of the web from their own perspective. In the survey of 2010, all web archives selected specific sites for archiving. This selection is determined by multiple factors, such as consent by the authors or relevance for inclusion in thematic collections (e.g. elections or natural disasters). However, 80% of the web archives exclusively held content related to their hosting country, region or institution. Of the 42 initiatives, 11 (26%) also performed broad crawls of the web, including all sites hosted under a given domain name or geographical location. The IMF survey reported that 23% of European web archives run domain crawls, while 71% performed thematic or selective crawls. The NDSA survey reported that all USA initiatives archived web content from their own institution, as well as content from other organizations or individuals for future research.

In 2014, at least 45 initiatives (66%) perform selective crawls and 20 (29%) TLD or broad crawls of the web. Almost all initiatives exclusively hold content related to their hosting country, region or institution. There are three initiatives that archive TLD of other countries besides their own. The Internet Archive and the Internet Memory Foundation share a vision to preserve web content from all over the world. The PWA preserves content from four countries that speak native Portuguese.

Size

Figure 3.3 presents the distribution of the size of archived collections measured in total volume of data and number of contents. Notice that one HTML page containing three embedded images results in the archive of four contents. Selective web archiving is frequently focused on preserving individual sites. Thus, although the number of archived sites could also be an interesting metric, the size of web sites significantly varies and the number of archived sites by itself is not descriptive of the volume of archived data. Therefore, I decided not to include this metric to simplify the questionnaire.

3.2 Results

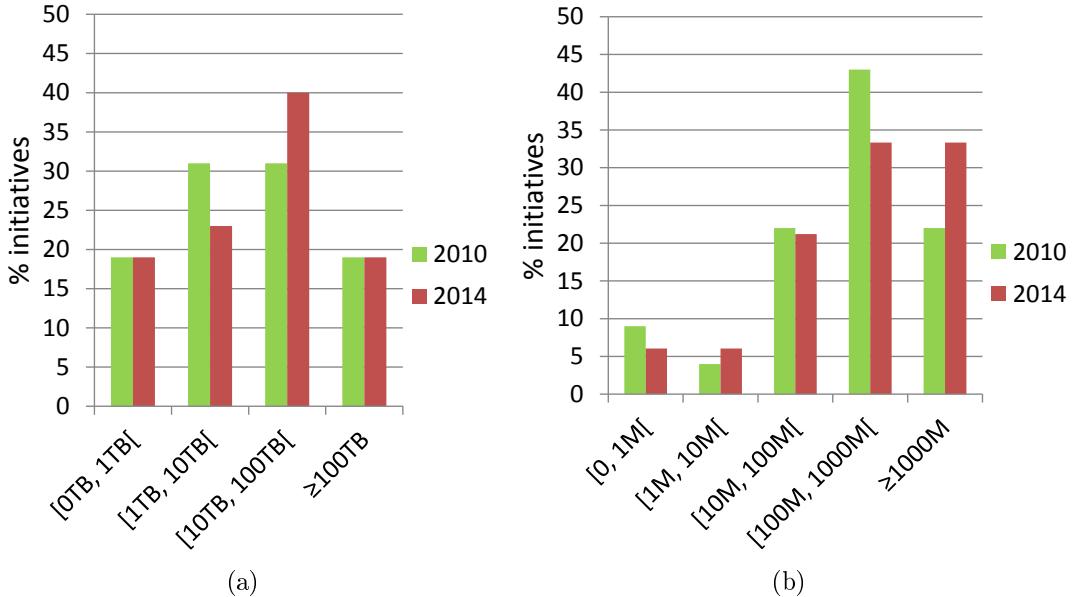


Figure 3.3: Size of archived collections measured in: (a) volume of data (Terabytes) and (b) number of contents (e.g. images, pages, videos).

The results of 2010 show that 50% of the collections are smaller than 10 TB and 78% have less than 1 000 million contents. The volume of data in replicas to ensure preservation was not considered in this measurement. The average content size was 46 KB and ranged between 14.2 KB and 119.4 KB. There are several reasons for this difference. Some web archives are focused on specific contents that are typically large, such as video, PDF documents or images. Web archives also use different formats for archiving web data that may contain additional meta-data or use compression. Another reason is that the size of contents tends to grow ([Miranda & Gomes, 2009a](#)). Therefore, older archived contents tend to be smaller than recent ones.

Web archives world-wide preserved from 1996 to 2010 a total of 181 978 million contents (6.6 PB). The Internet Archive by itself held 150 000 million contents (5.5 PB). In 2014, all initiatives have archived together at least 534 604 million contents, which sums around 17 PB of data. This represents an increase from 2010 to 2014 of 294% on contents and 258% on volume of data. The Internet Archive continues to be by far the web archive with the largest collection with

3. CHARACTERIZING WEB ARCHIVES

376 000 million contents. The information of its volume of data was not available in the Wikipedia page. Hence, I extrapolated from the 2010 results and estimated 13.8 PB of data. The size of the current web cannot be accurately determined. However, in 2008 Google announced that one single snapshot of the web comprised 1 trillion unique URLs (10^{12}) (Google Inc., 2008). Notice that this number refers only to web pages and does not include contents, such as images or videos, that are also preserved by web archives. The obtained results show that the amount of archived data is small in comparison with the volume of data that is being published on the web.

There was an increase of initiatives with collections between 10 TB and 100 TB in detriment of collections between 1 TB and 10 TB. While in 2010, 50% of the initiatives preserved collections smaller than 10 TB and 31% preserved collections between 10 TB and 100 TB, in 2014 these percentages were 42% and 40%, respectively. The percentage of initiatives with collections larger than 100 TB continues to be 19%. In accordance with this finding, the percentage of initiatives with collections between 100 million and 1 000 million contents decreased from 43% to 33%, mostly because the percentage of initiatives with collections with more than 1 000 million contents increased from 22% to 33%. The main conclusion is that the archived collections grew significantly in volume of data and number of contents.

Formats

Figure 3.4 presents the evolution of the distribution of file formats used to store archived content. The ARC format defined by the Internet Archive was the *de facto* standard in 2010 (Burner & Kahle, 1996). In 2009, the WARC format was published by the International Organization for Standardization (ISO) as the official standard format for archiving web contents (ISO 28500:2009, 2009) and it was exclusively used by 10% of the initiatives in 2010. The ARC and WARC formats were dominant, being used by 54% of the initiatives. I found that there was a decrease, from 26% in 2010 to 13% in 2014, of initiatives using exclusively the ARC format. These initiatives likely changed to the WARC format that increased 3 percentage points and the ARC/WARC formats that also increased 3

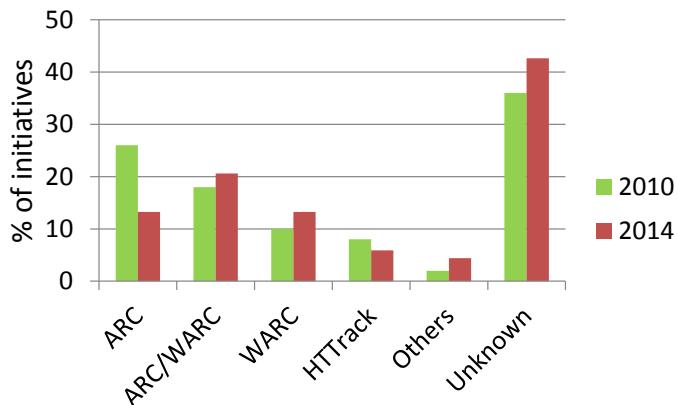


Figure 3.4: Usage of file formats to store web contents.

percentage points. The ARC and WARC formats continue to be by far the most predominant, being used today by 47% of web archiving initiatives against the 54% in 2010. There are only 10% of initiatives using other file formats, such as the HTTrack format. Still, 43% of the initiatives did not reported the adopted format in the Wikipedia page.

The usage of standard formats for web archiving facilitates the collaborative creation of tools, such as search engines or replication mechanisms, to process the archived data. Besides historical reasons, the widespread of the ARC/WARC formats was motivated by the Archive-Access project, which freely provides open-source tools to process this type of files ([IIPC, 2009](#)).

3.2.3 Access and Technologies

Access Type

Figure [3.5](#) presents the types of access provided by the initiatives over their collections in 2010 and 2014. The obtained results of 2010, show that 89% of the initiatives support access to the multiple versions of a given URL published over time, 79% enable searching through meta-data and 67% provide full-text search over archived contents. These results differ from the IMF survey, which reported 68%, 65% and 70% of European initiatives supporting URL, meta-data and full-text search, respectively. The percentage of European web archives offering URL

3. CHARACTERIZING WEB ARCHIVES

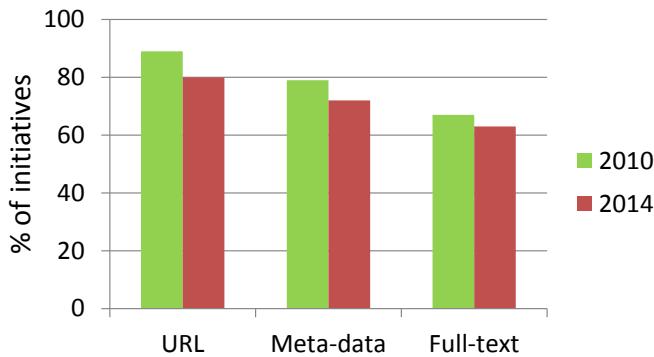


Figure 3.5: Access type provided by web archives.

and meta-data search are significantly lower, but slightly higher in full-text search. The NDSA survey shows similar results. The URL search (62%) and full-text search (67%) are also the most provided data access types to users. The NDSA survey reported other access types, namely meta-data search (51%) and browsing by URL (48%) and title (55%).

The results of 2014 are almost the same with a small relative decrease in all access types. The most predominant access type is the search by URL, then the search by meta-data and last, by full-text search. There were 2 initiatives that provided full-text, but only to a part of their collections (one 30% and the other 15%). The DILIMAG initiative reported the lack of resources to implement full-text search.

Access Restrictions

In 2010, some initiatives held the copyright of the archived contents (e.g. German Bundestag, UK WA, Canada WA) or explicitly required the consent of the authors before archiving (e.g. UK WA, OASIS). The Tasmanian WA operated since its inception under the assumption that web sites fall within the definition of book. Thus, no permission to capture from publishers was required. The Internet Archive and the PWA proactively archive and provide access to contents, but remove access on-demand. On the other hand, for 16 initiatives (38%) the access to collections was somehow restricted. The Library of Congress, WebArchiv

3.2 Results

and Australia WA provided public online access to part of their collections. Netarkivet.dk provided online access on-demand only for research purposes. The Finnish WA provided online access to meta-data, but not to archived contents. BnF, Web@rchive and Preservation .ES granted access exclusively through special rooms on their facilities. The IMF survey found that 50% of the European initiatives perform web archiving protected by a law enacted or passed. Regarding the policy for accessing archived data, 41% of the initiatives provide access for everyone, 28% online access with restrictions, 18% on-site access for anyone, 21% on-site access with restrictions and 21% do not provide any access of their contents. Maintaining the accessibility level of the original information is mandatory to make web archives useful for citizens. If a content is publicly available on the current web, it should continue to be publicly available when it becomes a historical content. However, this policy collides with national legislations that restrict access or even inhibit proactive web archiving. The web broke economical and geographical barriers to information, but legislations are raising them against historical content. It is economical unattainable for most people to travel, possibly to a foreign country, to investigate if an information published in the past exists in a web archive.

The information available on the Wikipedia page about the access restrictions is not sufficient for a statistical analysis. Still, some initiatives recorded their restrictions. The WebArchiv of Czech Republic provides unlimited access only from public terminals in the National Library. The Chinese WA and the Web@rchive of Austria provide access to content in their National Libraries. The Finnish WA also provides on-site access to contents. For the Netarkivet.dk of Denmark, the online access is granted only to researchers and the BnF Web Legal Deposit of France grants access only to authorized users.

Technology

Figure 3.6 depicts the technologies being used by the initiatives that manage their own systems. In 2010, the Archive-Access tools were dominant (62%), including the Heritrix, NutchWAX and Wayback projects, that support content harvesting, full-text and URL search, respectively. However, respondents frequently

3. CHARACTERIZING WEB ARCHIVES

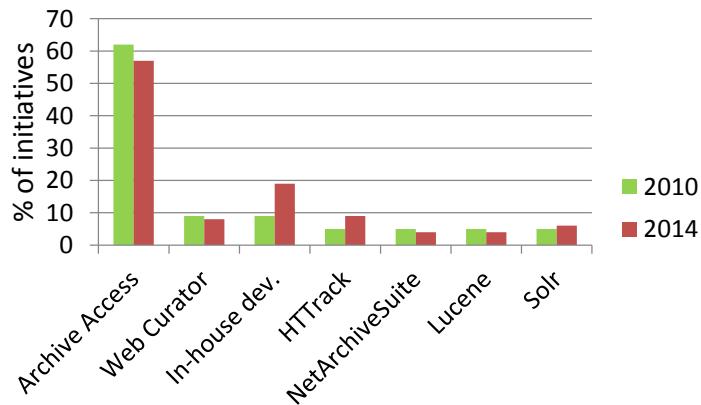


Figure 3.6: Technologies used by web archives.

mentioned that full-text search was hard to implement and that the performance of NutchWAX was unsatisfactory, being one reason for the partial indexing of their collections. Nonetheless, in 2010, NutchWAX supported full-text search for the Finnish WA (148 million), Canada WA (170 million), Digital Heritage of Catalonia (200 million), California Digital Library (216 million) and BnF (15% of a collection of 200TB). It was estimated that the largest web search engine is Google and that it indexes 38 000 million pages ([Kunder, 2011](#)). Creating a search engine over the archived data (534 604 million contents), would imply indexing 14 times more data. The IMF survey indicates that 80% of European initiatives use Heritrix to crawl web content. The NDSA survey reported that 76% of the USA initiatives were using Wayback to provide data access via URL search. There is no mention to full-text search tools. However, since 60% of these initiatives were using WAS, especially Archive-It, they were likely using the full-text provided by NutchWAX.

Despite the increase from 3 in 2010 to 11 in 2014 of web archive services (WAS), the number of initiatives that used WAS increased just 3 percentage points, from 16% to 19%. The Archive-It is the service most used, totaling 7 initiatives. There was an increase from 10% to 19% of initiatives doing some in-house development. This software was mostly developed by WAS, such as the Hanzo Archives' access tools, or curation tools developed by libraries, such as the DigiBoard of the Library of Congress Web Archives. These increases con-

tributed to the decrease of the use of Archive-Access tools, which include Heritrix, NutchWAX and Wayback projects. Still, the Archive-Access tools continue to be predominant, with 57% of the initiatives using at least one of these tools in 2014, against the 62% in 2010. Lucene and Solr together continue to be used by 10% of the initiatives.

3.3 Summary

Web archiving has been gaining interest and recognition from modern societies around the world. Still, there is a lack of knowledge in the research community about the most recent developments in web archiving and the existing initiatives. This chapter provides an updated and global overview on these issues.

Based on two conducted surveys, I observed that web archiving initiatives are typically hosted on developed countries, but we can find them spread all over the world in almost every continent. Web archives are generally composed by small teams that mainly work on the acquisition and curation of data. Almost all initiatives exclusively hold content related to their hosting country, region or institution, which stresses the need for each country to finance at least one initiative at national level.

Web archiving initiatives have been in existence since 1996 and their number has been growing since then. Particularly, from 2010 to 2014 there was a large increase in the number of initiatives, hosting countries, number of contents and volume of archived data. Currently, web archiving initiatives hold 17 PB (534 604 million contents), which shows a growing awareness of the importance of web archiving all over the world and a continued effort of the community in mitigating the web ephemerality problem.

On the other hand, despite the social and economic impact of losing the information that is being exclusively published on the web, the obtained results show that the human resources invested in web archiving are still scarce and the size of teams are even decreasing. The lack of resources will probably originate a historical void in the future about our current times. The results already show that only a small part of the web has been preserved.

3. CHARACTERIZING WEB ARCHIVES

Most web archiving initiatives use Lucene-based solutions to support full-text search, such as NutchWAX or Solr. Other surveys also showed that the predominant types of access to archived content are the URL and full-text search, usually supported by Wayback and NutchWAX, respectively. To the best of my knowledge, these are the most advanced IR technology currently used in web archives. However, the respondents of the surveys mentioned that the existing technology provides unsatisfactory search results and full-text is hard to implement. With the fast growth of archived data, this problem only tends to aggravate. Hence, efficient and effective search mechanisms are required to access the massive data already in web archives.

Chapter 4

Characterizing Information Needs

The previous chapter provided an updated and global overview of web archiving initiatives, surveying the type of access provided to their users and the technologies used for that end. The main conclusion is that web archiving information retrieval (WAIR) technology is in its early stages, being essentially based on commonly used web search technology that does not account for the specificities of WAIR. For instance, the time dimension present in the web data archived over the years is completely ignored when searching. This leads to questioning whether traditional web search technology can effectively support the information needs of web archive users, which in turn leads to another unanswered question: what are the information needs of web archive users?

Understanding what users need is the first step to the success of any information technology (IT) system. However, sometimes users only have a vague idea of what they want the system to do. I faced this problem when I started developing the access functionalities for the Portuguese Web Archive (PWA). People had a great difficulty in suggesting anything without seeing the system working. Showing other national web archives helped them to understand the concept of the project. Nevertheless, without real information needs over past documents and subjects that they could remember and explore, the responses remained too vague. The only feedback I received was on whether functionalities of other systems were a good or bad idea. For instance, everyone agreed that full-text and URL search over web archive collections were good ideas and I implemented

4. CHARACTERIZING INFORMATION NEEDS

them with the state-of-the-art WAIR technology described in the previous chapter. However, full-text and URL search are not an end in themselves. They are mechanisms to obtain specific information, such as details on a subject written in the past.

With the public release of the PWA experimental version in 2010, it was finally possible to collect valuable feedback from users and enrich our understanding of their information needs, i.e. the goals/intents behind their queries. Identifying the users' underlying goal is important for three main reasons. First, it points out directions for developing technology that can better satisfy web archive users. Different intents may require different solutions. Second, it enables us to provide full-text search results tailored toward the user goal. Studies over web search engines clearly show that tuning the ranking model for that goal can significantly improve results ([Geng et al., 2008](#); [Kang & Kim, 2003](#)). I expect the same behavior in full-text search over web archive collections. Third, it structures the elaboration of a representative WAIR evaluation, which is essential to compare approaches and measure progress.

In this chapter, I draw the first profile of why and what users search. I used three methods to collect quantitative and qualitative data from users, namely, search log mining, an online questionnaire to be answered by the users while they were searching and a laboratory study. All findings and their implications on the development of future web archives were discussed.

The main contributions of this chapter are:

1. the first characterization of web archive users about their information needs and searched topics;
2. an analysis of whether the web search engine technology currently used in web archives can effectively support the information needs of web archive users;
3. the identification of functionalities that users would like to see implemented, but are not currently supported.

The remainder of this chapter is organized as follows. Section [4.1](#) describes the user interface of the PWA. Section [4.2](#) details the methodology employed

4.1 The PWA User Interface



Figure 4.1: Search interface after a full-text search.

in the present study. The results obtained from the conducted experiments are presented in Section 4.3. The chapter ends with a summary in Section 4.4.

4.1 The PWA User Interface

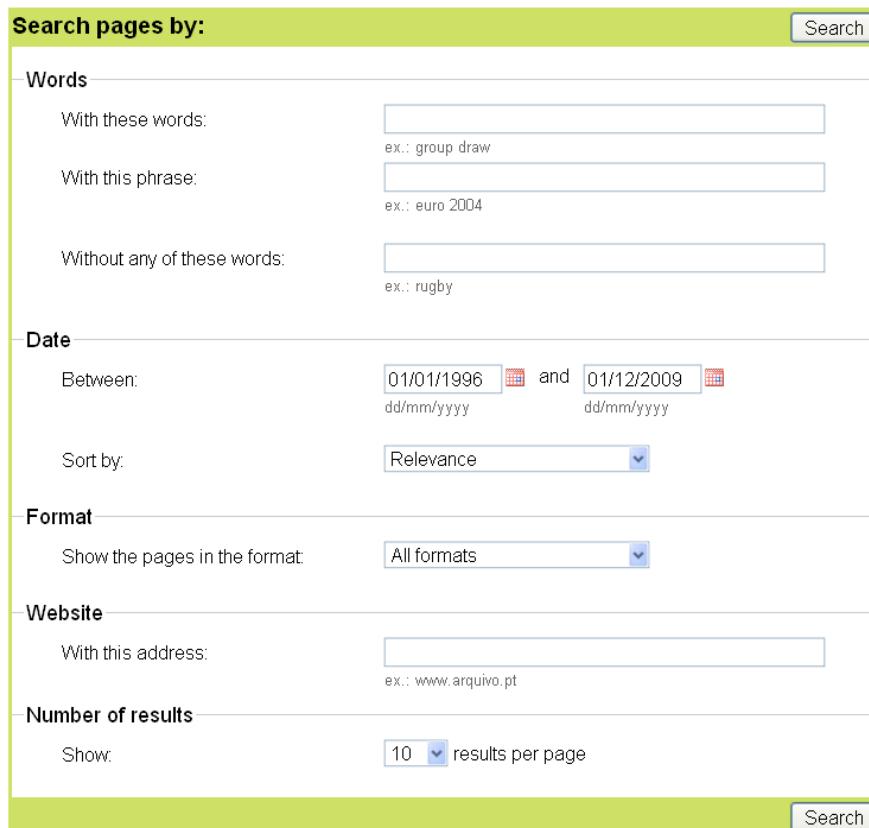
The experimental version of the PWA is a public service since January 2010. It is accessible from <http://archive.pt>, providing both Portuguese and English language interfaces. In 2010, it contained nearly 150 million documents searchable by full-text and URL via an interface complemented with a date range filter to narrow the results to a time period. Other web archives, such as Padicat²⁶ and Pandora²⁷, provide similar access. However, in the PWA user interface both full-text and URL queries are submitted from the same text box as in Google omni bar. The PWA interprets the type of query and presents the results accordingly. The archived documents ranged between 1996 and 2009.

The interaction with the users and the layout of the results is similar to web search engines, such as Google. In a typical session, a user can submit a full-text query and receive a search engine results page (SERP) containing a list of 10 results matching the query. Figure 4.1 illustrates this case. Each result includes the title of the web page and its crawled date, a snippet of text containing the query terms and the URL. The user can then click on the results to see and navigate in the web pages as they were in the past. If the desired information is

²⁶ <http://www.padicat.cat>

²⁷ <http://pandora.nla.gov.au>

4. CHARACTERIZING INFORMATION NEEDS



The image shows a screenshot of an advanced search interface. At the top, there is a green header bar with the text "Search pages by:" on the left and a "Search" button on the right. Below this, the interface is divided into several sections:

- Words**:
 - With these words: ex.: group draw
 - With this phrase: ex.: euro 2004
 - Without any of these words: ex.: rugby
- Date**:
 - Between: and
 - Sort by:
- Format**:
 - Show the pages in the format:
- Website**:
 - With this address: ex.: www.arquivo.pt
- Number of results**:
 - Show: results per page

At the bottom right of the form area is another "Search" button.

Figure 4.2: Advanced search interface.

not found, the user can repeatedly modify and resubmit the query. In addition, the user can click on the navigation links to explore other SERPs or use the advanced search interface to restrict the query with advanced search operators. Figure 4.2 shows the available operators. It is possible to restrict the result set by format and sort it by one of the three criteria: relevance, newest first or oldest first. These advanced operators can also be added to the query directly in the text box.

The PWA user interface has some specificities. First, the text box is complemented with a date range filter to narrow the results to a time period. Second, each result has an associated link to see all versions throughout time of the respective URL. When clicked, the PWA presents the same search engine versions page (SEVP) as when a user submits that URL on the text box. Figure 4.3

4.2 Methodology



Figure 4.3: Search interface after a URL search.

depicts a SEVP with a table, where each column contains all the versions of a year sorted by date. The user can then click on any version to see it as it was on that date.

4.2 Methodology

User study methods can be classified in three dimensions:

1. client-side ([Fox et al., 2005](#)) or server-side ([Jansen & Spink, 2006](#)) log analysis of the users interactions with the system;
2. surveys based on interviews ([Teevan et al., 2004](#)) or questionnaires ([Aula et al., 2005](#)) conducted on users;
3. experiments with users in a laboratory ([Aula et al., 2010](#)) or in their natural environment (in situ) ([Kellar et al., 2007](#)).

All methods have pros and cons, so I experimented one of each group as complementary ways of analysis. Next, I synthesize the chosen methods employed on the PWA in 2010.

4.2.1 Data Collecting Methods

Search logs capture a large and varied amount of interactions between users and IR systems. This enables the generalization of strong relationships between data.

4. CHARACTERIZING INFORMATION NEEDS

Another advantage of this method is its unobtrusiveness, i.e. non interference in the users' normal behavior. Most users are not even aware during a search that their interactions are being logged. On the other hand, search logs are limited to what can be registered. They ignore the contextual information about users, such as their demographic characteristics, the motivations that lead them to start searching, and their degree of satisfaction with the system.

When analyzing logs, contextual information must be collected using other methods. A possibility is to ask users directly, displaying online interactive questionnaires while the users are performing or concluding a critical function. This allows them to enter fresh opinions on the system usability and functionality. However, interactive questionnaires force users to engage in additional activities beyond their normal searching behavior, where the benefits are not always apparent. This interference on search can bias results. It is challenging to define a simple and fast mechanism that encourages users to provide feedback without significantly disrupting their main task.

A significant part of behavioral information is not registered neither in logs, nor described by the users in questionnaires. This information can be only collected through observation. Laboratory studies involve observing users in a controlled setting, conducting searches in response to a simulated information need. Specialized equipments, such as video/screen capture or eye-trackers, are used to gather different types of data for analysis. As result, this method provides the best insight on the system usability and user satisfaction. As disadvantage, the time spent observing the participants and the costs of acquiring specialized equipments often lead researchers to reduce the users sample to a size smaller than required to obtain statistically significant results. Another problem is their intrusiveness in the search process. The fact that the users are aware of being observed can affect their normal behavior.

Potentially valuable datasets include large and diversified data to generalize results, and rich data to explain them. Figure 4.4 represents the relation between the three chosen data collecting methods. The y-axis represents the richness of the collected data, where the richest is obtained by the laboratory studies. The x-axis represents the degree of generalization of the results in Figure 4.4(a) and the

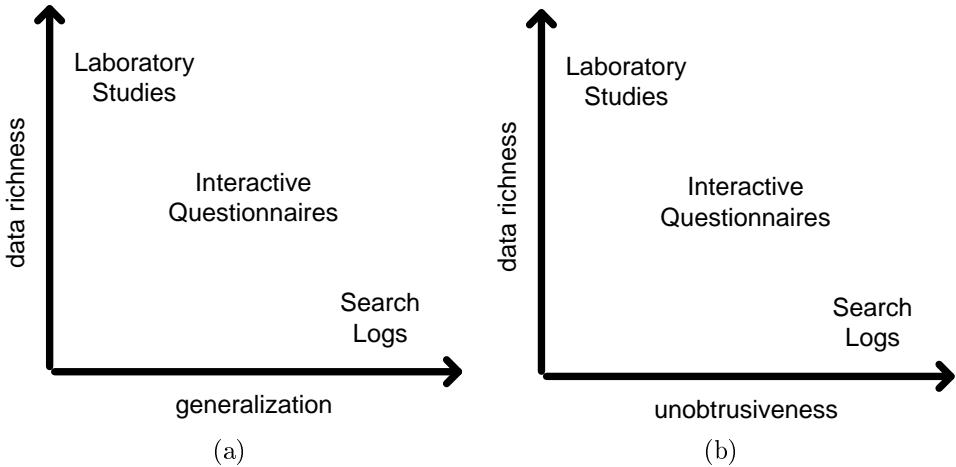


Figure 4.4: Data collecting methods used.

degree of unobtrusiveness in Figure 4.4(b), where search logs surpass all others. The next section details the experiments.

4.2.2 Experiment #1: Search Logs

Procedure

I started by preparing the log fields for analysis through a series of data cleansing steps. All incomplete entries, empty queries and sessions without any query were discarded. Internal queries submitted by the PWA monitoring system, the queries by example displayed on the PWA entry page and sessions conducted by clients identified as web crawlers have also been excluded. Additionally, sessions with more than 100 queries were removed too. Sessions with many queries were likely to come from web crawlers and only queries submitted by human users were considered. This cutoff value of 100 was used in some other studies, thus enabling a more direct comparison with my results ([Jansen & Spink, 2005](#)).

A proper delimitation of a session is important, since a session represents the set of interactions that belong to the same user when attempting to satisfy an information need. Like in most studies that analyze search logs, I used the users' IP address and session identifier to delimit sessions ([Jansen & Spink, 2006](#)). I also used a time interval t of inactivity. Two consecutive interactions are included on

4. CHARACTERIZING INFORMATION NEEDS

different sessions if they have an inactivity between them of at least t . Without this interval, sessions could have several days of duration, which would hardly represent the reality. Studies diverge on the choice of this interval, from 5 to 120 minutes ([Jansen & Spink, 2006](#)), while others argue that no time boundary is effective in segmenting sessions ([Jones & Klinkner, 2008](#)). I have selected a 30 minute interval, because 98% of the PWA sessions were shorter and it is the session default timeout on most web applications. This interval also produced results close to the ones of Support Vector Machine (SVM) classifiers used for delimiting sessions ([Radlinski & Joachims, 2005](#)).

After delimiting the sessions, I followed the [Rose & Levinson \(2004\)](#) idea of developing a software tool for assisting session classification. Using this tool, the queries and clicks of 400 sessions were manually analyzed to infer their information needs and addressed topics. Needs and topics were target of discussion and brainstorming, followed by an iterative process of refinement. It is necessary to clarify that the needs are inferred from the sessions without certitude. However, the sessions were individually classified by two evaluators and then their discrepancies resolved. The agreement between the two evaluators measured with the Cohen's kappa coefficient was 0.71, which is conventionally interpreted as a substantial agreement ([Landis & Koch, 1977](#)). The taxonomy of the topics was based on the study of [Jansen & Spink \(2006\)](#).

Participants

The PWA contains all kind of contents from the Portuguese web. Moreover, the PWA is a public service, so I believe that the logs contain searches from all kind of users, with a variety of interests, ages and professions. These logs are related to a period from May 17 to July 2, 2010.

The log data was never used to match a real identity. However, the location of the users' IP addresses was checked. I counted 81% of PWA users with IP addresses assigned to Portugal and near 94% of the interactions were submitted through the Portuguese language user interface. This strongly indicates that users were mostly Portuguese.

4.2 Methodology

Thank you for helping us improve our service. Your answers are confidential.

Which of the following phrases describe best what you were doing?

- * Seeing how a web page or site, that I know, was in the past (e.g. my homepage).
- * Collecting information about a subject written in the past (e.g. Iraq war).
- * Downloading an old file (e.g. music, video, image or software).
- * Recovering a web page or site that disappeared (e.g. to recover my Blog).
- * Seeing the evolution over time of a web page or site (e.g. the Google.pt page).
- * Seeing the evolution over time of the popularity of a subject (e.g. crisis).
- * Other:

Were you searching between specific dates (e.g. between 2000 and 2002)?

- * Yes
- * No

What other functionalities would you like our service to offer?

Give examples of how our service could help in your profession or daily activities:

Suggestions and critics:

Figure 4.5: Survey about the search of the Portuguese Web Archive.

4.2.3 Experiment #2: Interactive Questionnaire

Procedure

My goal was to receive responses from real information needs, motivated by the users, instead of asking them to imagine a scenario that could be handled using the web archive. Hence, my solution was to invite users to participate in an online questionnaire while they were searching. The invitation appeared in a form of a short message, placed close to the top right corner of the results page. Figure 4.1 shows this message: *Help us improve! It only takes 30s.*

The questionnaire presented in Figure 4.5, was designed based on existing guidelines described by Jansen *et al.* (2008b). It was implemented using the Google Forms framework²⁸, with some changes to attach the session identifier to the responses sent by each user. The questionnaire has a very short introduction on the top, thanking the participants and guaranteeing the confidentiality of their responses. It was followed by five questions, two of multiple-choice and three open-ended. The first question intends to identify the user's information need from those I suggest or new ones that I did not envision. The second focuses in determining if the need is restricted to a specific date range. The third asks

²⁸<http://docs.google.com>

4. CHARACTERIZING INFORMATION NEEDS

for functionalities that the user would like to see implemented. The fourth tries to get user-cases where the web archive could help in the user's profession or daily activities. The fifth is a generic question for suggestions and critics. I chose to restrict the number of questions to five, without demographic or experience related questions, because the participation rate on this type of experiments tends to be low. Increasing the number of questions, especially open-questions, would further reduce this rate.

Two pre-evaluation studies with five users each were performed to verify if all the questions were clearly understood. The studies were also an opportunity to detect problems and refine the questionnaire. To control the submitted data, I manually validated all responses. To guarantee that the same user had not submitted the questionnaire multiple times, I checked the users' IP addresses and session identifiers.

Participants

Of the six users that opened this questionnaire through the search interface, no one answered it. This indicates problems in the design adopted to captivate users and in the questionnaire itself. I detected that users contacted via email spent between 1 and 4 minutes from the time they opened the questionnaire until submission. These times seem prohibitive to receive a large number of answers. However, according to [Eysenbach \(2004\)](#), it is not unusual to have view rates of web-based questionnaires of less than 0.1%.

Due to lack of responses, I asked people to try the PWA and then to answer the questionnaire. This request was disseminated through the social networks associated to the project, Facebook and Twitter, and via email to acquaintances. As result, 21 participants responded to the online questionnaire, from the 75 that opened its URL. This means a participation rate of 28%. All 21 were recruited via email, which can bias results. I think that most people that came through Twitter and Facebook, which were 60%, only saw the questionnaire out of curiosity, since some of the followers work on similar projects. From the 21 responses, 2 were rejected because they were empty. This gives the questionnaire a completion rate of 90%. The answers were collected from June 18 to July 2, 2010.

4.2 Methodology

4.2.4 Experiment #3: Laboratory Study

Procedure

The experiment was conducted by the LaSIGE Human-Computer Interaction and Multimedia Research Team²⁹ on participants individually. Six steps were followed. First, an introduction of the project was presented and then the goal of the study explained. Second, a pre-questionnaire was provided to the participants to gather their demographics and experience background about computers and Internet. Third, a set of well defined tasks was presented with the goal to measure the usability of the PWA. I will not discuss these usability tests, since they are out of scope of this dissertation, but information about them can be found at ([Gomes et al., 2013](#)). The usability tests enabled the participants to become familiarized with the system.

On the fourth step, the participants were instructed to choose their own task based on their real information needs. It is known that allowing people to search for information that they are interested in stimulates their motivation and elicits realistic behavior ([Russell & Grimes, 2007](#)). Participants could stop whenever they wanted and were encouraged to search as they normally would at home or work. All interactions of the participants with the system were logged and also recorded on video with the Camtasia software³⁰. The participants were also observed by two researchers with minimal intrusion and without asking them to *think-aloud* about whatever they were looking at, doing and feeling. The goal was to achieve the closest to a normal searching behavior.

Fifth, after finishing the task, a post-questionnaire was given to each user containing the questions presented in Figure 4.5. The questionnaire was anonymous. Sixth, the researchers thanked the participant's help.

Participants

A total of 21 participants were recruited, 8 male and 13 female. Their ages ranged between 19 and 53 years, with an average of 30. The participants had a variety of

²⁹<http://hcim.di.fc.ul.pt/>

³⁰<http://www.techsmith.com/camtasia>

4. CHARACTERIZING INFORMATION NEEDS

Q1	Which of the following phrases describe better what you were doing?	Information Need	Exp. #1	Exp. #2	Exp. #3
1	Seeing how a web page or site, that I know, was in the past.	Navigational	47.70%	31.58%	47.62%
2	Seeing the evolution over time of a web page or site.	Navigational	9.21%	21.05%	33.33%
3	Collecting information about a subject written in the past.	Informational	37.83%	31.58%	14.29%
4	Downloading an old file.	Transactional	5.26%	10.53%	4.76%
5	Recovering a web page or site that disappeared.	Transactional	0%	5.26%	0%
6	Seeing the evolution over time of the popularity of a subject.	Informational	0%	0%	0%
7	Other	-	0%	0%	0%
Q2	Were you searching between specific dates (e.g. between 2000 and 2002)?				
1	Yes		15.79%	47.37%	9.52%
2	No		84.21%	52.63%	90.48%

Table 4.1: Distribution of information needs in the three experiments.

professions, interests and academic degrees. I believe that this diversity reflects the population of potential web archive users.

All participants presented a significant experience with computers, 17 had been using them for more than 10 years and the remaining 4 for more than 5 years. These participants also had been using the Internet for many years, 15 for more than 10 years, 5 for more than 5 years and 1 for more than 1 year. All the participants selected Google as the preferred search engine, using occasionally other search engines, such as Yahoo!.

4.3 Results

All information needs of web archive users focus on past data and match a class from the taxonomy proposed by [Broder \(2002\)](#), which is described in Section [2.3.2](#). As result, I aggregated options 1 and 2 from the first question (Q1) presented in Table [4.1](#). Both options refer a web page or site in mind, so they were considered navigational. Option 3 match the informational need, since users wanted to collect information about a subject, and options 4 and 5 the transactional, because both options focus on downloading or recovering old files. I will not discuss the other options, since the results show that they are not likely real or statistically significant in frequency.

4.3 Results

Q2	Were you searching between specific dates (e.g. between 2000 and 2002)?	Navigational Need	Informational Need	Transactional Need
1	Yes	86.95%	79.71%	93.75%
2	No	13.05%	20.29%	6.25%

Table 4.2: Distribution of sessions searched between dates per information need.

4.3.1 Experiment #1: Search Logs

The navigational needs were predominant, especially searching for a known page or site. This need led users to start 47.70% of the sessions. The other 9.21% of the navigational sessions, resulted from the exploration of several versions of web pages throughout the years. Sometimes, users expressed their navigational need in a very clear way through URL queries. It was counted 16.12% of navigational sessions containing only URL queries.

The second most frequent need was collecting information about a subject written in the past. A total of 37.83% of the sessions were initiated due to this informational need. Downloading an old file, i.e. the transactional need, originated 5.26% of the sessions. In this case, users searched mostly for images, but also searched for software, music, TV commercial jingles and BitTorrent files.

The PWA users only restricted queries by date range in 15.79% of the sessions, as shown in Table 4.1. Table 4.2 shows the distribution of sessions searched between dates per information need. In all information needs users mostly searched without narrowing searches by date range. Nevertheless, the informational needs were the most narrowed, occurring in 20.29% of the sessions.

Topics

The searched topics were separately classified for the navigational and informational needs. For the navigational, the sessions were classified according to the topics to which the sites are mostly about. Table 4.3 shows that sites about *Commerce* were searched in 28.31% of the sessions, while *Computers or Internet*, such as blogs, and *Education*, such as universities, were searched 14.46% each. For the informational needs, the sessions were classified according to the topics of the information searched. Table 4.4 shows that *People* was the most searched

4. CHARACTERIZING INFORMATION NEEDS

Topic	%
Commerce	28.31
Computers or Internet	14.46
Education	14.46
Government	8.43
Entertainment	7.23
Sciences	6.02
Society	5.42
Things	3.01
Health	2.41
Sports	1.81
Performing or Fine arts	1.81
Unknown or Other	1.20
People	1.20
Culture	1.20
Economy	0.60
Places	0.60
Employment	0.60
Sex or Pornography	0.60
Religion	0.60

Table 4.3: Topics searched per navigational needs.

Topic	%
People	36.52
Health	14.78
Entertainment	9.57
Things	6.96
Sports	6.09
Places	4.35
Sciences	4.35
Education	3.48
Travel	2.61
Economy	2.61
Commerce	2.61
Performing or Fine arts	2.61
Computers or Internet	1.74
Culture	0.87
Religion	0.87

Table 4.4: Topics searched per informational needs.

topic, corresponding to 36.52% of the sessions. 14.78% were about *Health* and 9.57% about *Entertainment*.

These results are in accordance with other results from users of web search engines in USA, Europe and Portugal ([Costa & Silva, 2010a](#); [Jansen & Spink, 2006](#)). For instance, *People, places or things* were the most searched topics in 2001 and 2002 by users of the AlltheWeb.com search engine, mostly from Germany and Norway. The same topics were also the most searched by users of AltaVista in 2002, mostly from the USA. The Excite users, which are also mostly from the USA, searched in 1999 and 2001 with a predominance of topics about *Commerce, Travel, Employment or Economy*. This same category of topics was the most searched by Portuguese users in 2003 and 2004. The categories of *Commerce* and *People* are at the top 3 most searched by users of web search engines and users of the PWA.

However, the most frequent queries are different between users of the two type of systems. Figure 4.6(a) displays a tag cloud of the search queries submitted to the PWA in 2010. We can see that the most frequent queries are names of Portuguese politicians at the time, such as the prime minister *José Sócrates* and the president of the republic *Cavaco Silva*. These queries are not present in the most frequent queries submitted to the Portuguese web search engine Tumba! in 2003 and 2004 (Costa & Silva, 2010a). Figure 4.6(b) displays a tag cloud of the search queries submitted to Tumba!. *Sexo* (sex) was the most searched query like in most web search engines. Notice, however, that sex represents only 2% of the total queries. Other queries were *emprego* (job) and the *eMule* P2P program. These differences can be, however, due to the temporal difference of the search logs analyzed.

4.3.2 Experiment #2: Interactive Questionnaire

Options 1 and 3 from the first question (Q1) presented in Table 4.1 were the prevalent choices of the participants. Both were selected in 31.58% of the questionnaires submitted. Option 2 was chosen 21.05%, increasing the navigational needs to a total of 52.63%. Options 4 and 5, i.e. the transactional needs, correspond to 15.79% of the participants choices. The second question (Q2) whether users searched between dates, almost divided the answers. Around 47% answered *Yes*.

I compiled some answers from the third question: *What other functionalities would you like our service to offer?* A specialized search engine for images was referred to twice, while a search engine for videos and another for old news was mentioned once. Seeing the evolution of a page or site was suggested three times, for instance to compare layouts. An example given was a side-by-side comparison between two versions of a page. Participants also proposed functionalities already supported by web search engines, such as a safe search to filter adult contents, an alert service such as the Google Alerts, auto-completion of queries on the search box, and a personal area with the user's search history.

I then collected several use-cases from the fourth question: *Give examples of how our service could help in your profession or daily activities.* The most

4. CHARACTERIZING INFORMATION NEEDS



Figure 4.6: Tag clouds of search queries submitted to (a) the Portuguese Web Archive in 2010 and (b) the Portuguese web search engine Tumba! in 2003 and 2004.

4.4 Summary

usual was the research of old information, such as political events. The interest of seeing curiosities, such as old photos, downloading software and manuals was also mentioned. Another suggested use-case was the creation of trustability profiles, based on the companies and employers background published on the past web.

4.3.3 Experiment #3: Laboratory Study

Table 4.1 shows that the prevalent choices of the participants on the first question (Q1) were options 1 and 2 with 47.62% and 33.33%, respectively. Both options represent navigational needs that together are present in 80.95% of all the tasks chosen by the participants. Option 3, which represents an informational need, was chosen 14.29%. The transactional need, i.e. option 4, was selected 4.76%. The second question (Q2) showed surprising results. Around 90% of the participants did not search between dates.

Based on the third question, the participants suggested several functionalities. Three indicated a specialized search of images or photos. Others intended to see old information, such as old events, or to compare the knowledge of today with the past. An example given was seeing the evolution of a law. Participants also suggested seeing the evolution of a page or downloading old articles or magazines currently unavailable. Four participants said that the PWA had all the necessary functionalities.

On the fourth question, users mostly answered that the PWA could help them in the research of old information, for instance to conduct studies. Another scenario was to satisfy curiosities.

4.4 Summary

Web archiving technology has been serving users without knowing nothing about them. This inevitably creates unsatisfied users, since the technology is not designed and optimized for them. Thus, in this chapter, I presented the first characterization on why and what web archive users search. Three instruments to collect quantitative and qualitative data about users were used, namely, search log mining, an online questionnaire and a laboratory study. The obtained results

4. CHARACTERIZING INFORMATION NEEDS

Studies on Users of Web Search Engines	Information Need		
	Informational	Transactional	Navigational
Broder user survey (Broder, 2002)	39%	36%	24.5%
Broder log analysis (Broder, 2002)	48%	30%	20%
Rose et al. 1st log analysis (Rose & Levinson, 2004)	60.9%	24.3%	14.7%
Rose et al. 2nd log analysis (Rose & Levinson, 2004)	61.3%	27%	11.7%
Rose et al. 3rd log analysis (Rose & Levinson, 2004)	61.5%	25%	13.5%
Jansen et al. log analysis (Jansen et al., 2008a)	65%	20%	15%
Studies on Users of Web Archives	Information Need		
	Informational	Transactional	Navigational
Experiment #1 log analysis	37.83%	5.26%	56.91%
Experiment #2 questionnaire	31.58%	15.79%	52.63%
Experiment #3 laboratory study	14.29%	4.76%	80.95%

Table 4.5: Distribution of information needs on several studies.

indicate similar tendencies, despite the percentage variations. I believe these variations are mostly due to the small number of participants in experiments #2 and #3. The results show that:

1. Information needs from users of web archives and web search engines are different. In web search engines, the users' intents are mainly informational, then transactional and lastly, navigational. In web archives, the users' intents are mainly navigational, then informational and lastly, transactional. Results of several studies in Table 4.5 attest this. This changing of needs should be reflected in the retrieval technology. For instance, the ranking of results should be tuned for navigational queries when the query type is unknown.
2. Most users do not restrict searches by date range, despite all information needs are focused on the past. This could be an interface problem. Different interfaces, such as the temporal distribution of documents matching a query or timelines, could create a richer perception of time for the user. Nevertheless, I found that informational needs are more restricted than navigational and transactional needs. This can be used, for instance, to help identifying the information need of a user and provide search results tailored toward the user goal.
3. Nearly half of the informational needs are focused on names of people, places or things. Many navigational queries only contain companies or institutions

4.4 Summary

names. The most searched queries are names of politicians. Named entity recognition can be a valuable technique to identify the best pages referring those names.

4. Web archives fail in supporting some important needs. The most commonly sought was seeing and exploring the evolution of a web page or site. This need represents 4.2% of all user sessions of the Internet Archive's Wayback Machine, but users have to request one version at a time in this system ([AlNoamany *et al.*, 2013](#)). Tools to support fast comparisons between pages and sites should be researched, such as the Diff-IE Add-on for Internet Explorer ([Teevan *et al.*, 2009](#)). Another need that is not supported, but that was significantly mentioned, is image search.

Chapter 5

Characterizing Search Patterns

A complete characterization of web archive users must respond to three questions: why, what and how do users search? The previous chapter covered the first two questions and showed that, despite current web archives are built using web search engine technology, the users of web archives have different information needs than the users of web search engines. Hence, different search patterns and behaviors are expected, which without a proper response, could degrade the search effectiveness and negatively influence user satisfaction.

In this chapter, I draw a profile on how web archive users search. It is based on the quantitative analysis of the Portuguese Web Archive (PWA) search logs, from which I have obtained detailed statistics about the user sessions, queries, query terms and clicks. I have also compared the web archive users with the users of web search engines from different world regions to analyze whether the web search engine technology can be adopted to work on web archives. Nevertheless, the identification of specificities of web archive users provides insights on search behavior and might contribute to better support the architectural design decisions of web archives.

The main contributions of this chapter are:

1. the first characterization of the search patterns of web archive users;
2. an analysis of the similarities and specificities between users of web archives and users of web search engines, aimed to better adapt web search engine technology for web archive search.

5. CHARACTERIZING SEARCH PATTERNS

The remainder of this chapter is organized as follows. Section 5.1 describes the logs dataset used in this study. Section 5.2 details the methodology followed and Section 5.3 presents the results. Section 5.4 finalizes with a summary of the chapter.

5.1 Logs Dataset

The analysis presented in this chapter is based on the logs of the PWA, covering seven months of user interactions, from June to December, 2010. By interactions, I mean all queries and clicks submitted by the users while using the user interface described in Section 4.1, and recorded by the PWA search engine (server side). The seven month span has the advantage of being less likely to be affected by ephemeral trends. The PWA system at the time provided access to nearly 150 million archived web documents ranging between 1996 and 2009.

The logs follow the Apache Common Log Format³¹. Each entry corresponds to an interaction with the search engine in the form of a HTTP request. It contains the user's IP address and the user's session identifier. Each entry also contains a timestamp indicating when the interaction occurred and the HTTP request line that came from the client.

The log data was never used to match a real identity. However, IP addresses were geographically mapped for a better characterization of the users. I found that 72% of PWA's users had IP addresses assigned to Portugal. Near 89% of the interactions were submitted through the Portuguese language interface. The remaining were submitted through the English language interface. This strongly indicates that most users were Portuguese.

5.2 Methodology

The analysis focused on four dimensions: sessions, queries, terms and clicks, defined in the following way:

³¹ <http://httpd.apache.org/docs/2.0/logs.html>

5.2 Methodology

- A *session* is a set of interactions by the same user when attempting to satisfy one information need. The session is the level of analysis in determining the success or failure of a search. It is composed by one or more queries and zero or more clicks.
- A *query* is a search request composed by a set of terms. The *initial query* is the first query submitted in a session, while all the following queries are called *subsequent*. An *identical query* is a query with exactly the same terms as the previous one submitted in the same session. A *unique query* corresponds to one query regardless of the number of times it was logged. The set of unique queries is the set of query variations. An *advanced query* is a query with at least one advanced search operator.
- A *term* is a series of characters bounded by white spaces, such as words, numbers, abbreviations, URLs, symbols or combinations between them. There are also advanced search operators, but they do not count as terms. A *unique term* is defined as one term on the dataset regardless of the number of times it was logged. The set of unique terms is the submitted lexicon.
- A *click* in this context refers to the following of a hyperlink to immediately view a query result (i.e. archived web page). Depending on where the user clicks, it can be a click on a search engine results page (SERP) or a click on a search engine versions page (SEVP). Figure 4.1 illustrates an example of a SERP, while Figure 4.3 displays an example of a SEVP.

Next, I briefly present the methods used on the search log analysis.

5.2.1 Log Preparation

The log fields for analysis were prepared through a series of data cleansing steps already described in Section 4.2.2. The delimitation of user sessions is also described in the same section. Additionally, the queries that resulted from navigation clicks to see another SERP were not counted as a new query. These are the same queries parameterized to show more results.

5. CHARACTERIZING SEARCH PATTERNS

	full-text	URL
Sessions	6 177	3 237
Queries	13 770	4 986
Terms	39 132	-
SERPs	19 812	-
Clicks on SERPs	14 664	-
Clicks on SEVPs	-	3 861
Queries per Session	2.23	1.54
Terms per Query	2.84	-
SERPs per Query	1.44	-
Clicks on SERP per Query	1.06	-
Clicks on SEVP per Query	-	1.56
Characters per Term	6.42	27.27

Table 5.1: General statistics of user interactions.

All terms were normalized to lowercase. Extra white spaces were removed. Since the PWA did not perform stemming, all variations of a query term were considered as different terms. The set of query terms also includes misspellings.

5.3 Results

Statistics were computed from the logged interactions. The first detected pattern was that users mostly conducted two types of sessions: with only full-text queries and with only URL queries, in 59.34% and 31.10% of the times, respectively. I call these *full-text sessions* and *URL sessions*. In the analysis, the remaining 9.56% sessions with mixed queries were ignored for simplification.

Table 5.1 shows the general statistics of user interactions. The users of the PWA performed 6 177 full-text sessions, averaging 2.23 queries per session. The number of terms per query was 2.84, with 6.42 characters per term. The users saw 1.44 search engine results page (SERP) per query and clicked 1.06 times on their hyperlinks to view a result. They hardly clicked in the SERPs to see all versions of a result. This only happened in 0.06 times per query. Overall, these results mean that for each query, the users saw mostly the first and sometimes the next SERP, where they clicked once.

Session duration	% full-text sessions	% URL sessions
[0, 1[59.93	81.19
[1, 5[23.07	12.42
[5, 10[6.22	2.97
[10, 15[2.77	1.95
[15, 30[4.95	1.02
[30, 60[2.18	0.45
[60, 120[0.73	0.00
[120, 180[0.10	0.00
[180, 240[0.05	0.00
[240, ∞ [0.00	0.00

Table 5.2: Session duration (minutes).

The users also submitted 3 237 URL sessions, roughly half of the full-text sessions. On average, each session had 1.54 queries with 27.27 characters. Half of the URLs submitted, 50.24%, were not found in the PWA. For the URLs found, the users clicked on 1.56 versions to see them as they were on past. Basically, a user submitted a URL and saw one or two versions of that URL.

Next, I will detail the analysis and explain the remaining results.

5.3.1 Session Level Analysis

Session duration

The duration of a session is measured from the time the first query is submitted until the last time the user interacted with the PWA. I ignored if the user spent more session time viewing the archived web pages after the last interaction or used part of the time doing parallel tasks ([Ozmutlu et al., 2003](#)). I have assigned a 0 minutes duration to sessions composed by only one query.

The large majority of sessions ended quickly, as shown in Table 5.2. Around 60% of the full-text sessions lasted less than 1 minute and 89% less than 10 minutes. Only about 3% of the sessions had a longer than an half hour duration. Each session took on average 4 minutes and 8 seconds. URL sessions lasted even less time than full-text sessions. On average, each session lasted 1 minute and 14

5. CHARACTERIZING SEARCH PATTERNS

# queries	% full-text sessions	% URL sessions
1	64.98	72.10
2	12.53	15.57
3	7.48	6.21
4	5.00	3.06
5	2.72	1.11
6	1.65	0.56
7	1.12	0.74
8	0.68	0.28
9	1.26	0.19
≥ 10	2.58	0.18

Table 5.3: Number of queries per session.

seconds. Around 81% of the sessions had a duration of less than 1 minute and only 6% took longer than 5 minutes.

Query distribution

Table 5.3 shows that the majority of the users only submitted one query. Around 85% of the full-text sessions had up to 3 queries and less than 3% had 10 or more queries. This last number can represent highly motivated users searching for special topics (e.g. porn) ([Markey, 2007](#)).

When users submitted URL sessions, 72% were composed by only one query, while 94% up to three queries. Only 2% had five or more queries. A URL query is a very specific query, where users know exactly what they are searching for. This can explain why users submitted fewer queries than in full-text sessions.

5.3.2 Query Level Analysis

Modified queries

Sometimes users submit sequences of queries as a way to refine or reformulate the search in a trial and error approach. I consider that two sequential queries submitted on the same session have the same information need if they share at

	% full-text	% URL
Initial Queries	44.86	64.92
Subsequent Queries	55.14	35.08
- Modified	44.53	-
- Identical	20.35	21.44
- Terms Swapped	3.75	-
- New	31.37	78.56
Unique Queries	68.82	73.95
Unique Terms	26.66	-
Queries never repeated	54.38	59.99
Terms never repeated	13.88	-

Table 5.4: General statistics of modified queries and terms.

least one term. In this case, the second query is called a modified query. However, the stopwords (too common terms) were ignored in this analysis. A modified query could be a specialization of the query (adding terms), a generalization (removing terms) or both at the same time.

As shown in Table 5.4, 44.53% of all subsequent full-text queries are modified queries. Table 5.5 shows that around 71% of the modified queries are the result of a zero or one change on the number of terms. A zero-length change means that the users modified some terms, but their count remained the same. Users tend to add more terms in the modified queries rather than remove them. I counted around 42% versus 25%. As in web search engines, PWA’s users tend to go from broad to narrow queries ([Costa & Silva, 2010a](#); [Jansen *et al.*, 2000](#); [Silverstein *et al.*, 1999](#)).

Identical and New queries

A variety of reasons can lead users to repeat queries, such as a refresh of the SERP or SEVP, a back-button click or the submission of the same query more than once due to a network or search engine delay. When analyzing the subsequent full-text queries, I counted 20.35% of identical queries, where each query has exactly the same terms as the previous one made in the same session (see Table 5.4). I also counted the subsequent queries with the same terms, but written in a different

5. CHARACTERIZING SEARCH PATTERNS

# terms	% modified queries
≤ -5	1.51
-4	1.33
-3	3.46
-2	6.12
-1	13.04
0	32.21
+1	25.64
+2	10.12
+3	3.11
+4	2.13
$\geq +5$	1.33

Table 5.5: Number of terms changed per modified full-text query.

order. For instance, a query *Web Archive* followed by a query *Archive Web*. Only a small number of subsequent queries, 3.75%, had the terms swapped. Besides the modified and identical queries, the users also submitted 31.37% of subsequent queries with only new terms. This indicates that at most this percentage of subsequent queries were the result of a new information need.

I divided the subsequent URL queries in identical and new queries. As shown in Table 5.4, 78.56% of the subsequent URL queries were new. The remaining 21.44% were the result of the same URL submission.

Advanced queries

In the PWA, users could use four advanced search operators:

NOT to exclude all results with a term in their text (e.g. *-web*);

PHRASE to match all results with a phrase in their text (e.g. *"web archive"*);

SITE to match all results from a domain name (e.g. *site:wikipedia.org*);

TYPE to match all results from a media type (e.g. *type:PDF*).

Table 5.6 presents the percentages of advanced queries (i.e. with at least one advanced search operator). It shows that 25.87% of the queries included

advanced operator	% advanced queries	% total queries
NOT	3.61	0.94
PHRASE	78.10	20.20
SITE	12.81	3.31
TYPE	5.48	1.42
total	100.00	25.87

Table 5.6: Advanced operators per full-text query.

some operators. This is a significantly higher percentage when compared with studies over web search engines ([Hölscher & Strube, 2000](#); [Jansen *et al.*, 2000](#); [Silverstein *et al.*, 1999](#)). The reason is the PHRASE operator, which represents 78.10% of the choices. The PWA suggested a URL within quotes for each URL submitted, to inform the users that they could match the URL in the text, as shown in Figure 4.3. However, even when ignoring the URLs within quotes, the percentages are roughly the same. The second most used operator was the SITE, occurring in 12.81% of the advanced queries. The TYPE and NOT operators were insignificantly used when compared to the total number of queries.

Term distribution

The distribution of the terms per full-text query listed in Table 5.7 shows that the majority of the queries had 1 or 2 terms. This is also visible by the 2.84 average of terms per query (see Table 5.1). Around 87% of the queries had up to 5 terms and less than 3% had 10 or more terms. These results indicate that the users tend to submit short queries. These values are useful, for instance, to optimize index structures ([Lucchese *et al.*, 2007](#)) or to determine the adequate length of the input text boxes on the user interface ([Hearst, 2009](#)).

SERPs

The users saw on average about 1.44 SERPs per full-text query. Table 5.8 shows that all users saw the first SERP as expected, since the PWA always returned it after a query. Then, the users followed the natural order of the SERPs, but

5. CHARACTERIZING SEARCH PATTERNS

# terms	% full-text queries
1	35.77
2	24.99
3	15.14
4	7.54
5	3.55
6	4.47
7	2.40
8	1.92
9	1.46
≥ 10	2.76

Table 5.7: Number of terms per query.

SERP viewed	% full-text queries
1	100.00
2	14.44
3	8.08
4	5.29
5	3.75
6	2.88
7	2.33
8	1.72
9	1.59
≥ 10	3.79

Table 5.8: SERPs viewed per query.

in a sharp decline. For instance, the second SERP was viewed in 14.44% of the queries. This indicates that prefetching the second SERP would not significantly improve web archive performance. On the other hand, the close percentages of the following SERPs indicate that prefetching them can bring improvements as shown in other studies ([Fagni *et al.*, 2006](#)).

Clicks on SERPs

The users clicked on 1.06 times per query to access an archived web page listed on the SERPs. About 66% of the clicks occurred on the first SERP. Figure 5.1 displays that users clicked on the rank of results following a power law distribution, with a 0.88 correlation. These results are similar to web search engine studies, which also present a discontinuity of clicks in the last ranking position of each SERP (multiples of 10 considering that each SERP has 10 search results) ([Baeza-Yates *et al.*, 2005](#)).

Query frequency distribution

I ranked the full-text unique queries by their decreasing frequency and verified that their distribution fits a power law with a 0.96 correlation. This power law distribution was also observed in web search engines ([Baeza-Yates *et al.*, 2008](#); [Fagni *et al.*, 2006](#)). It means that a small number of queries was submitted many

5.3 Results

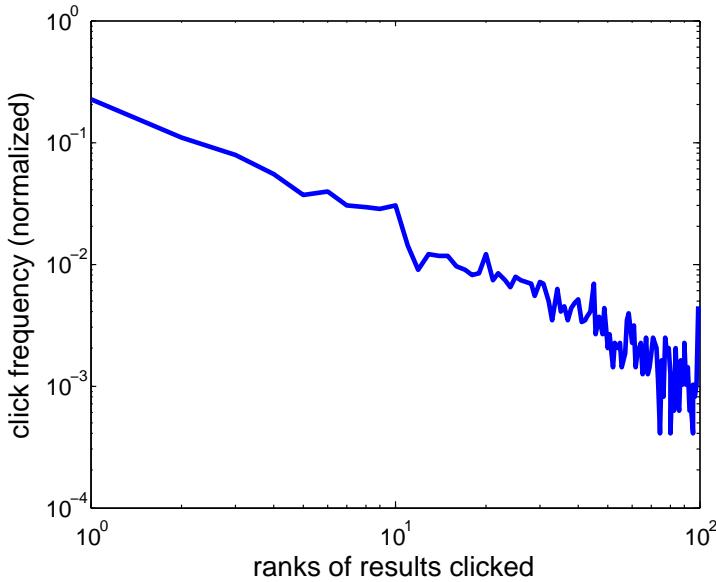


Figure 5.1: Distribution of ranks clicked on SERPs.

times, while a large number of queries were submitted just a few times. This finding offers several possibilities, for instance, for caching purposes. Figure 5.2 depicts the cumulative distribution of queries. By caching around 27% of the most frequent queries, the PWA could respond to 50% of the total query volume.

I also ranked the URL unique queries by their decreasing frequency. Their distribution, once again, fits a power law with a 0.96 correlation. By caching around 32% of the most frequent URL queries, the PWA could respond to 50% of the queries. Although satisfactory, the percentage of queries cached is much superior than in previous studies (Costa & Silva, 2010a). This is likely due to the small number of sessions analyzed, which leads to a reduced repetition.

As a consequence of the number of users' queries and clicks following a power law distribution, the number of archived pages seen by the users also follows a power law distribution, with a 0.94 correlation. This applies to both full-text and URL sessions.

5. CHARACTERIZING SEARCH PATTERNS

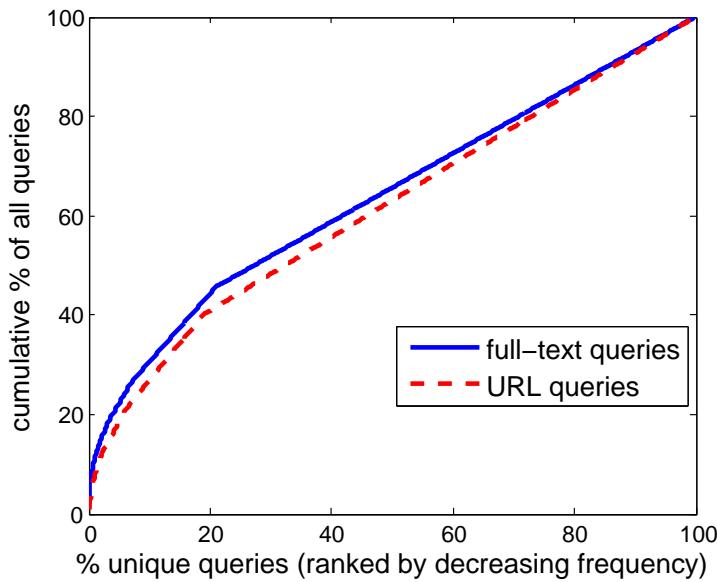


Figure 5.2: Cumulative distributions of queries.

5.3.3 Term Level Analysis

Term frequency distribution

Analogous to the query frequency distribution, I ranked the full-text unique terms by their decreasing frequency. Their distribution fits the power law with a 0.97 correlation. As depicted in Figure 5.3, the cumulative distribution shows that it is necessary to cache just around 6% of the most frequent terms to handle 50% of the queries. Much less RAM is necessary to cache terms than queries for a similar hit rate. These results are consistent with others presented for web search engines (Baeza-Yates *et al.*, 2008; Costa & Silva, 2010a). However, caching the terms instead of the queries adds extra processing over the posting lists of the inverted index, to evaluate the documents matching the query. A proper trade-off must be found.

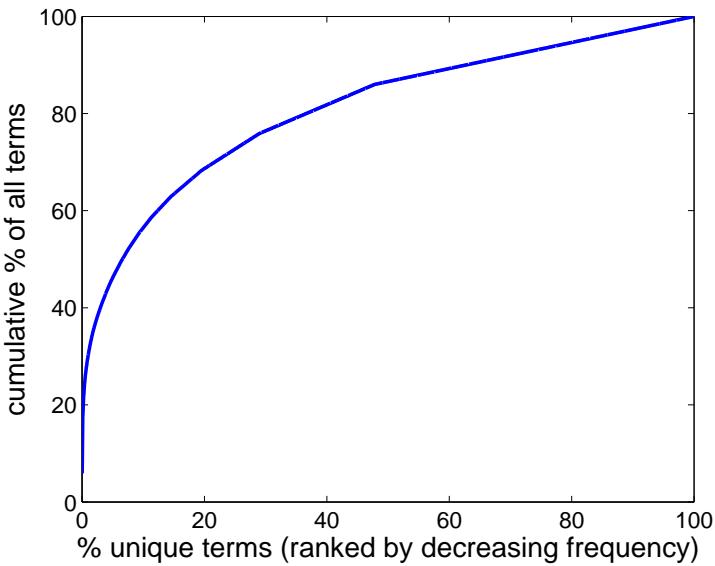


Figure 5.3: Cumulative distribution of full-text query terms.

5.3.4 Temporal Level Analysis

Queries restricted by date

The users restricted by the end date 23.55% of the full-text queries, while only 1.64% by the start date, as shown in Table 5.9. The start and end dates were both changed in 12.98% of the queries. The same pattern exists in URL queries, where the start date was changed almost only when the end date also was. This indicates that users are more interested in old documents. The idea is reinforced with the analysis of the years included in the full-text queries restricted by date. As it can be seen in Figure 5.4, the older the years, the more likely they are of being included in queries. However, the URL queries have an almost constant rate.

Clicks on temporal versions

Documents tend to have just a few years with archived versions, not always from the same time interval. Thus, segmenting the number of clicks per year would likely bias the results. Instead, I computed for all URL queries, the percentage

5. CHARACTERIZING SEARCH PATTERNS

restriction	% full-text queries	% URL queries
start date	1.64	1.34
end date	23.55	30.16
start & end date	12.98	4.88

Table 5.9: Queries restricted by date.

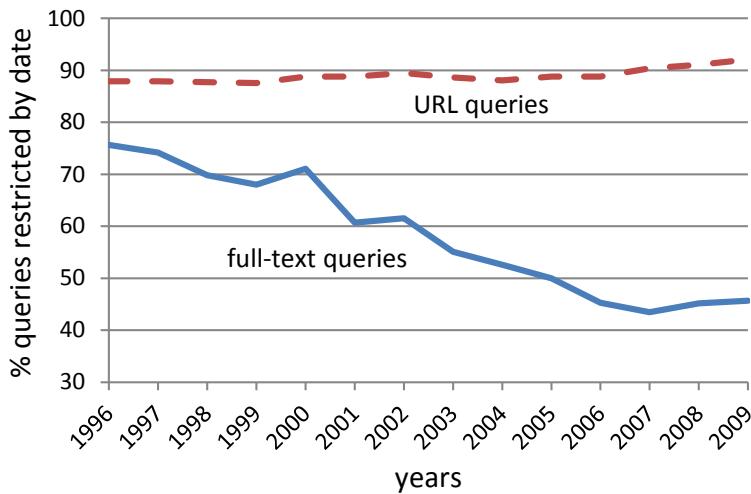


Figure 5.4: Years included in queries restricted by date.

of clicks in each year y_i with at least one version. I measured it as:

$$\frac{\text{clicks}(y_i)}{\text{times}(y_i)}$$

where the denominator represents the number of times the year y_i was displayed to the user, and the numerator the number of clicks in y_i . For instance, the first year y_1 is 1997 if there are no archived versions for that URL in 1996. Otherwise, y_1 is 1996.

In Figure 5.5 it is visible that users clicked much more on versions of documents of the initial year of archiving than on versions of the remaining years. The versions of the first year were clicked 55% of the times, while all the other years were clicked at most 20%. With the exception of the eighth year, the first

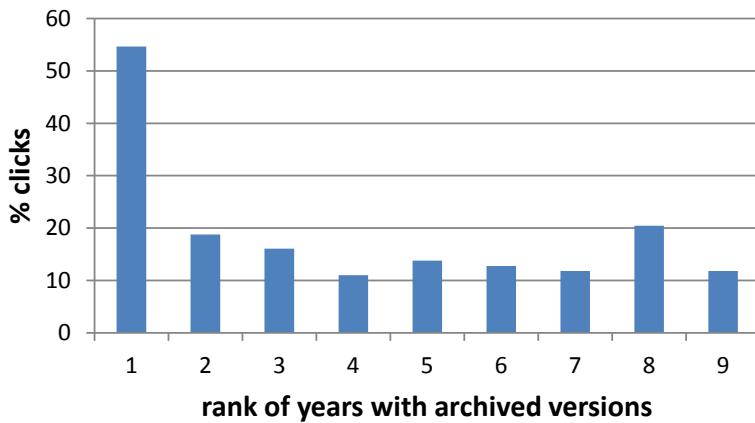


Figure 5.5: Clicks on years with archived versions (from oldest to newest).

three years had the higher percentages. This shows a preference for the older documents. A posterior study about the user access patterns on the Internet Archive’s Wayback Machine corroborates this finding ([AlNoamany *et al.*, 2013](#)). The proportion of requests out of the number of versions available in a year is higher in the older years.

Implicit temporal queries

I analyzed the number of queries with temporal expressions, since they represent a temporal dependent intent. I started by experimenting named-entity recognition tools for Portuguese, such as REMBRANDT ([Freitas *et al.*, 2010](#); [Mota & Santos, 2008](#)). However, as queries are not grammatical, the tools presented a low precision. Instead, I used a string match of all the queries with years, months and day patterns. Then, I classified a random subset of 1 000 queries to validate the detection patterns. Surprisingly, they worked very well. The patterns achieved a precision, recall and accuracy, of 89%, 100% and 98%, respectively. The patterns created some false positives, but unexpectedly no false negatives. This was mostly because there were no temporal expressions in the logs without date patterns (e.g. last decade).

All matches were manually validated, from which I excluded the false positives. In the end, I counted 3.49% of queries with temporal expressions. Almost all were related with past events, such as *world cup 2006*. This is a small percentage in

5. CHARACTERIZING SEARCH PATTERNS

IR system type	web search engine			web archive
	USA	Europe	Portugal	
world region	Excite	FAST	Tumba!	Portugal
name				PWA
single query session	55% - 60%	53% - 59%	41% - 50%	65%
queries per session	2.3	2.9	2.5 - 2.9	2.2
single term queries	20% - 30%	25% - 35%	40%	36%
terms per query	2.6	2.3	2.2	2.8
advanced queries	11% - 20%	2% - 10%	11% - 13%	26%
SERPs viewed per query	1.7	2.2	1.4	1.4

Table 5.10: Comparison between users of web search engines and web archives.

line with the 1.5% of temporal expressions found in the logs of the AOL web search engine ([Nunes *et al.*, 2008](#)).

5.4 Summary

This chapter analyzed the search patterns of web archive users and compared them with the search patterns of users of web search engines in Table 5.10. Excite from USA ([Jansen & Spink, 2006](#); [Spink *et al.*, 2002](#)), FAST from Europe ([Jansen & Spink, 2006](#); [Spink *et al.*, 2002](#)) and Tumba! from Portugal ([Costa & Silva, 2010a](#)) were the web search engines considered for comparison. This last study was conducted as part of this thesis research with the goal of comparing users of web archives and web search engines of the same country, aimed to minimize cultural bias in the results.

The results show that, as in web search, web archive users do not spend much time and effort searching the past. They also prefer short sessions, composed of short queries and few clicks. On the other hand, web archive users iterate less than users of web search engines. They submit more single query sessions, which explains the smaller number of queries per session. This finding reflects the results of the previous chapter, which show that most of the information needs of web archive users are navigational, contrary to the needs of web search engine users. Moreover, web archive users search for known-items using names, titles and URLs, some within quotes, which give good clues of the desired information.

5.4 Summary

Another explanation is that web archive users submit longer queries, which could lead to better results. On the other hand, the single term queries and the SERPs viewed per query are in conformity with web search engine results (Jansen & Spink, 2005, 2006; Jansen *et al.*, 2000; Markey, 2007).

Overall, the analyzed search patterns show no evidence precluding the adoption of web search engine technology for web archive search. This was a surprise to me, because web archive users have different information needs. For instance, web archive users said they wanted to see the evolution of a page throughout time, but they tend to click on just one or two versions of each URL. All information needs of the users are focused on the past, but most of the user queries are not restricted by date, neither contain temporal expressions. Web archive users search as in web search engines. This behavior may be the consequence of having offered a similar interface, leading them to search in a similar way. Hence, new types of interfaces must be experimented, such as the temporal distribution of documents matching a query or timelines, which could create a richer perception of time for the user and eventually trigger different search behaviors.

Nevertheless, the identification of the users' specificities might contribute to the development of better adapted web archives. I observed that half of the URLs submitted in queries were not archived. These URL queries are a good source of *seeds* for the web crawler to start. There is strong preference in searching and seeing the oldest documents over the newest. This finding can be used in ranking results, when no other temporal data is given. Queries, terms, clicked ranks and seen archived pages follow a power law distribution. This means that all have a small fraction that is repeated many times and can be exploited to increase the performance of web archives. For instance, caching around 6% of the most frequent query terms enables response to 50% of the full-text queries and caching the last query of a user in a session enables response to 20% of full-text queries and 21% of URL queries. The power law pattern can also be exploited to improve the search effectiveness with clickthrough features for ranking (Joachims, 2002; Joachims *et al.*, 2005). Other examples of the use of the findings reported in this chapter include the redesign of index structures considering the temporal dimension (Costa *et al.*, 2013a) and designing better web interfaces, such as

5. CHARACTERIZING SEARCH PATTERNS

highlighting the most used functionalities or replacing the unused functionalities ([Gomes *et al.*, 2013](#)).

An important finding is that full-text search is the preferred access type. The URL and meta-data queries are about one third and one quarter of the full-text queries, respectively. This stresses the importance of providing a high quality full-text search service to web archive users.

Chapter 6

Evaluating WAIR systems

The previous chapter showed that full-text search has become the dominant form of finding information in web archives. It gives users the ability to quickly search through vast amounts of unstructured text, powered by sophisticated ranking tools that order results based on how well they match user queries. However, the poor quality of search results still remains a major hurdle in the way of turning web archives into a usable source of information. As the amount of archived data continues to grow, this problem only tends to aggravate.

The research community and users agree that it is imperative the improvement of information search in web archives. In turn, the search improvement greatly depends on the availability of suitable evaluation methodologies and test collections. These have been a driver of research and innovation in information retrieval (IR) throughout the last decades ([Voorhees & Harman, 2005](#)), enabling to:

1. compare multiple systems and approaches, demonstrating their effectiveness and robustness;
2. measure progress and produce sustainable knowledge for future development cycles;
3. predict how well a system will perform when deployed in an operational setting;
4. research under a set of controlled conditions.

6. EVALUATING WAIR SYSTEMS

Unfortunately, methodologies and test collections have been missing for web archive information retrieval (WAIR) evaluation. On the other hand, existing evaluation methodologies and test collections from IR evaluation campaigns, such as TREC (Voorhees & Harman, 2005), are not useful for web archives, because they have different task goals and characteristics.

WAIR differs from typical IR and web IR in addressing the retrieval of document versions from web archives according to topical and temporal criteria of relevance. Temporal IR³² also considers both criteria of relevance. However, a web archive corpus is distinctively composed by a stack of content collections harvested from the web over time. Thus, each document may have several versions and the relevance of a version depends on the user's period of interest. Another main difference of WAIR is that its multi-version web collections have different characteristics over time, which causes variations in the discriminative power of features used in ranking.

This chapter presents an evaluation methodology specifically developed to measure the search effectiveness of WAIR technology. The methodology is based on a list of requirements compiled from the characterizations of web archives in Chapter 3 and their users in Chapters 4 and 5, which are essential to providing reliable and representative results tailored for the user information needs. The methodology includes the design of a test collection and the selection of evaluation measures to support reproducible experiments. I demonstrate the usefulness of the methodology through an experiment, which measured, for the first time, the search effectiveness of web archives using state-of-the-art methods. The results confirm the poor quality of search results retrieved with such technology.

The main contributions in this chapter are:

1. the first evaluation methodology proposed to measure the search effectiveness of WAIR systems and models;
2. the empirical validation of the methodology with the creation of a test collection made available to the research community;

³²http://en.wikipedia.org/wiki/Temporal_information_retrieval

3. the first measurement of the search effectiveness of state-of-the-art WAIR technology.

The remainder of this chapter is organized as follows. Section 6.1 describes the web archive characteristics that guide the design of the evaluation methodology proposed in Section 6.2. A case study applying the methodology is presented in Section 6.3. The obtained results are reported in Section 6.4 and Section 6.5 ends with a summary of this chapter.

6.1 Web Archive Characteristics

Wrong assumptions lead to wrong conclusions. Hence, before evaluating a WAIR system or model, it is necessary to understand their characteristics. Based on previous characterizations, I have compiled a list of requirements to drive the design of one such evaluation.

6.1.1 Corpus

A web archive corpus is composed by a stack of content collections harvested from the web over time. These collections are typically very heterogeneous in scope and size. Still, common characteristics across the content collections of web archives can be found, as seen in Chapter 3:

Selective and broad national crawls. Of the 68 world-wide web archive initiatives surveyed in 2014, almost all exclusively hold content related to their country, region or institution. Selective crawling was performed by 66% of the initiatives, for instance, focusing in one sub-domain or topic. These collections are narrower, but deeper, trying to crawl every URL about the topic. Broad crawling was also performed by 29% of the initiatives, including all documents hosted under a country code top-level domain or geographical location. These collections are wider, but shallower. In another survey on European web archives, 71% of them operate selective crawls and 23% broad crawls of domains ([Internet Memory Foundation, 2010](#)).

6. EVALUATING WAIR SYSTEMS

Variable number of versions per document. Some documents and sites are visited more often by crawlers due to digital preservation policies and, as result, are more frequently collected. The kind of content also influences the number of versions. For instance, newspapers have a higher change rate, while scientific articles tend to be static for long periods.

Diverse set of media types. The characterization of web collections shows that all common media types are included in web archive collections, such as text, image, sound and video, but with predominant presence of HTML, PDF, JPEG and GIF formats, which comprise over 95% of all web contents ([Baeza-Yates *et al.*, 2007a; Miranda & Gomes, 2009b](#)).

Volume of data between 1 TB and 100 TB. Most of the web archive collections in 2014 have a volume of data smaller than 100 TB (81%). The predominant volume of data is between 10 TB and 100 TB (40%), while 23% of collections have a volume of data between 1 TB and 10 TB.

Between 100 million and 1 billion documents. Most of the web archive collections in 2014, more precisely 67%, contain less than 1 billion documents (i.e. files). The predominant number of documents is between 100 million and 1 billion (33%).

Large time span of at least 7 years. Four web archives were created in 1996 and their number has been growing since then. Assuming that the oldest web collections are from the creation year of web archives, 62% of the web archives contain collections of at least 7 years old. The average age of the oldest collections preserved by web archives is 8 years. An evaluation corpus should have a large time span to not bias future WAIR technology to a specific period when some design patterns and technologies prevailed.

6.1.2 Search Topics

The evaluation of a web archive, as any other information system, must take into account the characteristics and needs of its user community. Characteriza-

tions of web archive users, mostly in Chapters 4 and 5, provide insights on the characteristics that search topics should include:

Generic use cases. Despite some professional categories being more prone to use web archives, such as historians, ordinary people also access them occasionally. There are numerous everyday life use cases that web archives can fulfill, as exemplified by [Ras & van Bussel \(2007\)](#), the [IIPC Access Working Group \(2006\)](#) and the results of Chapter 4.

Navigational and informational queries. The predominant information needs of web archive users are navigational, i.e. users intend to see how a web page or site was in the past or how it evolved throughout time. The second most usual information need is informational, i.e. users intend to collect information about a topic written in the past, usually from multiple pages without a specific one in mind. Both represent more than 90% of all information needs.

Queries about commerce and people. *Commerce* is the predominant topic category searched by users when they are trying to fulfill a navigational need, while *People* is the most predominant topic category for informational needs. The most frequent queries are names of politicians.

1/3 of queries restricted by date range. Despite user information needs being focused on the past, the ratio of queries temporally restricted in web archives is only 1/3. Another aspect is that older years are more likely of being included in such queries.

Queries without temporal clues. Only 3% of queries have expressions that could indicate a temporal dependent intent, such as *Euro 2004*.

Short queries, each with 1 to 3 terms. A typical full-text session is composed by 1 or 2 queries, each having 1 to 3 terms. Queries and terms follow a power law distribution, which means that a small fraction of each is submitted many times, while a large fraction is submitted just a few times.

6. EVALUATING WAIR SYSTEMS

6.1.3 Relevance Propagation

A document d collected at n periods has n archived versions $\{v_{t_1}^d, \dots, v_{t_n}^d\}$. A web archive enables searching over all these versions and may retrieve one or multiple versions of d . This deeply influences our understanding of relevance in two ways. First, the relevance granularity is the document's version identified by the pair $\langle URL, timestamp \rangle$. Second, the relevance is bi-dimensional. Each version has associated a temporal relevance along with a topical relevance.

Topical relevance

A navigational query intends to find an archived document for some purpose. Thus, if one version of a document d is relevant, we may assume that any version $v_{t_i}^d$ of d has the same topical relevance. Knowing this, we can propagate the topical relevance between versions of the same document. Only one version of each document needs to be assessed for navigational queries. All the other versions receive the same relevance degree.

For informational queries, the topical relevance of a version $v_{t_i}^d$ is measured according to how well it describes the searched topic in detail. Hence, since all versions $v_{t_i}^d$ of a document d may have different content, they all may have different topical relevance. We cannot propagate the topical relevance between versions of the same document, except when the content of versions $v_{t_i}^d$ is very similar (e.g. near-duplicates).

Temporal relevance

The relevance of archived versions depends also on the period of interest of the user query. Users explicitly express a date range that acts as a filter and exclude all versions with timestamps outside this range. This is the users' expected behavior, so I assume that the excluded versions are temporally not-relevant. All the others are considered equally relevant in the temporal dimension, because in web archives, highly relevant documents for a topic may exist throughout the entire search period, despite being known that some periods tend to concentrate more relevant documents ([Jones & Diaz, 2007](#)).

6.2 Evaluation Methodology

Search-equivalent versions

Summarizing what was previously discussed, I assume that two versions $v_{t_i}^d$ and $v_{t_j}^d$ of a document d , where $i \neq j$, have identical:

- topical relevance for a given navigational topic;
- topical relevance for a given informational topic if their content is very similar (e.g. near-duplicates);
- temporal relevance for a given topic if the timestamps t_i and t_j are both inside or outside the search interval.

Two versions $v_{t_i}^d$ and $v_{t_j}^d$ are defined as **search-equivalent** for a search topic u if they have the same topical and temporal relevance.

6.2 Evaluation Methodology

The proposed methodology extends the Cranfield paradigm, described in Section 2.5, to support the ad-hoc retrieval task for web archives. The Cranfield paradigm establishes the creation of a test collection, which is a laboratory testbed representing real users searching in real systems. A WAIR test collection is composed of three parts: a multi-version corpus, search topics with or without temporal restrictions, and relevance judgments. The effectiveness of a WAIR system is then measured with representative evaluation measures by comparing its search results against the known relevant documents for each search topic. A great advantage of a test collection based evaluation is that it enables to evaluate system or approach changes in a very short time.

The methodology for building a WAIR test collection, depicted in Figure 6.1, has the following steps:

1. Characterization of web archives along with their collections and users.
With the knowledge compiled in the previous section, it is now possible to build a representative test collection to draw valid conclusions.

6. EVALUATING WAIR SYSTEMS

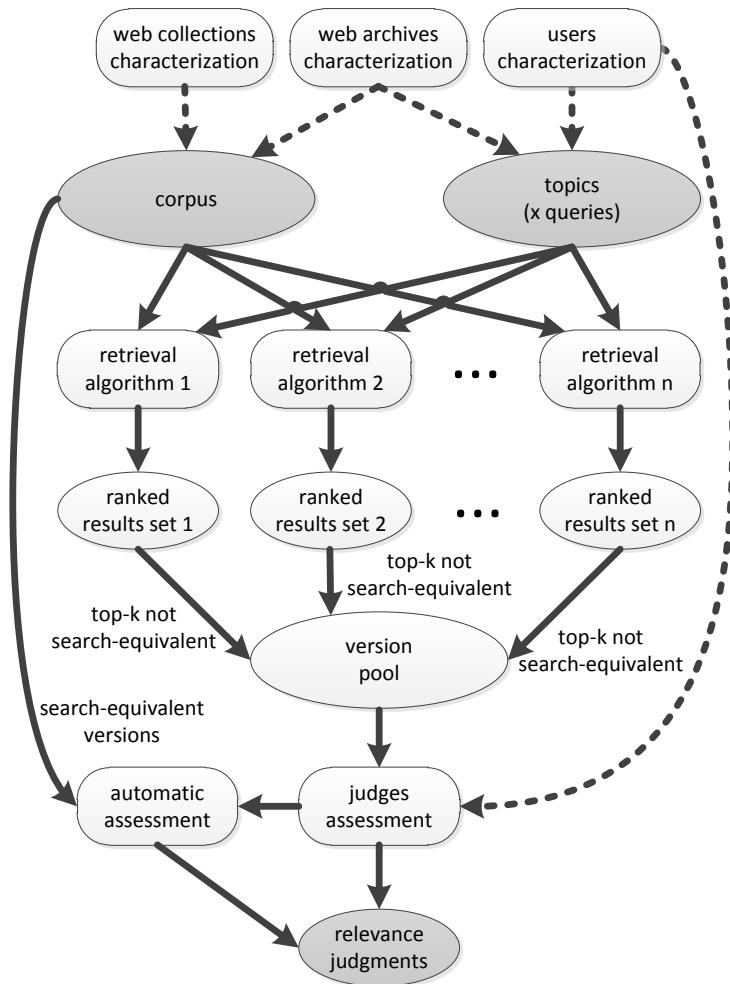


Figure 6.1: Methodology for building a WAIR test collection.

2. Selection of a representative corpus of the documents that will be encountered in a real search environment. The corpus must fit the characteristics observed in world-wide web archives, such as their size and time span.
3. Selection of search topics based on the users' information needs and search patterns. Topics are created from queries sampled from a query log of an operational web archive. These queries represent real and diverse information needs.
4. Development of several and diversified retrieval algorithms for matching

6.2 Evaluation Methodology

and ranking document versions for each search topic. These algorithms should contemplate topical and temporal features to exploit both search dimensions.

5. Aggregation of all top-k versions returned by each retrieval algorithm for each search topic into a version pool, ignoring the search-equivalent versions. The aggregated versions have their timestamps within the search interval of interest specified on topics. The versions with timestamps outside the interval are ignored, since they are considered temporally not-relevant.
6. Manual assessment of all items in the version pool by a set of judges according to the user information need defined for each search topic. The information needs are defined taking into account the characteristics of the user community when using a web archive. All versions in the pool are within the search interval and, thus, are assumed as temporally relevant.
7. Automatic assessment of all versions of a document d with a manually assessed version $v_{t_i}^d$. Each version $v_{t_j}^d$ of d receives the same topical relevance degree given to $v_{t_i}^d$ if their relevance can be inferred (i.e. if they are search-equivalent).

6.2.1 Evaluation Measures

The manual and automatic assessments form the ground-truth used to evaluate the effectiveness of all retrieval algorithms and systems. There is now the issue of selecting evaluation measures that reflect the users' search behavior. Results of Chapter 5 have shown that the measures should focus on the top 10 results, since web archive users mostly see and click on the first page of results. Results have also shown that the clicks on the rank position of results follow a power law distribution, which indicates that users click from top to bottom, as the users of web search engines. Thus, the measures should consider the relevant versions ranked ahead of the not-relevant and give a higher weight to relevant documents at higher rank positions. The dependency between search-equivalent versions must also be considered. The past experience in web archives has shown that users do not want to see multiple versions of a URL on the search results,

6. EVALUATING WAIR SYSTEMS

but rather only one URL with a link to a list of all the other versions of that URL. This corresponds to the common behavior already implemented in the user interfaces of existing WAIR systems, as shown in Section 4.1.

We have two choices to model the dependency between search-equivalent versions. The first is to design or adopt a measure, such as α -NDCG, that penalizes the relevance of search-equivalent versions (Clarke *et al.*, 2008). The second is to use a standard measure, such as NDCG, after ignoring the search-equivalent versions. I chose the second case, because it is: (1) preferable to use standard measures widely adopted within the community that were already thoroughly researched; (2) easier to optimize an IR system for one objective, than for a bi-objective where relevance is traded-off with diversity. Notice that search result diversification is an NP-hard optimization problem (Agrawal *et al.*, 2009). As a drawback, the WAIR systems should collapse these search-equivalent versions before presenting the results to the users. However, this corresponds to the common behavior already implemented in the user interfaces of existent WAIR systems (Niu, 2012a).

Concluding, I promote diversity in search results by ignoring easily identifiable search-equivalent versions before applying a standard evaluation measure. Any IR measure that can make use of relevance judgments can be used. However, these measures should have a maximum cut-off of k (e.g. NDCG@ k), where k is the number of top ranked results assessed.

6.3 Test Collection Construction

This section presents the design of a test collection from the Portuguese Web Archive (PWA) as a case study to empirically validate the proposed evaluation methodology.

6.3.1 Corpus Selection

The corpus is composed by six crawls of the Portuguese web, broadly considered the subset of the web of interest to the Portuguese people. Since the goal is to

6.3 Test Collection Construction

#	Years	#Documents (K)	Size (GB)	Description
1	1996	75	0.316	selective crawl of most popular sites
2	1996 - 2000	5 047	48	broad crawls periodically made by the Internet Archive
3	2000 - 2008	118 842	1 900	broad crawls periodically made by the Internet Archive
4	2004 - 2006	14 374	165	selective crawls made by the National Library of Portugal
5	2008	48 718	1 600	exhaustive crawl of mostly the .PT domain
6	2009	68 776	2 500	exhaustive crawl of mostly the .PT domain
Total		255 832	6 213	

Table 6.1: Web crawls that compose the corpus of the test collection.

create a corpus representative of the documents encountered in a real search environment, it only includes collections indexed and searchable through the public interface of the PWA at <http://archive.pt>. The main characteristics of the corpus are detailed in Table 6.1, showing a significant heterogeneity in age, size and type. They result from different crawls, which obtained 256 million documents, corresponding to 6.2 TB of compressed data (8.9 TB uncompressed) in ARC format (Burner & Kahle, 1996). This corpus contains some of the first documents published in the Portuguese web in 1996 and go until 2009. It includes all common types of textual formats, such as HTML, PDF and Microsoft Office, and other media formats (image, video and audio) to support a faithful rendering of document versions, which are no longer available on the live web. I consider this corpus sufficiently comprehensive and representative, but not too large to discourage its use. For comparison, the ClueWeb09³³ is the largest corpus made available to support research on IR. It contains over 1 billion web pages, which sums 5 TB compressed (25 TB uncompressed). This size is superior to the size of my corpus and several research groups have demonstrated that their IR systems scale to this order of magnitude, for instance, in the TREC web tracks since 2009.

6.3.2 Search Topics Selection

I focused on selecting navigational topics, since they represent the predominant information need of web archive users. Thus, I randomly sampled queries from the PWA query log fitting the general search patterns presented in Section 6.1. From these queries I created 50 navigational topics, where one third have temporal

³³<http://lemurproject.org/clueweb09/>

6. EVALUATING WAIR SYSTEMS

restrictions. IR evaluation campaigns generally use 50 topics, since this number gives a high confidence in the comparison between evaluated systems, especially for statistically significant differences ([Voorhees, 2009](#)). I aimed at selecting topics with different levels of complexity for IR systems, guaranteeing that a substantial part of the query terms are not present in the title or URL of the searched versions, nor all queries try to find site homepages, despite these being common. I also guaranteed that all topics have at least one relevant document archived and are not ambiguous in any sense.

The advantage of selecting queries instead of creating topics from scratch is that the submitted queries capture the real and diverse user information needs, as opposed to manually creating artificial needs. The disadvantage is that the original intent of queries is not directly available. Topic creators had to examine each query within its user session, together with all the other queries and clicks, to infer the query's underlying need. Topic creators also browsed results from related queries to identify possible interpretations of the selected query.

Each topic is composed by three fields: query, period and description. The query is the set of terms entered by a user when searching in the web archive. The period defines the range of dates of interest to the user. These two fields are the ones submitted to the WAIR system. The description specifies the user information need. This field is important to help assessors judging the relevance of a version and aid future experimenters understanding the topic. An example of a navigational topic with a search period would be:

```
<topic number="1" type="navigational">
  <query>benfica</query>
  <period>
    <start format="dd/mm/yyyy">01/01/2007</start>
    <end format="dd/mm/yyyy">31/12/2007</end>
  </period>
  <description>
    Sport Lisboa e Benfica sports club in 2007.
  </description>
</topic>
```

A set of informational topics could be created in an analogous way.

6.3.3 Retrieval

WAIR system

The corpus was indexed by the IR system of the PWA, which has been released as an open source project at <http://code.google.com/p/pwa-technologies/>. The PWA IR system executes three steps in pipeline after receiving a query with a search period:

1. versions are topically matched with the query;
2. matched versions are temporally filtered according to the search period;
3. the remaining versions are ranked by topical and temporal similarity to the query and search period.

Ranking Models

A ranking model computes a score to each matching version that is an estimate of its relevance to a query. Matching versions are then ranked by score. I implemented 9 models. The first was the Lucene's term-weighting function³⁴, which is computed over 5 fields (anchor text of incoming links, text body, title, URL and hostname of URL) with different weights. The second was a small variation of Lucene used in NutchWAX, with a different normalization by field length. These two models can be considered the state-of-the-art in WAIR, since most of the IR technology currently used in web archives is based on the Lucene and Nutch-WAX search engines, as shown in Chapter 3. As a baseline and third model, I selected the Okapi BM25 with default parameters $k_1=2$ and $b=0.75$ ([Robertson & Zaragoza, 2009](#)).

I also implemented two time-aware models that will be studied in more depth in the next chapter. The two models give a higher score to: (1) documents with more versions; (2) documents with a longer lifespan between the first and last archived versions. Both are defined by the same function:

$$f(d) = \log_y(x) \tag{6.1}$$

³⁴http://lucene.apache.org/java/2_9_0/api/all/org/apache/lucene/search/Similarity.html

6. EVALUATING WAIR SYSTEMS

where, for the first case, x is the number of versions of document d and, for the second case, x is the number of days between the first and last versions of document d . The y is the maximum possible x for normalization. Thus, $f(d)$ is normalized to a value between 0 and 1.

Each of these two time-aware models, f_1 and f_2 , was linearly combined with the NutchWAX's term-weighting function, f_3 , using three different weights: 0.1, 0.25, 0.5. That is, f_1 and f_3 were linearly combined in the following three models generally denoted by TVersions:

1. $0.1 \times f_1 + 0.9 \times f_3$;
2. $0.25 \times f_1 + 0.75 \times f_3$;
3. $0.5 \times f_1 + 0.5 \times f_3$.

while f_2 and f_3 were linearly combined in other three models generally denoted by TSpan:

4. $0.1 \times f_2 + 0.9 \times f_3$;
5. $0.25 \times f_2 + 0.75 \times f_3$;
6. $0.5 \times f_2 + 0.5 \times f_3$.

6.3.4 Relevance Assessment

Chosen Paradigm

Initially, I tried to congregate efforts from the research community for a joint IR evaluation on web archives ([Costa & Silva, 2009](#)). However, the IR community was not very aware and motivated to address the problems of the web archiving community, and the web archiving community has given priority to other issues beyond IR, such as preservation. Hence, I have explored the three most used assessment paradigms described in Section [2.5](#).

First, I tried using *implicit feedback*, but the search logs of the PWA did not have enough user interactions to extract accurate relevance assessments. For instance, few $\langle query, click \rangle$ pairs were repeated by different users. Second, I

6.3 Test Collection Construction

1. Imagine that to find the page of:
José Saramago, Nobel Prize-Winning Writer in 1998.
2. You submit the query:
jose saramago
3. And you obtain as result the:
archived page of 03-24-2007 with the <http://www.caleida.pt/saramago/> address.
4. Open the archived page and evaluate its relevance as:
 - * Highly relevant: it is exactly the page I was searching for.
 - * Relevant: it is a good alternative, but it is not the page I was searching for.
 - * Not relevant: it is not the page I was searching for.
 - * Don't know / Can not answer.
5. Justify your judgment. Your comments are valuable to us (optional):

Figure 6.2: Form used to assess navigational topics.

experimented the *crowdsourcing* paradigm with the Amazon Mechanical Turk³⁵ and CrowdFlower³⁶ services. I included several processes to control the quality of results, such as a pre-qualification test to validate the ability to perform the task. This led me to realize that almost no worker in these two services spoke Portuguese, which was necessary to understand the archived documents, and thus, the obtained assessments were too few. In the end, I followed the *pooling* paradigm, which is the most popular and widely used in major IR evaluation campaigns, such as TREC, CLEF, NTCIR and INEX.

Manual Assessment

Three judges, including the topics creator, assessed on a three-level scale of relevance, each of the 1 979 $\langle URL, timestamp, topic \rangle$ triplets aggregated in the version pool. They followed strict guidelines and document versions were presented in a random order, hiding from the judges the algorithm that retrieved the versions and their ranking order. Figure 6.2 shows the form used for collecting the relevance assessments for the navigational topics.

The usefulness of the test collection depends heavily on the level of agreement of relevance judgments. Hence, I analyzed their level of consensus. The inter agreement between judges measured by Fleiss' kappa was 0.46 when considering a ternary relevance scale or 0.55 when considering a binary scale (the highly and partially relevant were considered relevant). This shows a moderate level of

³⁵<http://www.mturk.com>

³⁶<http://crowdflower.com>

6. EVALUATING WAIR SYSTEMS

Grade	very relevant	relevant	not relevant
# manual judgments	69	91	1 819
# automatic judgments	5 168	5 571	257 083

Table 6.2: Relevance judgments in the WAIR test collection per relevance grade.

agreement, lending confidence to the judgment quality. These inter agreement values are inline with the ones of TREC judges ([Al-Maskari *et al.*, 2008](#)).

Automatic Assessment

The relevance assessment is the most time-consuming part of creating a test collection. To speed up the process, I took advantage of the characteristics of the collection to automatically assess 267 822 versions, such as described in Section 6.1. For each manually assessed version, I used the PWA IR system to find all search-equivalent versions of the same document for each topic. Then, the same topical relevance degree was propagated to all these search-equivalent versions. Table 6.2 shows the number of relevance judgments per relevance grade. As expected, the number of relevant and very relevant versions is much smaller than the not-relevant. Notice that for each navigational query there is usually only one relevant or/and very relevant result.

Extrapolating from the time spent in manual assessments, the automatic assessments reduce assessment time by more than 4 000 hours per judge.

6.3.5 General Statistics

The general statistics of the test collection are detailed in Table 6.3. It includes a corpus with about 256 million web document versions (8.9 TB of uncompressed data) archived between 1996 and 2009. The test collection also includes 269 801 document versions assessed using a three-level scale of relevance (not-relevant, relevant and very relevant). The assessed document versions were returned by 9 different ranking models in response to 50 navigational queries randomly sampled from a public web archive. This selection strategy enables to get a high coverage

document versions	256 million
data volume	8.9 TB
date range	1996 to 2009
navigational queries	50
average query length	2.23
assessed document versions	269 801
assessment scale of relevance	3-level

Table 6.3: Test collection statistics.

of relevant documents, especially because navigational queries tend to have only one (very) relevant document. The queries have 2.23 terms on average and 1/3 are restricted by date range.

6.4 Results

Table 6.4 presents the results of the ranking models described above and evaluated with the test collection. The bold entries indicate the best result achieved in each measure. We can see that BM25 and Lucene present the worst results and their effectiveness is close. The NutchWAX model has a NDCG@1, NDCG@5 and NDCG@10 superior in 3%, 5.8% and 4.1%, respectively, when compared with the Lucene model. The other measures used, Precision at cut-off k (P@k) and Success at rank k (S@k), show similar results.

The obtained results determine, for the first time, how effective is the IR technology typically used in web archives. For instance, the Lucene and NutchWAX’s results achieved an S@1 value of 0.28 and 0.32, respectively, which is less than half of the best results achieved in the 2004 Web Track, i.e. an S@1 of 0.65 ([Craswell & Hawking, 2005](#)). Despite these values not being directly comparable due to the different test collections, there is a considerable gap to the S@1 value of 0.84 obtained by Google ([Lewandowski, 2011](#)).

A promising finding is that the time-aware models are significantly better than the time-unaware. The best configuration of the two models, TVersions and TSpan, presented better NDCG@1, NDCG@5 and NDCG@10 values than the BM25 and Lucene models, for a statistical significance level of 0.01 using a

6. EVALUATING WAIR SYSTEMS

Metric	time-unaware models			time-aware models	
	BM25	Lucene	NutchWAX	TVersions	TSpan
NDCG@1	0.250	0.220	0.250	0.430 †	0.450 †
NDCG@5	0.145	0.157	0.215	0.266 †	0.263 †
NDCG@10	0.119	0.133	0.174	0.202 †	0.193
P@1	0.300	0.280	0.320	0.500 †	0.520 †
P@5	0.140	0.164	0.236	0.264	0.256
P@10	0.108	0.132	0.168	0.172	0.158
S@1	0.300	0.280	0.320	0.500 †	0.520 †
S@5	0.480	0.500	0.680	0.780	0.760
S@10	0.620	0.600	0.780	0.840	0.760

† shows a statistical significance of $p < 0.01$ against NutchWAX with a two-sided paired t-test. The bold entries indicate the best result achieved in each measure.

Table 6.4: Results of the tested ranking models.

two-tailed paired Student's t-test. When compared with NutchWAX, the TVersions model achieved NDCG@1, NDCG@5 and NDCG@10 values of 18%, 5.1% and 2.8% higher, respectively. These increases have a statistical significance of $p < 0.01$, which strongly indicates that the use of temporal information improves the effectiveness of web archives. Notice that these models could only be evaluated with a multi-version corpus as the one built. I will explore this research direction in the next chapter.

6.4.1 Topic difficulty

Figure 6.3 plots the NDCG@5 and NDCG@10 averages over the 9 tested ranking models for each of the 50 navigational topics. The topics are sorted by NDCG@5 and it is visible that the topic difficulty varies significantly, between 0 and 0.54. This variance is desirable for a test collection in order to provide topics with different levels of challenge. For instance, there are topics that present very poor results, because the query terms did not match the searched document. The query of topic 21 was *Dona Maria Segunda (second) Theatre*, but the text and link references only contained the terms *Dona Maria II Theatre*.

6.5 Summary

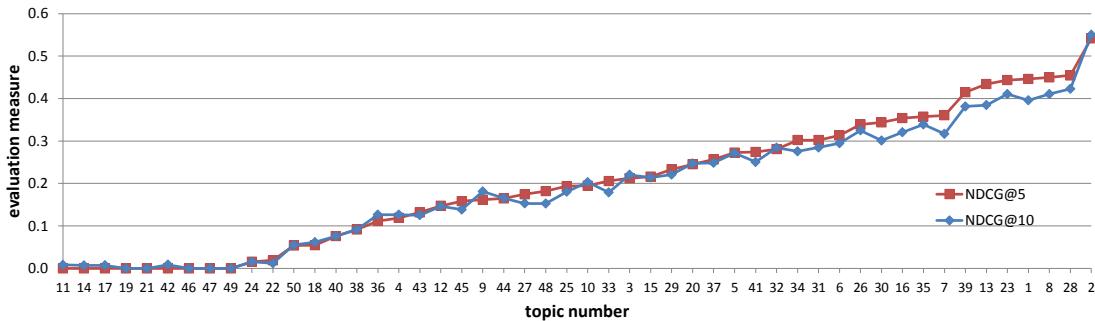


Figure 6.3: Navigational topics sorted by the average of the 9 tested ranking models.

6.4.2 Reusability

A test collection is reusable if it provides accurate measurements of the search effectiveness of systems that did not contribute with their results to the document pool. Otherwise, a new system returning relevant documents not previously identified would have its effectiveness underestimated. A test collection using only one IR system, such as this, is very likely to miss relevant documents and is biased toward that system. Nevertheless, this problem is mitigated because all topics are navigational, which tend to have only one (very) relevant document. Moreover, researchers can use this collection to accurately evaluate a new system after assessing their results and adding them to the version pool. The fact that the pool will have versions assessed by different judges over time is not a problem. The ranking between the judged systems will be the same as if judges would have assessed all documents in the same day ([Blanco *et al.*, 2011](#)).

6.5 Summary

Users cannot find the desirable information, because web archives present visibly low quality search results. It is therefore of crucial importance to improve WAIR technology, which in turn requires a systematic and reproducible evaluation to measure progress. Such evaluation methodology has been missing up to now. I believe the reason is mostly due to the lack of knowledge about the WAIR

6. EVALUATING WAIR SYSTEMS

systems and their users. It is impossible to evaluate something that we do not understand.

Previous characterizations gave us that knowledge. I have compiled a list of WAIR specificities that guided throughout this chapter the design of an evaluation methodology for WAIR systems. The proposed methodology extends the Cranfield paradigm to create a test collection composed by three representative components: a multi-version corpus, search topics and relevance judgments. Previous characterizations have enabled me to answer the following questions to build these components:

- What are the typical web collections? This answer is necessary to create a corpus.
- Why, what and how do users search? These answers are necessary to create a set of search topics.
- Where do users click (or what results do users see)? This answer is necessary to create relevance judgments.
- What and how many results do users see? These answers are necessary to design evaluation measures.

I also took advantage of the characteristics of the corpus to propagate the relevance degree of a version to all search-equivalent versions of the same document. This enabled saving more than 4 000 hours per judge, which could be a great help in the creation of future test collections.

In the end, I was able to measure, for the first time, the effectiveness of state-of-the-art WAIR technology. As anticipated, the quality of results were not satisfactory, showing that there is a large room for improvement, especially when compared with the effectiveness of existing web search engine technology. The poor quality of results motivates the development of a common evaluation framework to foster research in WAIR and thus, may one day lead to a novel IR task in a major evaluation campaign, such as TREC or CLEF.

I also experimented two time-aware ranking models for navigational queries. They are based on the idea that the more versions a document has or the longer

6.5 Summary

they existed, the more likely it is of being relevant. I observed statistically significant improvements in both models over the state-of-the-art IR typically used in web archives, which shows that WAIR can be improved by exploiting temporal information intrinsic to web archives. This is just the first step in leveraging temporal information to improve WAIR systems.

The test collection is available for research at <http://code.google.com/p/pwa-technologies/wiki/TestCollection>. Despite its specificities, such as the language, I believe that this collection could be used as a starting point to tune the WAIR technology handling other national webs.

Chapter 7

Improving WAIR systems

In the previous chapter, I proposed an evaluation methodology for web archive search systems based on a list of requirements compiled from previous characterizations of web archives and their users. The methodology includes the design of a test collection and the selection of evaluation measures that enabled, for the first time, to measure the effectiveness of state-of-the-art WAIR technology. We are now able to measure the impact of new developments.

This chapter describes how to cope with the poor search effectiveness of web archives by addressing three identified limitations. First, the ranking relevance of document versions in a web archive is currently computed based only on the similarity of their content with the query, ignoring many other features which have shown to improve the search effectiveness of web search engines. I have experimented state-of-the-art learning to rank (L2R) algorithms on such features aimed to improve the search effectiveness of state-of-the-art WAIR. Second, web archives preserve many years of collected web snapshots, but current WAIR approaches ignore the time dimension in such collections. I researched what relevant information to WAIR can be extracted from this time dimension, by exploiting, for the first time, the long-term persistence of web documents. In the conducted experiments, over 14 years of web snapshots, I found that for navigational queries, relevant documents tend to have a longer lifespan and more versions. Based on this finding I modeled the persistence of web documents into novel ranking features. These features are especially important in web archives, because the

7. IMPROVING WAIR SYSTEMS

query-independent features typically used to identify popular or important documents based on click-through data and the web-graph, are not available in this context. Web archives receive a much smaller volume of queries and clicks than web search engines, and the web-graphs are sparser since only a small part of the web is commonly collected and preserved by each archive. Third, the characteristics of the web vary over time. For instance, the sites in the 90s did not have the richer layouts and more interactive interfaces of the early 00s with CSS and JavaScript. Other examples include the dynamics of the web link structure, which grows following a power law ([Leskovec et al., 2007](#)), and the dynamics of language in web contents, which have many terms appearing and disappearing every year ([Tahmasebi et al., 2012](#)). I believe that a single general ranking model cannot predict the variance of web characteristics over such long periods of time. As a result, I have developed an approach that learns and combines multiple ranking models specific for each period, designated as temporal-dependent ranking.

The main contributions in this chapter are:

1. the first study on the use of the state-of-the-art L2R framework to improve the search effectiveness of WAIR technology. A dataset to support research on L2R applied to WAIR was made available to the research community;
2. the first analysis that exploits the correlation between the long-term persistence of web documents and relevance, from which I modeled novel ranking features that are good at discriminating relevant from not-relevant documents;
3. a novel temporal-dependent ranking framework that exploits the variance of web characteristics over time by learning and combining multiple ranking models specific for each period;
4. an empirical validation of the proposed features and framework, which in turn validates the thesis. Results show significant improvements over the search effectiveness of single-models that learn from all data independently of its time.

The remainder of this chapter is organized as follows. Section 7.1 analyzes the long-term persistence of web documents. Section 7.2 proposes a temporal-dependent ranking framework. Section 7.3 presents the experimental setup and the results are reported in Section 7.4. Section 7.5 concludes with a summary of this chapter.

7.1 Web Documents Persistence

Most ranking models have a static view of web documents and only consider their last version. I posit that web document persistence can be used to create discriminative features for improving the performance of ranking models. This section analyzes the correlation between the relevance of web documents and their long-term persistence.

7.1.1 Collection Description

The analysis uses the Portuguese Web Archive (PWA) test collection built for WAIR evaluation, described in Section 6.3. The general statistics are detailed in Table 6.3. The documents range over a period of 14 years, from 1996 to 2009. Such characteristics make this collection unique to study long-term persistence of web documents and their relation to relevance ranking. For instance, to study content change, Elsas & Dumais (2010) used a collection of 2 million documents crawled for a period of 10 weeks, Adar *et al.* (2009) used 55 thousand documents crawled during 5 weeks, Fetterly *et al.* (2003) crawled 150 million documents over a period of 11 weeks and Ntoulas *et al.* (2004) 150 web sites over the course of 1 year. These are much shorter periods of analysis not so adequate to this study.

7.1.2 Document Persistence

The persistence of web documents can be measured by their lifespan (i.e. difference in days between the first and last versions) and their number of versions. For simplification, the versions of a URL were identified by comparing their MD5 checksums. I first analyzed the distribution of the lifespan and number of versions

7. IMPROVING WAIR SYSTEMS

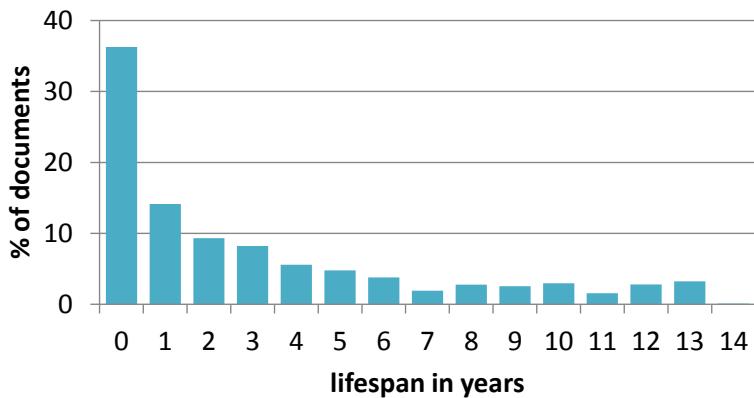


Figure 7.1: Distribution of the lifespan of documents in years.

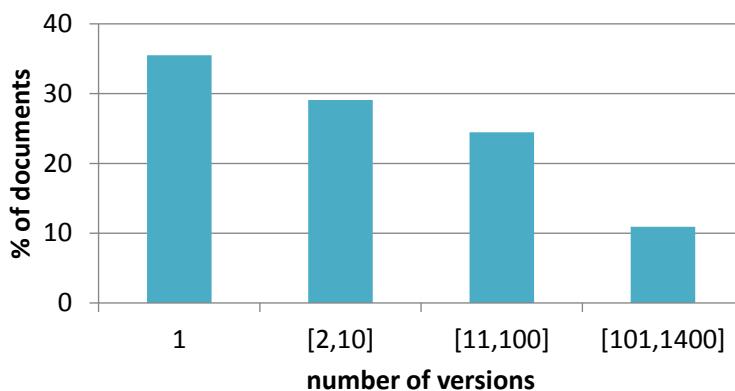


Figure 7.2: Distribution of the number of versions of documents over 14 years.

of documents in the PWA test collection. Figure 7.1 shows the lifespan distribution of web documents. Around 36% of documents have been online less than one year, to which I assigned a lifespan of 0 years. This percentage is inferior to the 50% reported by Ntoulas *et al.* (2004). 14% have a lifespan between 1 and 2 years and near 8% have a lifespan longer than 10 years. Figure 7.2 shows the distribution of the number of versions of documents. Around 36% have just 1 version, 29% have between 2 and 10, and 35% have more than 10.

The lifespan and number of versions present different distributions. While the number of versions fits a logarithmic distribution, the lifespan resembles a long tail distribution. When inspecting the documents, I saw that the document with most versions is the homepage of a newspaper (<http://www.correiomanha.pt/>)

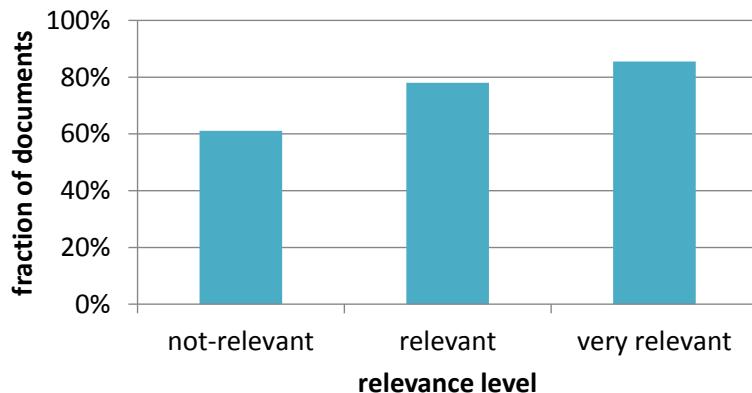


Figure 7.3: Fraction of documents with a lifespan longer than 1 year in each relevance level.

with 1 301 versions and a lifespan of 12.5 years. The document with the longest lifespan contains a list of scientific books for the younger (http://nautilus.fis.uc.pt/softc/Read_c/l_infantis/infantis.html) with a lifespan of 13 years and 2 months, but with just 8 versions. While all the documents with the highest number of versions have a long lifespan, the opposite is not true. In fact, the top ten documents with the longest lifespans have less than 15 versions. The Pearson correlation coefficient between the number of versions and the lifespan of web documents is 0.52.

7.1.3 Document Persistence & Relevance

I found some interesting patterns when analyzing the relationship between the long-term persistence of web documents and their relevance. Figure 7.3 shows the fraction of documents that have a lifespan longer than 1 year for each relevance level, i.e. the number of documents with a given relevance level and a lifespan longer than 1 year, divided by the total number of documents with that same relevance level. The figure shows that these documents are likely to have a higher relevance. The same correlation exists for documents between 1 and 5 years. The percentage of very relevant documents with more than 5 years is only 1% of the total documents for the 50 queries analyzed, which makes it difficult to identify any meaningful correlation. Nevertheless, the sum of the relevant and

7. IMPROVING WAIR SYSTEMS

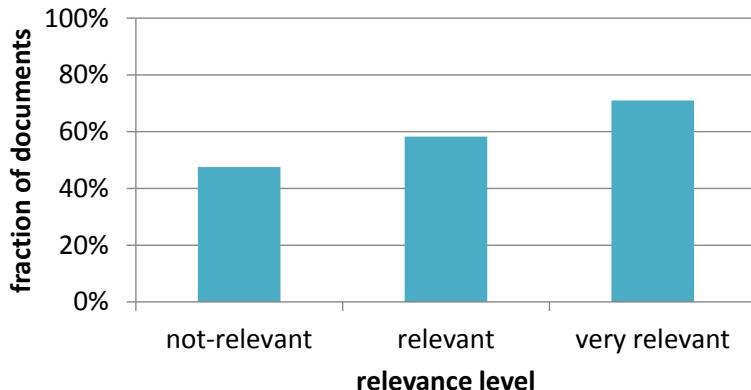


Figure 7.4: Fraction of documents with more than 10 versions in each relevance level.

very relevant fractions of documents is always superior to the not-relevant when considering the documents with a lifespan longer than 1 year. This indicates that the relevant documents tend to have a longer lifespan.

Figure 7.4 shows the fraction of documents that have more than 10 versions for each relevance level. These documents tend to have a higher relevance, such as the documents between 1 and 30 versions. The percentage of very relevant documents with more than 30 versions is only 1% of the total documents for the 50 queries analyzed. The 1% is the threshold where once again the correlation starts to be insignificant. However, the sum of the relevant and very relevant fractions of documents is always superior to the not-relevant when considering until 300 versions. After this number, the 4% of remaining documents present a different pattern. Even so, in general, these results indicate that relevant documents tend to have more versions.

7.1.4 Modeling Document Persistence

The lifespan and number of versions of documents are not correlated between them, but both are correlated with the relevance of documents. Hence, to leverage this correlation I modeled these measures of persistence with a logarithmic function that gives a higher score to: (1) documents with a longer lifespan;

7.2 Temporal-Dependent Ranking

(2) documents with more versions. Both are defined by the same function:

$$f(d) = \log_y(x) \quad (7.1)$$

where, for the first case, x is the number of days between the first and last versions of document d , and for the second case, x is the number of versions of document d . The y is the maximum possible x for normalization. The logarithmic function is just an example of a function that can be used to create ranking features, such as these two features that will be used ahead in this study.

7.2 Temporal-Dependent Ranking

This section presents the temporal-dependent ranking framework created for improving search effectiveness. First, the ranking problem is formalized. Second, it is explained how to divide the training data by time, and third, how to use these data to create temporal-dependent models. Fourth, it is described how to learn all models simultaneously and how to combine them to produce a final ranking score. Last, the implementation of this framework is detailed.

7.2.1 Ranking Problem

The traditional ranking problem is finding a ranking model f with parameters ω that receives X as input, where X is an $m \times d$ matrix of m query-document feature vectors of size d . This model f produces a vector \hat{y} of m ranking scores, one per query-document pair $< q, d >$, to predict the real relevance of document d for query q :

$$\hat{y} = f(X; \omega) \quad (7.2)$$

Manually finding and optimizing f is a laborious and prone to error work, especially when f combines multiple features. As a way to overcome this challenge, L2R algorithms automatically learn the best model \hat{f} , such that \hat{f} minimizes the given loss function L :

7. IMPROVING WAIR SYSTEMS

$$\hat{f} = \arg \min_{f \in F} \sum_{i=1}^m L(f(X_i; \omega), y_i) \quad (7.3)$$

where X_i represents the i^{th} query-document feature vector and y_i the corresponding relevance label. As Eq. 7.3 shows, the typical L2R outcome is a single general model that ranks documents independently of when they were created or updated.

7.2.2 Temporal Intervals

Instead of formulating the traditional ranking problem, we can learn multiple ranking models, each taking into account the specific characteristics of a period. In order to achieve that, a set of temporal intervals $T = \{T_1, T_2, \dots, T_n\}$ are first identified, from which multiple ranking models $M = \{M_1, M_2, \dots, M_n\}$ are then learned. Each interval T_k has associated a set of query-document feature vectors for training, where each feature vector X_i belongs to T_k if and only if the timestamp of the respective document version $t_i \in T_k$.

There are several timestamps associated to a document version, such as the dates of creation, modification, crawling or archiving. The creation and modification dates are good choices, since they refer to the time when a version was created. However, identifying them is not straightforward. The meta-data from the document's HTTP header fields, such as Date, Last-Modified and Expires are not always available, nor reliable. Studies estimate that from 35% to 64% of web documents have valid last-modified dates (Gomes & Silva, 2006), but these percentages can be significantly improved by using the dates of the web document's neighbors, especially of web resources embedded in the selected document (e.g. images, CSS, JavaScript) (Nunes *et al.*, 2007). Nevertheless, for simplification, in this work I adopted the crawling date.

7.2.3 Temporal-Dependent Models

It is hard to establish clear temporal boundaries in web data, because the ranking features tend to change gradually over time rather than abruptly. Thus, a model M_k is learned using all training instances of all intervals T , but each training instance contributes with a different weight to the learning of M_k . The

7.2 Temporal-Dependent Ranking

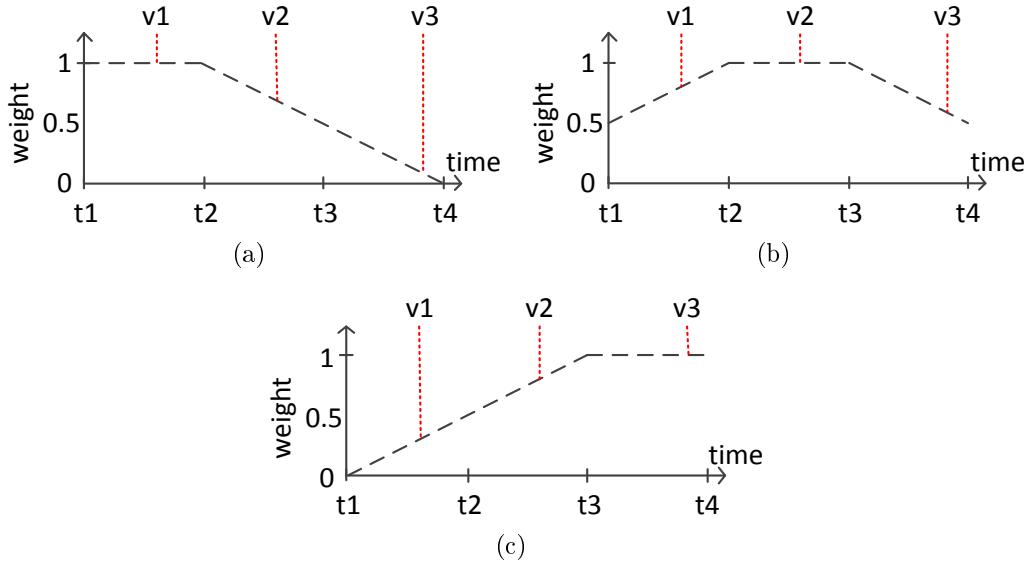


Figure 7.5: Weights of training instances, such as v_1 , v_2 and v_3 , when learning ranking models (a) M_1 , (b) M_2 and (c) M_3 .

instances of interval T_k contribute with a maximum weight, while the instances of other intervals $T_j \neq T_k$ contribute with a weight defined by their temporal distance to T_k . Consider Figures 7.5(a), 7.5(b) and 7.5(c) as illustrative examples. They depict the weights of a collection with web snapshots between time points t_1 and t_4 . Let's assume that we want to create 3 different models, $M = \{M_1, M_2, M_3\}$, taking into account the different characteristics of the web snapshots over time. For that, we divide the collection in 3 time intervals $T = \{T_1, T_2, T_3\}$ or $T = \{[t_1, t_2], [t_2, t_3], [t_3, t_4]\}$. Figure 7.5(a) shows that the training instances of interval T_1 , such as v_1 , are used with weight 1 when learning M_1 , while the other instances receive a weight that decreases as the timestamps of the instances move away from T_1 , such as v_2 and v_3 . Figures 7.5(b) and 7.5(c) show the values returned by temporal weight functions when learning M_2 and M_3 , respectively.

Contrary to typical learning to rank, my goal is to learn the best model \hat{f} for a temporal interval T_k , such that \hat{f} minimizes the following loss function L :

7. IMPROVING WAIR SYSTEMS

$$\hat{f} = \arg \min_{f \in F} \sum_{i=1}^m L(\gamma(X_i, T_k) f(X_i; \omega), y_i) \quad (7.4)$$

where γ is the temporal weight function. We can adopt several γ functions with the underlying idea that the weight decreases as the temporal distance increases, such as the following function:

$$\gamma(X_i, T_k) = \begin{cases} 1 & \text{if } X_i \in T_k \\ 1 - \alpha \frac{\text{distance}(X_i, T_k)}{|T|} & \text{if } X_i \notin T_k \end{cases} \quad (7.5)$$

s.t. $0 \leq \gamma \leq 1$

where $\text{distance}(X_i, T_k)$ is the absolute difference between the date of document version in X_i and the closer date to interval T_k , i.e. to the begin or end of T_k . $|T|$ denotes the total time covered by the collection. The γ function may have a larger or a smaller slope α to learn ranking models with higher or lower contribution of the training instances. For instance, by having a α of 2 instead of 1, the ranking model will be learned with half the contribution of the training instances and will ignore the instances in the half most distant intervals.

7.2.4 Multi-task Learning

A temporal-dependent model has two advantages over a model that only learns from data of a segment of time. First, solutions where each model learns from a part of the training data tend to present bad performance results, because more data usually beats better machine learning algorithms (Banko & Brill, 2001). Thus, each temporal-dependent model considers all training instances during learning, avoiding the problem of the lack of data. Second, a temporal-dependent model considers the dependency between datasets of different temporal intervals. A model will learn more from instances of closer intervals than from instances of intervals more far apart.

Another important aspect is that I want to minimize the overall prediction error of all temporal-dependent models, since all will be employed to rank query results. Hence, I minimize a global relevance loss function, which evaluates the

7.2 Temporal-Dependent Ranking

overall training error, instead of minimizing multiple independent loss functions without considering the correlation and overlap between models, i.e. instead of minimizing Eq. 7.4 for each model, I minimize:

$$\hat{f}_1, \dots, \hat{f}_n = \arg \min_{f_1, \dots, f_n \in F} \sum_{i=1}^m L\left(\sum_{j=1}^n \gamma(X_i, T_j) f_j(X_i; \omega), y_i\right) \quad (7.6)$$

where n is the number of temporal-dependent ranking models. The minimization of this global loss function enables learning all models simultaneously to optimize a unified relevance target. Notice that each training instance X_i is shared by each model f_j and the closer the time interval T_j to X_i the greater this sharing. Models based on data learned from time intervals far apart, will share little or no information of X_i . This is important for distant time intervals do not end up influencing negatively each other.

After learning all temporal-dependent models, an unsupervised ensemble method is employed to produce the final ranking score. I run each of the n ranking models f_j against a testing instance X_i multiplied by its temporal weight γ to the corresponding interval T_j . Then, all scores produced by all ranking models are summed:

$$score(X_i) = \sum_{j=1}^n \gamma(X_i, T_j) f_j(X_i; \omega) \quad (7.7)$$

This ensemble method follows the global loss function (Eq. 7.6) used in the learning phase, trying to minimize the overall prediction error and improve the final search effectiveness.

7.2.5 L2R Algorithm

The temporal-dependent ranking framework is quite flexible and can be implemented using different L2R algorithms as long as they are adapted to use the global loss function of Eq. 7.6. I followed the work of [Bian et al. \(2010a\)](#) and adapted the RankSVM algorithm.

The goal of RankSVM is learning a linear model that minimizes the number of pairs of documents ranked in the wrong relative order ([Joachims, 2002](#)). Formally,

7. IMPROVING WAIR SYSTEMS

RankSVM minimizes the following objective function:

$$\begin{aligned} & \min_{\omega, \xi_{q,i,j}} \frac{1}{2} \|\omega\|^2 + C \sum_{q,i,j} \xi_{q,i,j} \\ & \text{s.t. } \omega^T X_i^q \geq \omega^T X_j^q + 1 - \xi_{q,i,j}, \\ & \quad \forall X_i^q \succ X_j^q, \xi_{q,i,j} \geq 0 \end{aligned} \tag{7.8}$$

where $X_i^q \succ X_j^q$ implies that document i is ranked ahead of document j with respect to query q . C is a trade-off coefficient between the model complexity $\|\omega\|^2$ and the training error $\sum \xi_{q,i,j}$.

I modified the objective function of RankSVM following the global loss function, which takes into account the feature specificities of web snapshots over time. Each temporal-dependent ranking model M_k is learned by minimizing the following objective function:

$$\begin{aligned} & \min_{\omega, \xi_{q,i,j}} \frac{1}{2} \sum_{k=1}^n \|\omega_k\|^2 + C \sum_{q,i,j} \xi_{q,i,j} \\ & \text{s.t. } \sum_{k=1}^n \gamma(X_i^q, T_k) \omega_k^T X_i^q \geq \sum_{k=1}^n \gamma(X_j^q, T_k) \omega_k^T X_j^q + 1 - \xi_{q,i,j}, \\ & \quad \forall X_i^q \succ X_j^q, \xi_{q,i,j} \geq 0 \end{aligned} \tag{7.9}$$

7.3 Experimental Setup

This section presents the experimental setup, which enables to answer the following questions:

1. How much can we improve the search effectiveness of state-of-the-art WAIR using the L2R framework? I believe that the observations made in the context of L2R applied to document retrieval hold in relation to WAIR, but this hypothesis has not been tested.
2. Do temporal features intrinsic to web archives improve WAIR, such as the features based on the long-term persistence of web documents described in Section 7.1?

Grade	very relevant	relevant	not relevant
# judgments	4 610	4 357	30 641

Table 7.1: Relevance judgments in the L2R dataset per relevance grade.

3. Does the temporal-dependent ranking framework described in Section 7.2 improve WAIR over a single general model that fits all data independently of its time?

Next, I describe the dataset and the ranking features used in the experiments. Then, I present the compared ranking algorithms and models, and for last, the evaluation methodology and metrics.

7.3.1 L2R Dataset

For the experiments, I created a L2R dataset composed by a set of $\langle query, document\ version, grade, features \rangle$ quadruples, where the grade indicates the relevance degree of the document version to the query. The features represent a vector of ranking feature values, each describing an estimate of relevance for the $\langle query, document\ version \rangle$ pair.

From the 269 801 $\langle query, document\ version \rangle$ pairs assessed in the PWA test collection described in Section 6.3, I extracted 39 608 quadruples with 68 features. This is the size of the dataset, which has on average 843 versions per query. 3 queries were excluded from the 50, because their relevant and very relevant versions did not contain all features.

Table 7.1 shows the distribution of relevance judgments per relevance grade. As expected, the number of relevant and very relevant versions is much smaller than the not-relevant. Notice that for each of these navigational queries there is usually only one very relevant version and/or one relevant version. The dataset is publicly available for research at <http://code.google.com/p/pwa-technologies/wiki/L2R4WAIR>.

7. IMPROVING WAIR SYSTEMS

Format

The L2R dataset file format follows LETOR convention, which is based on the file format of the SVM-light software³⁷. Each of the following lines corresponds to a $\langle query, document\ version, grade, features \rangle$ quadruple, which represents one training example:

```
0 qid:21 1:0.10 2:0.233 ... 68:0.643 # id21968747index0
2 qid:21 1:0.70 2:0.344 ... 68:0.869 # id114746079index0
0 qid:22 1:0.05 2:0.112 ... 68:0.434 # id172346033index3
```

The first column is the relevance label of the $\langle query, document\ version \rangle$ pair. The larger value the relevance label has, the more relevant the version is. The second column is the query id (qid), and the following 68 columns are the feature ids with their values. The last column, after the # symbol, is the version identifier.

Feature normalization

The absolute values of the features in different queries might vary a lot. Hence, I followed LETOR guidelines and normalized the feature values across queries to make them comparable. All feature values were also normalized between 0 and 1 using a min-max normalization. Let $N^{(i)}$ be the number of document versions in the dataset with respect to a query i and $v_j^{(i)}$ a version where $1 \leq j \leq N^{(i)}$. A feature $x_j^{(i)}$ of a version $v_j^{(i)}$ was normalized as:

$$\frac{x_j^{(i)} - \min\{x_k^{(i)}, k=1, \dots, N^{(i)}\}}{\max\{x_k^{(i)}, k=1, \dots, N^{(i)}\} - \min\{x_k^{(i)}, k=1, \dots, N^{(i)}\}}$$

Partitioning

Following LETOR convention, I partitioned each dataset into five parts with the same number of queries, denoted as S1, S2, S3, S4, and S5. The idea is to evaluate results using a five-fold cross validation, where each folder contains three parts for training, one part for validation, and the remaining part for testing. The training set is used to learn ranking models. The validation set is used to tune

³⁷ <http://svmlight.joachims.org/>

Folder	Training Set	Validation Set	Test Set
1	S1, S2, S3	S4	S5
2	S2, S3, S4	S5	S1
3	S3, S4, S5	S1	S2
4	S4, S5, S1	S2	S3
5	S5, S1, S2	S3	S4

Table 7.2: Data partitioning for 5-fold cross validation.

the parameters of learning algorithms and the test set is used to evaluate the performance of the learned ranking models. The final results are the average over the five different folds described in Table 7.2.

7.3.2 Ranking Features

The performance of ranking models greatly depends on the quality of the features they use. Below it is shown an overview of the classes of the 68 features released in the L2R dataset. Each class exploits a different type of data:

term-weighting features estimate the similarity between the query and the different sections of a document version (anchor text of incoming links, text body, title and URL), such as Okapi BM25 ([Robertson & Zaragoza, 2009](#)).

term-distance features use the distance between terms in the different sections of a document version to quantify the relatedness between them, such as the Minimal Span Weighting function ([Monz, 2004](#)).

URL features compute an importance measure based on the probability of URLs representing an entry page, using the number of slashes, their length, or if they refer to a domain, sub-domain or page ([Kraaij et al., 2002](#)).

web-graph features estimate the popularity or importance of a document version inferred from the graph of hyperlinks between versions. These features include the number of inlinks to a version.

7. IMPROVING WAIR SYSTEMS

temporal features consider the time dimension of the web. They include the age of a document version and the two features described in Section 7.1.4 based on the long-term persistence of web documents.

Some of these features are typically used in web search engines and their results have been proven over time. The temporal features, however, were implemented specifically for this research. The complete list of features can be consulted in Appendix B.

7.3.3 Ranking Algorithms

The way L2R algorithms learn can be categorized into three approaches: pointwise, pairwise and listwise (Liu, 2009). I employed three state-of-the-art L2R algorithms that cover the three approaches:

pointwise: *Random Forests* consists of multiple regression trees, where each tree is built from a bootstrap sample of the training data and a random subset of features is selected to split each node of a tree (Breiman, 2001). The relevance score of each document is the average of the outputs of the individual regression trees.

pairwise: *RankSVM* (the original) which is described in Section 7.2.5.

listwise: *AdaRank* is a boosting algorithm that linearly combines "weak learners", which are iteratively selected as the feature that offers the best performance among all others (Xu & Li, 2007). Each new learner focus on the queries not ranked well on previous iteration, by giving more weight to them.

RankSVM and AdaRank produce linear models, while Random Forests produce nonlinear models. In all experiments I used the RankSVM implementation available in the *SVM^{rank}* software distribution³⁸ and the implementation of the other two L2R algorithms available in the *RankLib* software distribution³⁹.

³⁸http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

³⁹<http://www.cs.umass.edu/~vdang/ranklib.html>

7.3.4 Ranking Models Compared

To compare the search effectiveness of the proposed approaches against the state-of-the-art, I evaluated the following ranking models:

1. **Models with manually tuned features:** these are baseline models. For comparison I included the results of the three ranking models with manually tuned features, obtained from previous chapter. The first model is the Okapi BM25 with default parameters $k_1=2$ and $b=0.75$ ([Robertson & Zaragoza, 2009](#)). The second is Lucene's term-weighting function⁴⁰, which is computed over five fields (anchor text of incoming links, text body, title, URL and hostname of URL) with different weights. The third is a small variation of Lucene used in NutchWAX, with a different normalization by field length. These last two models can be considered the state-of-the-art in WAIR, since the most advanced IR technology currently used in web archives is based on the Lucene and NutchWAX search engines, as shown in Chapter [3](#).
2. **Models with regular features combined with L2R:** these are another class of baseline models, but based on the technology usually employed in web search engines. These models contain all ranking features of the L2R dataset referred in Section [7.3.2](#), except the temporal features. The regular features were automatically combined using the L2R algorithms to create a single ranking model. These models are denoted as the single-model approach with regular features.
3. **Models with all features combined with L2R:** these are the same models as in the previous point, but with all ranking features, regular and temporal. All these features were automatically combined by L2R algorithms to create a single ranking model. I refer to these models as the single-model approach with all features.
4. **Models with regular features combined with the temporal-dependent ranking framework:** unlike the previous models created independently of the time of each document version, these ranking models were created using

⁴⁰http://lucene.apache.org/java/2_9_0/api/all/org/apache/lucene/search/Similarity.html

7. IMPROVING WAIR SYSTEMS

the temporal-dependent ranking framework proposed in Section 7.2. The framework used equal intervals of time with an approximate number of training instances. The models only contain regular features.

5. **Models with all features combined with the temporal-dependent ranking framework:** these are the same models as in the previous point, but with all ranking features, regular and temporal.

7.3.5 Evaluation Methodology and Metrics

I performed a five-fold cross-validation, using the folders of the L2R dataset described in Section 7.3.1, to compare the average performance of the different ranking models.

Each of the 50 evaluated navigational queries may have one very relevant version and several relevant versions. Considering this fact, the ranking models were evaluated with two of the most used evaluation metrics: Precision at three cut-off values (P@1, P@5 and P@10) and the Normalized Discount Cumulative Gain at the same three cut-off values (NDCG@1, NDCG@5 and NDCG@10). Both metrics are described in Section 2.5.4. However, as explained in Section 6.2.1, I evaluate only the first document version shown in the search results and ignore all the other versions of the same URL, before applying P@k or NDCG@k.

7.4 Results

The results of the tested ranking models are summarized in Table 7.3.

Baselines. The NutchWAX model performs better than the Lucene and BM25 models. However, its performance is significantly worse than the models produced by the L2R algorithms using regular features. For instance, the model produced with the Random Forests algorithm, which presents the best results of the three L2R algorithms, has a NDCG@10 of 0.650, while NutchWAX gets 0.174. This is more than a three times increase. All models derived from L2R algorithms achieved better results than NutchWAX in all metrics with a statistical significance of $p < 0.01$ using a two-tailed paired Student's t-test. This strongly

7.4 Results

Metric	models with features manually tuned		
	BM25	Lucene	NutchWAX
NDCG@1	0.250	0.220	0.250
NDCG@5	0.145	0.157	0.215
NDCG@10	0.119	0.133	0.174
P@1	0.300	0.280	0.320
P@5	0.140	0.164	0.236
P@10	0.108	0.132	0.168
Metric	models with regular features combined with L2R		
	AdaRank	RankSVM	R. Forests
NDCG@1	0.380 †	0.500 †	0.550 †
NDCG@5	0.427 †	0.485 †	0.610 †
NDCG@10	0.470 †	0.523 †	0.650 †
P@1	0.460 †	0.560 †	0.640 †
P@5	0.264 †	0.276 †	0.390 †
P@10	0.182 †	0.194 †	0.236 †
Metric	models with all features combined with L2R		
	AdaRank	RankSVM	R. Forests
NDCG@1	0.400 †	0.530 ‡‡	0.650 ‡‡
NDCG@5	0.426 †	0.546 ‡‡	0.665 ‡‡
NDCG@10	0.476 †	0.571 ‡‡	0.688 ‡‡
P@1	0.480 †	0.580 ‡‡	0.760 ‡‡
P@5	0.260 †	0.324 ‡‡	0.396 ‡‡
P@10	0.182 †	0.196 †	0.238 †

Table 7.3: Results of the tested ranking models.

† shows a statistical significance of $p < 0.01$ against NutchWAX with a two-sided paired t-test, while ‡ shows a statistical significance of $p < 0.05$ against the models with regular features combined with L2R (i.e. the same model is compared with and without temporal features). The bold entries indicate the best result achieved in each metric.

7. IMPROVING WAIR SYSTEMS

indicates, as expected, that the use of L2R with ranking features typically used in web search engines, improves the search effectiveness of web archives, but also that the commonly used WAIR engines have a quite poor performance.

Temporal features. All previous models are baselines. The evaluation of temporal features is only compared against the strongest baseline, i.e. the models with regular features combined with L2R algorithms. I analyzed the discriminative power of the temporal ranking features by running the L2R algorithms with and without these features. We can see a clear pattern. The L2R algorithms almost always present statistically significant improvements for all metrics when using the temporal features. For instance, Random Forests has a NDCG@1 superior in 10% to the same algorithm learning without the temporal features and RankSVM increased 3 percentage points. Therefore, it shows that the temporal features intrinsic to web archives can indeed be used to improve WAIR.

Temporal-dependent ranking framework. Finally, I analyzed the single-model approach versus the temporal-dependent ranking framework, with and without temporal features. Figures 7.6 and 7.7 show the NDCG@1, NDCG@5 and NDCG@10 values obtained with the temporal-dependent ranking framework, when using regular features or all features. I tested the framework with different time intervals (1, 2, 4, 7 and 14) and different slopes α in the temporal weight function (0.25, 0.5, 0.75, 1, 1.25 and 1.5). Notice that the test collection has 14 years of web snapshots. Thus, when I use 14 or 7 time intervals, it means that a model is created for each year or two years, respectively. The use of 1 time interval is similar to creating just one model, i.e. the single-model approach.

The results show that the proposed temporal-dependent ranking framework outperforms the single-model approach, with and without temporal features. I achieved improvements for all time intervals, but the highest improvements were obtained when using 4 or 7 intervals. Results depicted in Figure 7.6 without temporal features, show that the major increase for NDCG@1 was from 0.500 to 0.560 (+6%) when using 4 and 7 intervals, while for a NDCG@5 was from 0.485 to 0.551 (+6.6%) and for NDCG@10 was from 0.523 to 0.572 (+4.9%), both when using 4 intervals. Results depicted in Figure 7.7 with temporal features,

7.4 Results

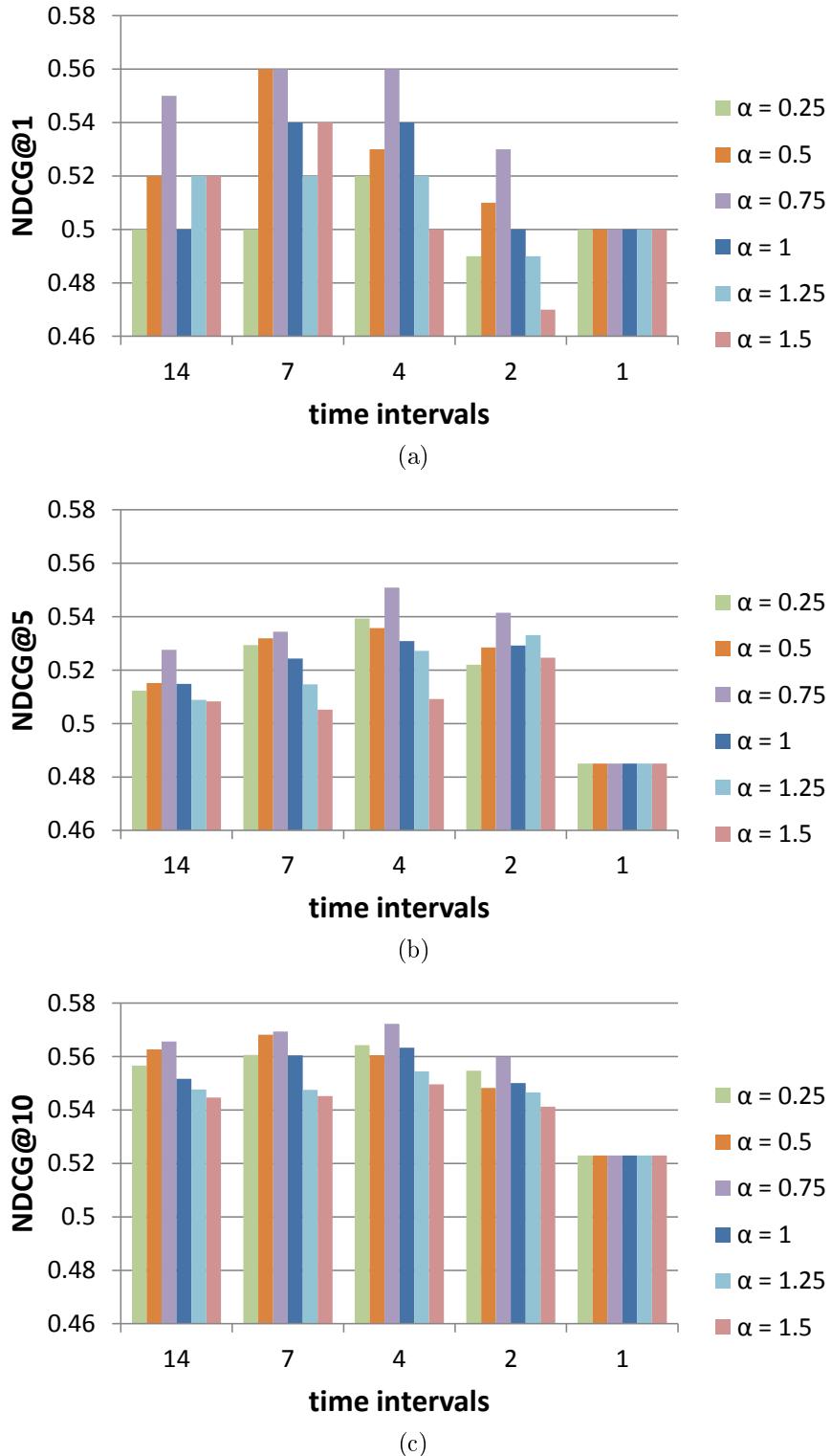


Figure 7.6: (a) NDCG@1, (b) NDCG@5 and (c) NDCG@10 results of the temporal-dependent ranking framework using different time intervals and α values of the temporal weight function. These models contain regular features.

7. IMPROVING WAIR SYSTEMS

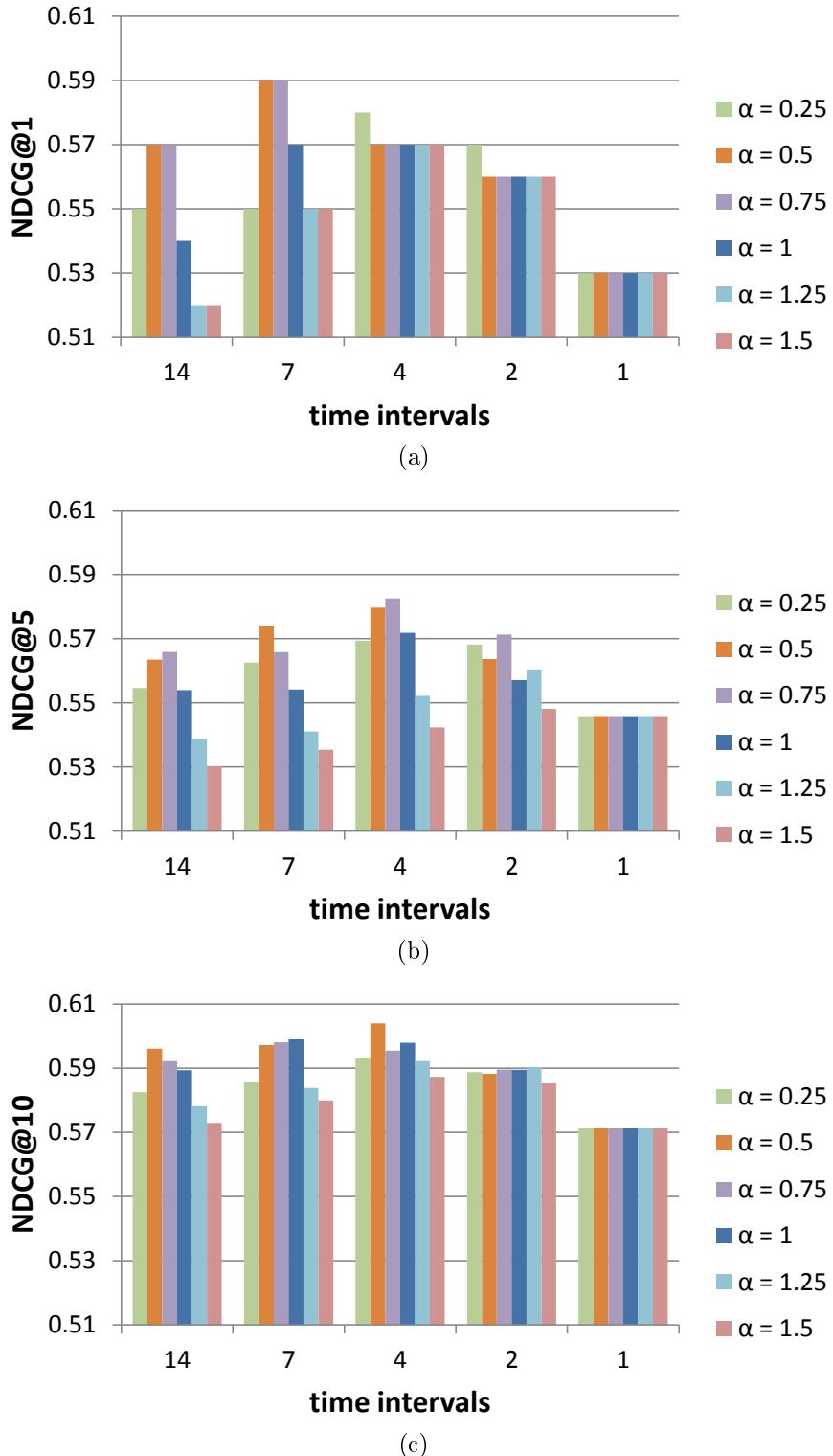


Figure 7.7: (a) NDCG@1, (b) NDCG@5 and (c) NDCG@10 results of the temporal-dependent ranking framework using different time intervals and α values of the temporal weight function. These models contain regular and temporal features.

show that the major increase for NDCG@1 was from 0.530 to 0.590 (+6%) when using 7 intervals, while for a NDCG@5 was from 0.546 to 0.583 (+3.7%) and for NDCG@10 was from 0.571 to 0.604 (+3.3%), both when using 4 intervals. The above results, which present a statistical significance ($p < 0.05$), indicate that the values of the ranking features change considerably over time in a way that can be learned by ranking models to better differentiate between relevant and not-relevant documents.

The slope α of the temporal weight function in Eq. 7.5 has an important impact in the final results. I obtained the worst results when α was larger than 1, i.e. when the contribution of the training instances is smaller. On the other hand, a small α , such as 0.25, caused a larger than desired contribution of the training instances. The best results were achieved with α between 0.5 and 1.

The temporal features and the temporal-dependent ranking framework are independent approaches that demonstrate promising results. However, both approaches also work well together. In fact, the results displayed in Figure 7.7 show that the best results can be achieved when combining them. The NDCG@1, NDCG@5 and NDCG@10 are superior in 9%, 10% and 8%, respectively, over the single-model approach using just regular features.

7.4.1 Results Analysis

To understand the superior effectiveness of the temporal-dependent ranking framework when compared against the typical single ranking models created by L2R algorithms, I sorted the ranking features by their importance, measured by the absolute weight assigned by RankSVM. The top features are almost the same, whether using just one model or multiple temporal-dependent models. The difference between ranking models created for different time intervals lies on small changes of weights of the features. This finding corroborates the observation that the characteristics of web documents evolve smoothly rather than abruptly and the temporal-dependent ranking models can adjust the feature weights to provide fine-grained ranking over time.

Table 7.4 shows the top 6 most important ranking features for the temporal-dependent ranking framework. From this table, we can see that BM25 and

7. IMPROVING WAIR SYSTEMS

BM25 over all fields
TF-IDF over all fields
Number of versions of a URL
TF-IDF over the hostname of URL
Length of the shortest text with all query terms in title
Days between the first and last versions of a URL

Table 7.4: Top 6 most important ranking features for the temporal-dependent ranking framework.

TF-IDF over all fields (anchor text of incoming links, text body, title, URL and hostname of URL) are the features with higher weight. The features based on long-term persistence of web documents, using the number of versions and the number of days between the first and last versions, are also at the top. RankSVM weighted some of these as the best features to identify relevant document versions for navigational queries.

7.5 Summary

This chapter presented a few important contributions to tackling the poor search effectiveness of state-of-the-art WAIR systems. First, the usefulness of the L2R framework in WAIR was demonstrated. The problem of finding the best version of a document to a web archive query was cast as a L2R problem. By employing state-of-the-art L2R algorithms on ranking features typically implemented in web search engines, I obtained significant improvements over the search effectiveness of state-of-the-art WAIR technology. The results show that the observations made in the context of L2R applied to document retrieval hold in relation to WAIR and suggest that future improvements in L2R technology could improve WAIR. Second, I have studied, for the first time, the effects of long-term web document persistence in relevance ranking. In the experiments, conducted over 14 years of web snapshots, relevant documents tend to have a longer lifespan and more versions. Significant gains were achieved by modeling these persistence characteristics of web documents as novel ranking features. Third, since the characteristics of the web vary over time, both in structure and content, I proposed

7.5 Summary

a temporal-dependent ranking framework. The underlying idea is that a model learned with web data from a period t will likely be more effective in ranking documents of that period t than documents of a different period u . Hence, the framework learns a different ranking model for each successive web period and combines them to produce a final ranking score. This framework tackles problems, such as how to establish temporal boundaries in web data, how to learn a period from all training instances to avoid the problem of the lack of data and how to learn more from instances of closer periods. The experimental results show that the proposed multi-model framework outperforms a simpler approach based on a single ranking model, when both use the same L2R algorithms.

The use of the proposed ranking features and temporal-dependent ranking framework achieved more than three times better results than the state-of-the-art WAIR technology, which will lead to a huge impact in the satisfaction of web archive users. The dataset, which was used in all the reported experiments, was made publicly available. It is described in this chapter and offers opportunities for several research topics in WAIR, such as feature engineering, feature selection or transfer learning.

Chapter 8

Conclusions

The goal of this thesis was to address the challenges of web archive information retrieval (WAIR) aimed to improve its state-of-the-art and fulfill the user information needs. The first challenge was to understand the status of web archiving initiatives in the world, especially the services they provide, the volume of data preserved and the state-of-the-art in WAIR. To overcome this lack of information, I conducted two surveys, in 2010 and 2014, which provide an updated and the most comprehensive characterization on web archiving initiatives. I have analyzed their evolution and found a significant growth in the number of initiatives, countries hosting these initiatives, volume of data and number of contents preserved, which indicates a growing effort that has been employed by the web archiving community to preserve the web. A cause for concern is that the amount of archived data is small in comparison with the amount of data that is permanently being published on the web. This will likely originate a knowledge gap regarding our current times. Still, the amount of archived data is larger and grows faster than the amount processed by any commercial web search engine, which raises scalability difficulties in giving efficient and effective data access.

The second challenge was to understand web archive users and whether the WAIR state-of-the-art is suitable for them. Understanding users is the first step to the success of any IT system, but surprisingly, web archiving technology has been serving users without knowing nothing about them. Hence, I conducted, for the first time, three user studies that characterize what are the user intents, which topics are most interesting to them, and how they search. The combined results of

8. CONCLUSIONS

a laboratory study, an online questionnaire and search log mining, produced the essential knowledge for guiding the development of web archives towards better user satisfaction. A major finding was that the information needs from users of web archives and web search engines are different, but both types of users are supported with the same web search engine technology. This raises the question whether web archive users should use technology not designed and optimized for them. Moreover, web archives fail in supporting some important needs, such as seeing and exploring the evolution of a web page or site, or fast comparisons between pages or sites. New developments are necessary to create these services.

The results obtained from the user studies showed that users prefer full-text as the main method for searching information in web archives. In turn, the respondents of the surveys frequently mentioned that full-text search is hard to implement and its performance is unsatisfactory. This stresses the importance of the third challenge addressed in this thesis of improving the WAIR state-of-the-art. Given the many years of collected web snapshots, I posited that the temporal information intrinsic to web archives can be exploited to improve WAIR. Thus, to prove this hypothesis, I have shown how to extract and model this temporal information. In particular, based on the assumption that the more relevant documents are maintained longer, I found a correlation between the long-term persistence of web documents and relevance for navigational queries, that was used to model novel ranking features. This persistence was measured with the number of versions and lifespan of documents. I also introduced and studied the problem of how to adapt ranking models to the successive periods covered by web archives. A single general ranking model, typically created by L2R algorithms, cannot predict the variance of web characteristics throughout long periods of time. In fact, L2R algorithms completely ignore when the documents were created or updated. Hence, I presented the concepts and techniques underlying a novel temporal-dependent ranking framework that learns and combines multiple ranking models specific for each period.

The superior performance of the novel ranking features and the temporal-dependent ranking framework, when compared with the WAIR state-of-the-art and even against the single-model approaches powered by state-of-the-art L2R algorithms, validates my thesis. The improvements are statistically significant

according to Student's paired t-test. The results were obtained through comprehensive experiments over a representative test collection and following an evaluation methodology for WAIR. The test collection, with distinct goals and characteristics, was created and made publicly available to the research community. The evaluation methodology, which extends the Cranfield paradigm to support WAIR, was proposed based on the findings gathered from all previous studies. The usefulness of the methodology and test collection was demonstrated through experiments where I measured progress and, for the first time, also measured the search effectiveness of web archives using state-of-the-art methods. In turn, the implementation of the proposed technologies in a large-scale web archive, i.e. the PWA, demonstrated their feasibility and utility in a real web archive system.

I believe that the findings of this thesis may be applied to other research domains. The proposed approaches can bring similar improvements to any digital libraries dealing with versioned content spanning long periods. Web IR can also benefit from this work if web systems, such as web search engines, will start storing the crawled web snapshots and focus in longer time horizons. However, experiments are necessary to validate this assumption. Improving the search results of web archives also brings improvements to other tools fed by these results, for instance, for temporal clustering ([Alonso *et al.*, 2009b](#)) and temporal snippets ([Alonso *et al.*, 2009a](#)), which allow users to further explore, analyze and visualize data in the time dimension.

The remainder of this chapter is organized as follows. Section 8.1 discusses the caveats to be considered when interpreting the results. Section 8.2 presents research directions for future work. Section 8.3 concludes with a brief list of the resources produced during this dissertation that can be used for further research.

8.1 Caveats

In this thesis, I used the PWA system and its data as the research environment. This choice may have biased results, since most of the data is from the Portuguese web and the users are mostly Portuguese. However, studies on national web domains show that the Portuguese web is similar to the web of any other country ([Baeza-Yates *et al.*, 2007a](#); [Gomes & Silva, 2005](#); [Miranda & Gomes, 2009a](#)).

8. CONCLUSIONS

Regarding users, I was the only conducting studies about why, what and how users search. Hence, it is not possible to compare Portuguese web archive users against web archive users of other countries. However, this thesis has shown that users from the PWA and a Portuguese web search engine have a similar search behavior (Costa & Silva, 2010a). Thus, the differences between both systems do not affect the way users search in them. Additionally, the results compiled about web search engine users across the USA and Europe, including Portugal, were also similar (Costa & Silva, 2010a; Jansen & Spink, 2006). Hence, the users' distinct language, vocabulary and culture have a small impact in the user search behavior. In conclusion, despite some nuances, it seems that users from both types of systems and different countries, have similar search behaviors. I believe that the results obtained in this thesis are general, but studies over other web archives with data from other countries and a different user population are necessary to confirm this.

Given the pioneering nature of this work, there were no evaluation resources available. I had to design an evaluation methodology and build a test collection for WAIR. Creating a test collection is a hard and laborious work. Hence, all the experiments were evaluated with just one collection. Despite my evaluations followed a five-fold cross-validation in order to get more accurate measurements and limit problems, such as overfitting, it would be desirable to test and achieve the same results with other test collections to provide a stronger validation of this thesis.

The corpus of the test collection, such as all corpora of web archives, may have several versions of documents missing due to crawling policies, errors accessing web servers or lack of web archiving initiatives during some periods. Regardless the cause, the missing versions may affect the measurement of the long-term persistence used by ranking features. This fact also suggests a limitation in the usefulness of this source of temporal data to enhance other IR systems, such as web search engines.

8.2 Outlook

This thesis presents some of the first steps in leveraging temporal information to improve WAIR systems. The obtained results are promising, but there is still much work ahead to turn web archives into usable sources of information. I briefly point out some directions for further research which could be carried out:

IR in web archives. User information needs are mostly navigational and informational. I researched generic searching tools for users to find and access information, supporting both information needs. Still, there is a large room for improvement in WAIR and plenty of opportunities for future research. The time dimension inherent to web archives likely conceals other information that can be exploited to design better ranking features. For instance, the persistence of query terms throughout document versions and anchor text of inbound links may help improve search results for navigational queries. The identification of *bursts* of documents and links about a topic may help improve and temporarily diversify search results for informational queries. I found that the proposed temporal-dependent ranking framework usually selects the same ranking features with different weights for different time intervals. This suggests that the evolution of the weights may be modeled in a way for a ranking model to automatically adapt to different time periods, instead of the solution of using multiple temporal-dependent models. Thus, better and faster search results may be computed. I think that the temporal-dependent ranking framework may be easily extended to work with other criterion to segment data, such as the geographic or demographic. Instead of creating ranking models for specific periods, they could be created for specific regions or age groups. However, it would be interesting to extend the framework to consider multiple criteria and thus, offer more personalized results.

Machine learning on web archives. IR tools require a substantial human effort when exploring and analyzing complex topics. Hence, analytical tools powered by machine learning algorithms should also be researched to fulfill informational needs for specific users requiring richer answers, such as historians or

8. CONCLUSIONS

journalists. Such tools would help explaining the stories of the past and predicting future events through the analysis and modeling of the evolution of data. Web archives are an exceptional data source to extract and leverage this evolution. A good example is the work of [Leskovec *et al.* \(2009\)](#) who tracked short units of information (e.g. phrases) from news as they spread over the web and evolve throughout time. This tracking provides a coherent representation of the news cycle, showing the rise and decline of main topics in the media. Another good example is the work of [Radinsky & Horvitz \(2013\)](#) who mined news and the web to predict future events. For instance, they found a relationship between droughts and storms in Angola that catalyze cholera outbreaks. Anticipating these events may have a huge impact in world populations. An interesting application of web archives would be extending the technology that supports sentiment analysis to determine the emotions over time when discussing specific topics ([Liu & Zhang, 2012](#)). Web archives could also be used as a source to extract entities, facts and events, which could be queried to analyze their evolution and validity time intervals, after integrated into a knowledge base ([Hoffart *et al.*, 2013](#)).

User interfaces for web archives. Web archive users search the same way as in web search engines, despite having information needs that are focused on the past. I suspect that the similar search behavior may be the consequence of having offered a similar user interface. Novel types of interfaces must be researched and experimented including, for example, presenting the temporal distribution of documents matching a query or timelines, which could create a richer perception of time for the user and eventually trigger different search behaviors. The Time Explorer is a good example for web archives, since it combines several interfaces integrated in the same application designed for analyzing how topics evolve over time ([Matthews *et al.*, 2010](#)). The core of the interface is a timeline with the main titles extracted from the news and a frequency graph with the number of news and entities most associated with a given query displayed over the time axis. The interface also displays a list of the more representative entities (people and locations) that occur on matching news and that can be used to narrow the search. The Zoetrope system also enables exploring archived data ([Adar *et al.*, 2008](#)). It introduces the concept of lenses that can be placed on any part of a web

page to see all its previous versions. These lenses can be filtered by queries and time, and combined with other lenses to compare and analyze archived data (e.g. check traffic maps at 6pm on rainy days). There are other examples, such as the visualization resources offered by the UK web archive⁴¹. However, the interfaces will always depend of the purpose of their applications. New purposes, such as the ones of the analytical tools referred above, will likely lead to new interfaces and an improved user experience.

8.3 Resources

This section presents a brief list of resources created during this dissertation that can be freely used for research:

Portuguese Web Archive system

<http://archive.pt>

Portuguese Web Archive OpenSearch API

<http://code.google.com/p/pwa-technologies/wiki/OpenSearch>

Test collection to support WAIR evaluation

<http://code.google.com/p/pwa-technologies/wiki/TestCollection>

L2R dataset for WAIR research

<http://code.google.com/p/pwa-technologies/wiki/L2R4WAIR>

All code available under the LGPL license

<http://code.google.com/p/pwa-technologies/>

⁴¹<http://www.webarchive.org.uk/ukwa/visualisation>

Appendix A

List of Web Archives Surveyed

This Appendix presents the list of the 42 web archiving initiatives identified across the world in 2010, ordered alphabetically by their hosting country.

A. LIST OF WEB ARCHIVES SURVEYED

Table A.1: List of web archives (WA). The names of the initiatives were shortened but the references contain the official ones. The description of initiatives marked with * was exclusively gathered from publicly available information.

Initiative short name	Hosting country
Australia WA (2011)	Australia
Tasmanian WA (2011)	Australia
Web@rchive (2011)	Austria
DILIMAG (2011)	Austria
Canada WA (2011)	Canada
Chinese WA (2011)*	China
Croatian WA (2011)	Croatia
WebArchiv (2011)	Czech Republic
Netarkivet.dk (2011)	Denmark
Finnish WA (2011)	Finland
BnF (2011)	France
INA (2011)*	France
Internet Memory Foundation (2011)	France, Netherlands
Baden-Württemberg (2011)	Germany
German Bundestag Web-Archiv (2011)*	Germany
Icelandic WA (2011)*	Iceland
WARP (2011)	Japan
OASIS (2011)	Korea
Koninklijke Bibliotheek WA (2011)	Netherlands
New Zealand WA (2011)	New Zealand
Norway WA (2011)*	Norway
PWA (2011)	Portugal
WA of Čačak (2011)	Serbia
WA Singapore (2011)*	Singapore
Slovenian WA (2011)	Slovenia
Preservation .ES (2011)	Spain
Digital Heritage Catalonia (2011)	Spain
Kulturarw3 (2011)*	Sweden
WA Switzerland (2011)	Switzerland
NTUWAS (2011)	Taiwan
WA Taiwan (2011)*	Taiwan
UK WA (2011)	UK
UK Gov WA (2011)	UK
Internet Archive (2011)	USA
Columbia University Libraries (2011)	USA
North Carolina WA (2011)	USA
Latin American WA (2011)*	USA
WA Pacific Islands (2011)	USA
Library of Congress WA (2011)	USA
Harvard University Library WA (2011)	USA
California Digital Library WA (2011)	USA
University of Michigan WA (2011)	USA

Table A.2: Creation year, staff and main scope of archived content of the web archiving initiatives. The description of initiatives marked with * was exclusively gathered from publicly available information.

Initiative short name	Creation year	Staff		Main scope of archived content
		Full-time	Part-time	
Australia WA	1996	4	4.25	National
Tasmanian WA	1996	0	1	Regional
Web@rchive	2008	0	2	National
DILIMAG	2007	2	0	German literature magazines
Canada WA	2005	0	2	National governmental
Chinese WA*	2003	n.a.	n.a.	National
Croatian WA	2004	4	3	National
WebArchiv	2000	5	0	National
Netarkivet.dk	2005	0	18	National
Finnish WA	2008	2	2	National
BnF	2006	9	0	National
INA*	2009	n.a.	n.a.	National audiovisual
Internet Memory Foundation	2004	21	0	International & service provider
Baden-Württemberg	2003	7.5	0	German literature
German Bundestag Web-Archiv*	2005	n.a.	n.a.	German parliament
Icelandic WA*	2004	n.a.	n.a.	National
WARP	2004	10	2	National
OASIS	2001	3	11	National
Koninklijke Bibliotheek WA	2006	1	1	National
New Zealand WA	1999	3	10	National
Norway WA*	n.a.	n.a.	n.a.	National
PWA	2007	4	1	National
WA of Čačak	2009	0	1	Regional
WA Singapore*	n.a.	n.a.	n.a.	National
Slovenian WA	2007	1	0	National
Preservation .ES	2006	2	2	National
Digital Heritage Catalonia	2006	4	0	Regional
Kulturarw3*	1996	n.a.	n.a.	National
WA Switzerland	2008	0	3	National
NTUWAS	2007	0	3	National
WA Taiwan*	2007	n.a.	n.a.	National
UK WA	2004	n.a.	0	National
UK Gov WA	2004	4	2	National governmental
Internet Archive	1996	12	0	International & service provider
Columbia University Libraries	2009	3	1	Thematic: human rights
North Carolina WA	2005	0	3	Regional
Latin American WA*	2005	n.a.	n.a.	International focused on Latin America
WA Pacific Islands	2008	0	4	International focused on Pacific Islands
Library of Congress WA	2000	6	80	National
Harvard University Library WA	2006	0	6	Institutional
California Digital Library WA	2005	4	1	International & service provider
University of Michigan WA	2000	0	2	Institutional

Appendix B

Ranking Features

This Appendix presents the complete list of 68 ranking features included in the L2R dataset for WAIR, which was used in the experiments conducted to validate this thesis. The dataset is publicly available at <http://code.google.com/p/pwa-technologies/wiki/L2R4WAIR>.

Feature	Description	Field	Comments
1	sum of the term frequency of all terms	body	
2	sum of the inverse document frequency of all terms		
3	field length		
4	average field length		
5	TF-IDF		(Salton, 1986)
6	BM-25		(Robertson & Zaragoza, 2009)
7	sum of the term frequency of all terms	URL	
8	sum of the inverse document frequency of all terms		
9	field length		
10	average field length		
11	TF-IDF		(Salton, 1986)
12	BM-25		(Robertson & Zaragoza, 2009)
13	sum of the term frequency of all terms	host of URL	
14	sum of the inverse document frequency of all terms		
15	field length		
16	average field length		
17	TF-IDF		(Salton, 1986)
18	BM-25		(Robertson & Zaragoza, 2009)
19	sum of the term frequency of all terms	anchor	
20	sum of the inverse document frequency of all terms		
21	field length		
22	average field length		

B. RANKING FEATURES

23	TF-IDF		(Salton, 1986)
24	BM-25		(Robertson & Zaragoza, 2009)
25	sum of the term frequency of all terms		
26	sum of the inverse document frequency of all terms		
27	field length	title	
28	average field length		
29	TF-IDF		(Salton, 1986)
30	BM-25		(Robertson & Zaragoza, 2009)
31	TF-IDF over all fields, having each the same weight		(Salton, 1986)
32	BM-25 over all fields, having each the same weight		(Robertson & Zaragoza, 2009)
33	Lucene	5 fields above	(Apache Lucene, 2011)
34	Lucene normalized		Lucene but with a normalized exponential decay
35	NutchWAX		Lucene but with a different normalization by field length
36	NutchWAX normalized		NutchWAX but with a normalized exponential decay
37	length of the shortest text segment with all query terms in the same order	body	(Tao & Zhai, 2007)
38	length of the shortest text segment with all query terms		(Tao & Zhai, 2007)
39	smallest distance among all pairs of matched query terms		(Tao & Zhai, 2007)
40	length of the shortest text segment with all query terms in the same order	URL	(Tao & Zhai, 2007)
41	length of the shortest text segment with all query terms		(Tao & Zhai, 2007)
42	smallest distance among all pairs of matched query terms		(Tao & Zhai, 2007)
43	length of the shortest text segment with all query terms in the same order	host of URL	(Tao & Zhai, 2007)
44	length of the shortest text segment with all query terms		(Tao & Zhai, 2007)
45	smallest distance among all pairs of matched query terms		(Tao & Zhai, 2007)
46	length of the shortest text segment with all query terms in the same order	anchor	(Tao & Zhai, 2007)
47	length of the shortest text segment with all query terms		(Tao & Zhai, 2007)
48	smallest distance among all pairs of matched query terms		(Tao & Zhai, 2007)
49	length of the shortest text segment with all query terms in the same order	title	(Tao & Zhai, 2007)
50	length of the shortest text segment with all query terms		(Tao & Zhai, 2007)

51	smallest distance among all pairs of matched query terms		(Tao & Zhai, 2007)
52	URL depth		(Kraaij <i>et al.</i> , 2002)
53	Number of URL slashes	URL	
54	URL length		
55	Number of inlinks		
56	Linearization of the number of inlinks		
57	Query issue time in days		
58	Timespan in days from the query issue time to the version date		
59	Age of the version in days		
60	Age of the oldest version of the same URL in days		
61	Age of the newest version of the same URL in days		
62	Days between the oldest and newest version of the same URL		
63	Normalized days between the oldest and newest version of the same URL		
64	Number of versions of the same URL		
65	Normalized number of versions of the same URL		
66	Exponential decay of the age of the version that boosts more recent versions		
67	Exponential decay of the age of the version that boosts older versions		
68	Exponential decay of the age of the version that boosts more recent and older versions		

Table B.1: List of ranking features of the L2R dataset for WAIR research.

In the future, more features can be extracted from the WAIR test collection publicly available for research at <http://code.google.com/p/pwa-technologies/wiki/TestCollection>. For instance, features from the query, such as the number of terms or the number of years within the search period. In addition to the features, I also released meta information, containing the mapping between the version id and the $\langle URL, timestamp \rangle$ pair. This pair can be used to locate a version in the WAIR test collection, which in turn can be used to research and derive new features from the versions, such as their type (e.g. news, spam, adult), their relationship in the corpus, sitemap information or even to extract temporally evolving web graphs.

References

- ACKLAND, R. (2005). Virtual observatory for the study of online networks (VOSON) - progress and plans. In *Proc. of the 1st International Conference on e-Social Science*. [1](#)
- ADAR, E., DONTCHEVA, M., FOGARTY, J. & WELD, D.S. (2008). Zoetrope: interacting with the ephemeral web. In *Proc. of the 21st Annual ACM Symposium on User Interface Software and Technology*, 239–248. [154](#)
- ADAR, E., TEEVAN, J., DUMAIS, S. & ELSAS, J. (2009). The web changes everything: understanding the dynamics of web content. In *Proc. of the 2nd ACM International Conference on Web Search and Data Mining*, 282–291. [30](#), [125](#)
- AGRAWAL, R., GOLLAPUDI, S., HALVERSON, A. & IEONG, S. (2009). Diversifying search results. In *Proc. of the 2nd ACM International Conference on Web Search and Data Mining*, 5–14. [110](#)
- AJI, A., WANG, Y., AGICHTEIN, E. & GABRILOVICH, E. (2010). Using the past to score the present: extending term weighting models through revision history analysis. In *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, 629–638. [30](#)
- AL-MASKARI, A., SANDERSON, M. & CLOUGH, P. (2008). Relevance judgments between TREC and non-TREC assessors. In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 683–684. [116](#)

REFERENCES

- ALCÂNTARA, O.D., PEREIRA JR, Á.R., ALMEIDA, H.M., GONÇALVES, M.A., MIDDLETON, C. & BAEZA-YATES, R. (2010). WCL2R: a benchmark collection for learning to rank research with clickthrough data. *Journal of Information and Data Management*, **1**, 551–566. [37](#)
- ALNOAMANY, Y.A., WEIGLE, M.C. & NELSON, M.L. (2013). Access patterns for robots and humans in web archives. In *Proc. of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 339–348. [27](#), [81](#), [97](#)
- ALONSO, O. & MIZZARO, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proc. of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 15–16. [40](#), [41](#)
- ALONSO, O., GERTZ, M. & BAEZA-YATES, R. (2007). On the value of temporal information in information retrieval. *ACM SIGIR Forum*, **41**, 35–41. [31](#)
- ALONSO, O., ROSE, D.E. & STEWART, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum*, **42**, 9–15. [40](#)
- ALONSO, O., BAEZA-YATES, R. & GERTZ, M. (2009a). Effectiveness of temporal snippets. In *Proc. of the WSSP Workshop at the World Wide Web Conference*, vol. 9. [151](#)
- ALONSO, O., GERTZ, M. & BAEZA-YATES, R. (2009b). Clustering and exploring search results using timeline constructions. In *Proc. of the 18th ACM Conference on Information and Knowledge Management*, 97–106. [17](#), [151](#)
- AMITAY, E., CARMEL, D., HERSCOVICI, M., LEMPEL, R. & SOFFER, A. (2004). Trend detection through temporal link analysis. *American Society for Information Science and Technology*, **55**, 1270–1281. [31](#), [32](#)
- APACHE LUCENE (2011). Lucene Similarity Function. http://lucene.apache.org/core/old_versioned_docs/versions/2_9_0/api/core/org/apache/lucene/search/Similarity.html, Accessed on March 2011. [162](#)
- ARMS, W., HUTTENLOCHER, D., KLEINBERG, J., MACY, M. & STRANG, D. (2006a). From Wayback Machine to Yesternet: new opportunities for social science. In *Proc. of the 2nd International Conference on e-Social Science*. [1](#)

REFERENCES

- ARMS, W.Y., AYA, S., DMITRIEV, P., KOT, B., MITCHELL, R. & WALLE, L. (2006b). A research library based on the historical collections of the Internet Archive. *D-Lib Magazine*, **12**. 1
- ASLAM, J., PAVLU, V. & YILMAZ, E. (2006). A statistical method for system evaluation using incomplete judgments. In *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 541–548. 39
- AULA, A., JHAVERI, N. & KÄKI, M. (2005). Information search and re-access strategies of experienced web users. In *Proc. of the 14th International Conference on World Wide Web*, 583–592. 27, 67
- AULA, A., KHAN, R.M. & GUAN, Z. (2010). How does search behavior change as search becomes more difficult? In *Proc. of the 28th International Conference on Human Factors in Computing Systems*, 35–44. 27, 37, 67
- AUSTRALIA WA (2011). Pandora Archive, National Library of Australia. <http://pandora.nla.gov.au/>, Accessed on March 2011. 158
- BADEN-WÜRTTEMBERG (2011). Bibliotheksservice-Zentrum. <http://www.bsz-bw.de/>, Accessed on March 2011. 158
- BAEZA-YATES, R. & RIBEIRO-NETO, B. (2011). *Modern information retrieval: the concepts and technology behind search*. Addison-Wesley Professional. 13, 28
- BAEZA-YATES, R., HURTADO, C., MENDOZA, M. & DUPRET, G. (2005). Modeling user search behavior. In *Proc. of the 3rd Latin American Web Congress*, 242. 92
- BAEZA-YATES, R., CASTILLO, C. & EFTHIMIADIS, E. (2007a). Characterization of national web domains. *ACM Transactions on Internet Technology*, **7**. 104, 151
- BAEZA-YATES, R., CASTILLO, C., JUNQUEIRA, F., PLACHOURAS, V. & SILVESTRI, F. (2007b). Challenges in distributed information retrieval (invited paper). In *Proc. of the 23rd International Conference on Data Engineering*. 7

REFERENCES

- BAEZA-YATES, R., GIONIS, A., JUNQUEIRA, F.P., MURDOCK, V., PLACHOURAS, V. & SILVESTRI, F. (2008). Design trade-offs for search engine caching. *ACM Transactions on the Web*, **2**, 1–28. [92](#), [94](#)
- BAILEY, P., CRASWELL, N., SOBOROFF, I., THOMAS, P., DE VRIES, A.P. & YILMAZ, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 667–674. [39](#), [41](#)
- BANKO, M. & BRILL, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proc. of the 39th Annual Meeting on Association for Computational Linguistics*, 26–33. [132](#)
- BAO, S., XUE, G., WU, X., YU, Y., FEI, B. & SU, Z. (2007). Optimizing web search using social annotations. In *Proc. of the 16th International Conference on World Wide Web*, 501–510. [29](#)
- BAR-YOSSEF, Z., BRODER, A.Z., KUMAR, R. & TOMKINS, A. (2004). Sic transit gloria telae: towards an understanding of the web's decay. In *Proc. of the 13th International Conference on World Wide Web*, 328–337. [32](#)
- BERBERICH, K., VAZIRGIANNIS, M. & WEIKUM, G. (2005). Time-aware authority ranking. *Internet Mathematics*, **2**, 301–332. [32](#)
- BERBERICH, K., BEDATHUR, S., ALONSO, O. & WEIKUM, G. (2010). A language modeling approach for temporal information needs. *Advances in Information Retrieval*, 13–25. [31](#)
- BIAN, J., LI, X., LI, F., ZHENG, Z. & ZHA, H. (2010a). Ranking specialization for web search: a divide-and-conquer approach by using topical RankSVM. In *Proc. of the 19th International Conference on World Wide Web*, 131–140. [35](#), [36](#), [133](#)
- BIAN, J., LIU, T.Y., QIN, T. & ZHA, H. (2010b). Ranking with query-dependent loss for web search. In *Proc. of the 3rd ACM International Conference on Web Search and Data Mining*, 141–150. [35](#)

REFERENCES

- BLANCO, R., HALPIN, H., HERZIG, D., MIKA, P., POUND, J., THOMPSON, H. & TRAN DUC, T. (2011). Repeatable and reliable search system evaluation using crowdsourcing. In *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information*, 923–932. [119](#)
- BNF (2011). Digital legal deposit, National Library of France. http://www.bnf.fr/en/professionals/digital_legal_deposit.html, Accessed on March 2011. [158](#)
- BREIMAN, L. (2001). Random forests. *Machine learning*, **45**, 5–32. [34](#), [138](#)
- BRIN, S. & PAGE, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, **30**, 107–117. [32](#)
- BRODER, A. (2002). A taxonomy of web search. *SIGIR Forum*, **36**, 3–10. [26](#), [74](#), [80](#)
- BUCKLEY, C. & VOORHEES, E.M. (2000). Evaluating evaluation measure stability. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 33–40. [38](#), [42](#)
- BUCKLEY, C. & VOORHEES, E.M. (2004). Retrieval evaluation with incomplete information. In *Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 25–32. [42](#)
- BURNER, M. & KAHLE, B. (1996). Arc File Format, <http://www.archive.org/web/researcher/ArcFileFormat.php>. [15](#), [47](#), [56](#), [111](#)
- CALIFORNIA DIGITAL LIBRARY WA (2011). California Digital Library. <http://webarchives.cdlib.org/>, Accessed on March 2011. [158](#)
- CANADA WA (2011). Library and Archives Canada. <http://www.collectionscanada.gc.ca/index-e.html>, Accessed on March 2011. [158](#)
- CAO, Z., QIN, T., LIU, T., TSAI, M. & LI, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Proc. of the 24th International Conference on Machine Learning*, 129–136. [34](#)

REFERENCES

- CHAPELLE, O. & CHANG, Y. (2011). Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research-Proceedings Track*, **14**, 1–24. [37](#)
- CHINESE WA (2011). WICP, National Library of China. <http://210.82.118.162:9090/webarchive>, Accessed on March 2011. [158](#)
- CHUNG, Y., TOYODA, M. & KITSUREGAWA, M. (2009). A study of link farm distribution and evolution using a time series of web snapshots. In *Proc. of the 5th International Workshop on Adversarial Information Retrieval on the Web*, 9–16. [1](#)
- CHURCH, K. & SMYTH, B. (2009). Understanding the intent behind mobile information needs. In *Proc. of the 13th International Conference on Intelligent User Interfaces*, 247–256. [27](#)
- CLARKE, C., KOLLA, M., CORMACK, G., VECHTOMOVA, O., ASHKAN, A., BÜTTCHER, S. & MACKINNON, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proc. of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 659–666. [110](#)
- CLAUSEN, L. (2004). Concerning etags and datestamps. In *Proc. of the 4th International Web Archiving Workshop*, vol. 16. [31](#)
- CLEVERDON, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, **19**, 173–193. [38](#)
- COHEN, D., AMITAY, E. & CARMEL, D. (2007). Lucene and Juru at Trec 2007: 1-million queries track. In *Proc. of the 16th Text REtrieval Conference*. [24](#)
- COLUMBIA UNIVERSITY LIBRARIES (2011). Web Resources Collection Program. https://www1.columbia.edu/sec/cu/libraries/bts/web_resource_collection/, Accessed on March 2011. [158](#)
- COSTA, M. (2004). *SIDRA: a Flexible Web Search System*. Master's thesis, University of Lisbon, Faculty of Sciences. [21](#)

REFERENCES

- COSTA, M. & SILVA, M.J. (2009). Towards information retrieval evaluation over web archives. In *Proc. of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 37–40. [6](#), [114](#)
- COSTA, M. & SILVA, M.J. (2010a). A search log analysis of a Portuguese web search engine. In *Proc. of the 2nd INForum - Simpósio de Informática*, 525–536. [5](#), [21](#), [76](#), [77](#), [89](#), [93](#), [94](#), [98](#), [152](#)
- COSTA, M. & SILVA, M.J. (2010b). Understanding the information needs of web archive users. In *Proc. of the 10th International Web Archiving Workshop*, 9–16. [6](#)
- COSTA, M. & SILVA, M.J. (2011). Characterizing search behavior in web archives. In *Proc. of the 1st International Temporal Web Analytics Workshop*, 33–40. [6](#)
- COSTA, M. & SILVA, M.J. (2012). Evaluating web archive search systems. In *Proc. of the 13th International Conference on Web Information Systems Engineering*, 440–454. [6](#)
- COSTA, M., GOMES, D., COUTO, F.M. & SILVA, M.J. (2013a). A survey of web archive search architectures. In *Proc. of the 3rd Temporal Web Analytics Workshop*. [7](#), [16](#), [19](#), [99](#)
- COSTA, M., MIRANDA, J., CRUZ, D. & GOMES, D. (2013b). Query suggestion for web archive search. In *Proc. of the 10th International Conference on Preservation of Digital Objects*. [7](#)
- COSTA, M., COUTO, F.M. & SILVA, M.J. (2014). Learning temporal-dependent ranking models. In *Proc. of the 37th Annual ACM SIGIR Conference*. [6](#)
- CRAMMER, K. & SINGER, Y. (2002). Pranking with ranking. *Advances in Neural Information Processing Systems*, **1**, 641–647. [34](#)
- CRASWELL, N. & HAWKING, D. (2005). Overview of the TREC-2004 web track. *NIST Special Publication*, 500–261. [42](#), [117](#)

REFERENCES

- CRASWELL, N., ROBERTSON, S., ZARAGOZA, H. & TAYLOR, M. (2005). Relevance weighting for query independent evidence. In *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 416–423. [32](#)
- CROATIAN WA (2011). National and University Library in Zagreb. <http://haw-nsk.hr/>, Accessed on March 2011. [158](#)
- CRUZ, D. & GOMES, D. (2013). Adapting search user interfaces to web archives. In *Proc. of the 10th International Conference on Preservation of Digital Objects*. [17](#)
- DAI, N. & DAVISON, B. (2010). Freshness matters: in flowers, food, and web authority. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 114–121. [32](#)
- DAI, N., SHOKOUI, M. & DAVISON, B. (2011). Learning to rank for freshness and relevance. In *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 95–104. [35](#)
- DAKKA, W., GRAVANO, L. & IPEIROTIS, P. (2010). Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*. [30](#)
- DELLAVALLE, R., HESTER, E., HEILIG, L., DRAKE, A., KUNTZMAN, J., GRABER, M. & SCHILLING, L. (2003). Going, going, gone: lost internet references. *Science*, **302**, 787–788. [1](#)
- DIGITAL HERITAGE CATALUNYA (2011). PADICAT, Library of Catalonia. <http://www.padicat.cat/>, Accessed on March 2011. [158](#)
- DILIMAG (2011). Innsbruck Newspaper Archive. <http://dilimag.literature.at/default.alo>, Accessed on March 2011. [158](#)
- DONG, A., CHANG, Y., ZHENG, Z., MISHNE, G., BAI, J., ZHANG, R., BUCHNER, K., LIAO, C. & DIAZ, F. (2010a). Towards recency ranking in web search. In *Proc. of the 3rd ACM International Conference on Web Search and Data Mining*, 11–20. [36](#)

REFERENCES

- DONG, A., ZHANG, R., KOLARI, P., BAI, J., DIAZ, F., CHANG, Y., ZHENG, Z. & ZHA, H. (2010b). Time is of the essence: improving recency ranking using twitter data. In *Proc. of the 19th International Conference on World Wide Web*, 331–340. [31](#)
- DOUGHERTY, M., MEYER, E., MADSEN, C., VAN DEN HEUVEL, C., THOMAS, A. & WYATT, S. (2010). Researcher engagement with web archives: state of the art. Tech. rep., Joint Information Systems Committee (JISC). [3](#), [24](#)
- ELSAS, J. & DUMAIS, S. (2010). Leveraging temporal dynamics of document content in relevance ranking. In *Proc. of the 3rd ACM International Conference on Web Search and Data Mining*, 1–10. [1](#), [3](#), [30](#), [125](#)
- EYSENBACH, G. (2004). Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). *Journal of Medical Internet Research*, **6**. [72](#)
- FAGNI, T., PEREGO, R., SILVESTRI, F. & ORLANDO, S. (2006). Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. *ACM Transactions on Information Systems*, **24**, 51–78. [92](#)
- FETTERLY, D., MANASSE, M., NAJORK, M. & WIENER, J.L. (2003). A large-scale study of the evolution of web pages. In *Proc. of the 12th International Conference on World Wide Web*, 669–678. [125](#)
- FINNISH WA (2011). The National Library of Finland. <http://verkkoarkisto.kansalliskirjasto.fi/>, Accessed on March 2011. [158](#)
- FOOT, K. & SCHNEIDER, S. (2006). *Web campaigning*. The MIT Press. [1](#)
- FOX, S., KARNAWAT, K., MYDLAND, M., DUMAIS, S. & WHITE, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, **23**, 147–168. [27](#), [39](#), [67](#)
- FRANKLIN, M. (2004). *Postcolonial politics, the internet, and everyday life: pacific traversals online*. Routledge. [1](#)

REFERENCES

- FREITAS, C., MOTA, C., SANTOS, D., OLIVEIRA, H.G. & CARVALHO, P. (2010). Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In *Proc. of the 7th International Conference on Language Resources and Evaluation*. 97
- FREUND, Y., IYER, R., SCHAPIRE, R. & SINGER, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4, 933–969. 34
- GARZÓ, A., DARÓCZY, B., KISS, T., SIKLÓSI, D. & BENCZÚR, A.A. (2013). Cross-lingual web spam classification. In *Proc. of the 22nd International Conference on World Wide Web*, 1149–1156. 7
- GENG, X., LIU, T., QIN, T., ARNOLD, A., LI, H. & SHUM, H. (2008). Query dependent ranking using k-nearest neighbor. In *Proc. of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 115–122. 35, 64
- GERMAN BUNDESTAG WEB-ARCHIV (2011). German Bundestag. <http://webarchiv.bundestag.de/cgi/kurz.php>, Accessed on March 2011. 158
- GOMES, D. & COSTA, M. (2014). The importance of web archives for humanities. *International Journal of Humanities and Arts Computing*. 8
- GOMES, D. & SILVA, M. (2006). Modelling information persistence on the web. In *Proc. of the 6th International Conference on Web Engineering*, 193–200. 31, 130
- GOMES, D. & SILVA, M.J. (2005). Characterizing a national community web. *ACM Transactions on Internet Technology*, 5, 508–531. 151
- GOMES, D., FREITAS, S. & SILVA, M.J. (2006). Design and selection criteria for a national web archive. In *Proc. of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, 196–207. 21
- GOMES, D., NOGUEIRA, A., MIRANDA, J. & COSTA, M. (2008). Introducing the Portuguese web archive initiative. In *Proc. of the 8th International Web Archiving Workshop*. 7, 15, 21

REFERENCES

- GOMES, D., MIRANDA, J. & COSTA, M. (2011). A survey on web archiving initiatives. In *Proc. of the International Conference on Theory and Practice of Digital Libraries*, 408–420. [5](#)
- GOMES, D., COSTA, M., CRUZ, D., MIRANDA, J. & FONTES, S. (2013). Creating a billion-scale searchable web archive. In *Proc. of the 3rd Temporal Web Analytics Workshop*. [6](#), [7](#), [21](#), [73](#), [100](#)
- GOOGLE INC. (2008). Official Google Blog: We knew the web was big... <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>. [56](#)
- GROTKA, A. (2008). IIPC - 2008 member profile survey results. Tech. rep., International Internet Preservation Consortium (IIPC). [20](#), [46](#)
- HARVARD UNIVERSITY LIBRARY WA (2011). Harvard University Library. <http://wax.lib.harvard.edu/collections/home.do>, Accessed on March 2011. [158](#)
- HEARST, M. (2009). *Search user interfaces*. Cambridge University Press. [91](#)
- HÖLSCHER, C. & STRUBE, G. (2000). Web search behavior of internet experts and newbies. *Computer networks*, **33**, 337–346. [91](#)
- HOFFART, J., SUCHANEK, F.M., BERBERICH, K. & WEIKUM, G. (2013). Yago2: a spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, **194**, 28–61. [154](#)
- HUURDEMAN, H.C., BEN-DAVID, A. & SAMMAR, T. (2013). Sprint methods for web archive research. In *Proc. of the 5th Annual ACM Web Science Conference*, 182–190. [20](#)
- ICELANDIC WA (2011). National and University Library of Iceland. <http://vefsafn.is/index.php?page=english>, Accessed on March 2011. [158](#)
- IIPC (2009). Internet Archive ARC access tools. <http://archive-access.sourceforge.net/>. [57](#)

REFERENCES

- IIPC ACCESS WORKING GROUP (2006). Use cases for access to Internet Archives. Tech. rep., International Internet Preservation Consortium. [25](#), [105](#)
- INA (2011). Institut National de l'Audiovisuel. <http://www.ina.fr/>, Accessed on March 2011. [158](#)
- INTERNET ARCHIVE (2011). Digital Library of Free Books, Movies, Music & Wayback Machine. <http://www.archive.org/>, Accessed on March 2011. [158](#)
- INTERNET MEMORY FOUNDATION (2011). Formerly European Archive. <http://internetmemory.org/en/>, Accessed on March 2011. [158](#)
- INTERNET MEMORY FOUNDATION (2010). Web archiving in Europe. Tech. rep., Internet Memory Foundation. [3](#), [22](#), [48](#), [103](#)
- IREM ARIKAN, S.B. & BERBERICH, K. (2009). Time will tell: leveraging temporal expressions in IR. In *Proc. of the 2nd ACM International Conference on Web Search and Data Mining*. [31](#)
- ISO 28500:2009 (2009). Information and documentation - WARC file format. http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717. [15](#), [56](#)
- JAFFE, E. & KIRKPATRICK, S. (2009). Architecture of the Internet Archive. In *Proc. of SYSTOR 2009: The Israeli Experimental Systems Conference*, 1–10. [22](#)
- JANSEN, B. & SPINK, A. (2005). An analysis of web searching by European AlltheWeb.com users. *Information Processing and Management*, **41**, 361–381. [69](#), [99](#)
- JANSEN, B. & SPINK, A. (2006). How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing and Management*, **42**, 248–263. [27](#), [28](#), [39](#), [42](#), [67](#), [69](#), [70](#), [76](#), [98](#), [99](#), [152](#)
- JANSEN, B., BOOTH, D. & SPINK, A. (2008a). Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, **44**, 1251–1266. [27](#), [80](#)

REFERENCES

- JANSEN, B., SPINK, A. & TAKSA, I. (2008b). *Handbook of research on web log analysis: surveys as a complementary method for web log analysis*. Information Science Reference. 71
- JANSEN, B.J., SPINK, A. & SARACEVIC, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, **36**, 207–227. 28, 89, 91, 99
- JÄRVELIN, K. & KEKÄLÄINEN, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, **20**, 422–446. 43
- JOACHIMS, T. (2002). Optimizing search engines using clickthrough data. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142. 29, 34, 39, 99, 133
- JOACHIMS, T., GRANKA, L., PAN, B., HEMBROOKE, H. & GAY, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 154–161. 40, 99
- JONES, R. & DIAZ, F. (2007). Temporal profiles of queries. *ACM Transactions on Information Systems*, **25**. 30, 106
- JONES, R. & KLICKNER, K.L. (2008). Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. of the 17th ACM Conference on Information and Knowledge Management*, 699–708. 70
- KAHLE, B. (2013). Wayback Machine: Now with 240,000,000,000 URLs, <http://blog.archive.org/2013/01/09/updated-wayback/>. 20
- KANG, I. & KIM, G. (2003). Query type classification for web document retrieval. In *Proc. of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 64–71. 35, 64
- KELLAR, M., WATTERS, C. & SHEPHERD, M. (2007). A field study characterizing web-based information-seeking tasks. *American Society for Information Science and Technology*, **58**, 999–1018. 27, 37, 67

REFERENCES

- KELLY, D. (2009). *Methods for evaluating interactive information retrieval systems with users*, vol. 3 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc. [27, 37](#)
- KITSUREGAWA, M., TAMURA, T., TOYODA, M. & KAJI, N. (2008). Socio-sense: a system for analysing the societal behavior from long term web archive. In *Proc. of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development*, 1–8. [1](#)
- KITTUR, A., CHI, E.H. & SUH, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proc. of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, 453–456. [40](#)
- KLEINBERG, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**, 604–632. [30](#)
- KONINKLIJKE BIBLIOTHEEK WA (2011). National library of the Netherlands. http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html, Accessed on March 2011. [158](#)
- KRAAIJ, W., WESTERVELD, T. & HIEMSTRA, D. (2002). The importance of prior probabilities for entry page search. In *Proc. of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 27–34. [30, 137, 163](#)
- KULTURARW3 (2011). National Library of Sweden. <http://www.kb.se/english/find/internet/websites/>, Accessed on March 2011. [158](#)
- KUNDER, M. (2011). WorldWideWebSize.com - The size of the World Wide Web. <http://www.worldwidewebsize.com/>. [60](#)
- LANDIS, J.R. & KOCH, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174. [70](#)
- LATIN AMERICAN WA (2011). Latin American Web Archiving Project (LAWAP), University of Texas at Austin. <http://lanic.utexas.edu/project/archives/>, Accessed on March 2011. [158](#)

REFERENCES

- LESKOVEC, J., KLEINBERG, J. & FALOUTSOS, C. (2007). Graph evolution: densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, **1**, 2. [124](#)
- LESKOVEC, J., BACKSTROM, L. & KLEINBERG, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. [154](#)
- LEWANDOWSKI, D. (2011). The retrieval effectiveness of search engines on navigational queries. *Aslib Proceedings*, **63**, 354–363. [117](#)
- LI, X. & CROFT, W.B. (2003). Time-based language models. In *Proc. of the 12th International Conference on Information and Knowledge Management*, 469–475. [30](#)
- LIBRARY OF CONGRESS WA (2011). Library of Congress. <http://www.loc.gov/webarchiving/>, Accessed on March 2011. [158](#)
- LIU, B. & ZHANG, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining Text Data*, 415–463. [154](#)
- LIU, T. (2009). *Learning to rank for information retrieval*, vol. 3 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc. [33](#), [34](#), [138](#)
- LIU, T., XU, J., QIN, T., XIONG, W. & LI, H. (2007). Letor: benchmark dataset for research on learning to rank for information retrieval. In *Proc. of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*. [32](#)
- LOPES, R., GOMES, D. & CARRIÇO, L. (2010). Web not for all: a large scale study of web accessibility. In *Proc. of the 2010 International Cross Disciplinary Conference on Web Accessibility*, 10. [7](#)
- LUCCHESE, C., ORLANDO, S., PEREGO, R. & SILVESTRI, F. (2007). Mining query logs to optimize index partitioning in parallel web search engines. In *Proc. of the 2nd International Conference on Scalable Information Systems*, 1–9. [91](#)

REFERENCES

- MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. [13](#), [28](#), [33](#), [42](#)
- MARKEY, K. (2007). Twenty-five years of end-user searching, part 1: research findings. *American Society for Information Science and Technology*, **58**, 1071–1081. [28](#), [39](#), [88](#), [99](#)
- MASANÈS, J. (2006). *Web Archiving*. Springer-Verlag New York Inc. [13](#), [19](#), [22](#)
- MASANÈS, J. (2011). LiWA news #3: Living web archives. http://liwa-project.eu/images/videos/Liwa_Newsletter-3.pdf. [20](#)
- MATTHEWS, M., TOLCHINSKY, P., BLANCO, R., ATSERIAS, J., MIKA, P. & ZARAGOZA, H. (2010). Searching through time in the New York Times. In *Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval*, 41–44. [154](#)
- METZLER, D., JONES, R., PENG, F. & ZHANG, R. (2009). Improving search relevance for implicitly temporal queries. In *Proc. of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 700–701. [31](#)
- MIRANDA, J. & GOMES, D. (2009a). Trends in web characteristics. In *Proc. of the 7th Latin American Web Congress*, 146–153. [3](#), [7](#), [55](#), [151](#)
- MIRANDA, J. & GOMES, D. (2009b). An updated portrait of the Portuguese web. In *14th Portuguese Conference on Artificial Intelligence (EPIA)*. [104](#)
- MONZ, C. (2004). Minimal span weighting retrieval for question answering. In *Proc. of the SIGIR 2004 Workshop on Information Retrieval for Question Answering*, 23–30. [137](#)
- MOTA, C. & SANTOS, D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o segundo HAREM*. Linguateca. [97](#)
- NDSA CONTENT WORKING GROUP (2012). Web archiving survey report. Tech. rep., National Digital Stewardship Alliance. [3](#), [48](#)

REFERENCES

- NETARKIVET.DK (2011). State and University Library. <http://netarkivet.dk/>, Accessed on March 2011. 158
- NEW ZEALAND WA (2011). National Library of New Zealand. <http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive>, Accessed on March 2011. 158
- NIU, J. (2012a). Functionalities of web archives. *D-Lib Magazine*, 18. 24, 110
- NIU, J. (2012b). An overview of web archiving. *D-Lib Magazine*, 18, 2. 16
- NORTH CAROLINA WA (2011). North Carolina State Archives and State Library of North Carolina. <http://webarchives.ncdcr.gov/>, Accessed on March 2011. 158
- NORWAY WA (2011). National Library of Norway. <http://www.nb.no/>, Accessed on March 2011. 158
- NTOULAS, A., CHO, J. & OLSTON, C. (2004). What's new on the web?: the evolution of the web from a search engine perspective. In *Proc. of the 13th International Conference on World Wide Web*, 1–12. 1, 125, 126
- NTUWAS (2011). National Taiwan University Library. <http://webarchive.lib.ntu.edu.tw/eng/default.asp>, Accessed on March 2011. 158
- NUNES, S., RIBEIRO, C. & DAVID, G. (2007). Using neighbors to date web documents. In *Proc. of the 9th Annual ACM International Workshop on Web Information and Data Management*, 129–136. 31, 130
- NUNES, S., RIBEIRO, C. & DAVID, G. (2008). Use of temporal expressions in web search. In *Proc. of the Advances in Information Retrieval, 30th European Conference on IR Research*, 580–584. 98
- OASIS (2011). National Library of Korea. http://www.oasis.go.kr/intro_new/intro_overview_e.jsp, Accessed on March 2011. 158

REFERENCES

- OZMUTLU, S., OZMUTLU, H. & SPINK, A. (2003). Multitasking web searching and implications for design. *American Society for Information Science and Technology*, **40**, 416–421. [87](#)
- PAGE, L., BRIN, S., MOTWANI, R. & WINOGRAD, T. (1998). The PageRank citation ranking: bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project. [30](#), [31](#)
- PAYNTER, G., JOE, S., LALA, V. & LEE, G. (2008). A year of selective web archiving with the web curator tool at the national library of New Zealand. *D-Lib Magazine*, **14**, 2. [15](#)
- PRESERVATION .ES (2011). National Library of Spain. <http://www.bne.es/es/LaBNE/PreservacionDominioES/>, Accessed on March 2011. [158](#)
- PWA (2011). Portuguese Web Archive, Foundation for National Scientific Computing. <http://www.archive.pt/>, Accessed on March 2011. [158](#)
- QIN, T., LIU, T., XU, J. & LI, H. (2010). LETOR: a benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, **13**, 346–374. [36](#)
- RADINSKY, K. & HORVITZ, E. (2013). Mining the web to predict future events. In *Proc. of the 6th ACM International Conference on Web Search and Data Mining*, 255–264. [1](#), [154](#)
- RADLINSKI, F. & JOACHIMS, T. (2005). Query chains: learning to rank from implicit feedback. In *Proc. of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 239–248. [29](#), [39](#), [70](#)
- RAS, M. & VAN BUSSEL, S. (2007). Web archiving user survey. Tech. rep., National Library of the Netherlands (Koninklijke Bibliotheek). [24](#), [25](#), [105](#)
- REYNOLDS, E. (2013). Web archiving use cases. Tech. rep., Library of Congress. [25](#)

REFERENCES

- RICHARDSON, M., PRAKASH, A. & BRILL, E. (2006). Beyond PageRank: machine learning for static ranking. In *Proc. of the 15th International Conference on World Wide Web*, 707–715. [30](#)
- RISSE, T. & PETERS, W. (2012). ARCOMEM: from collect-all ARchives to Community MEMories. In *Proc. of the 21st International Conference Companion on World Wide Web*, 275–278. [20](#)
- ROBERTSON, S. & ZARAGOZA, H. (2009). *The probabilistic relevance framework*, vol. 3 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc. [113](#), [137](#), [139](#), [161](#), [162](#)
- ROBERTSON, S.E. & JONES, K.S. (1976). Relevance weighting of search terms. *American Society for Information Science*, **27**, 129–146. [29](#)
- ROBERTSON, S.E., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M.M. & GATFORD, M. (1995). Okapi at TREC-3. In *Proc. of the 3rd Text REtrieval Conference*, 109–126. [29](#), [32](#)
- ROSE, D. & LEVINSON, D. (2004). Understanding user goals in web search. In *Proc. of the 13th International Conference on World Wide Web*, 13–19. [26](#), [70](#), [80](#)
- RUSSELL, D. & GRIMES, C. (2007). Assigned tasks are not the same as self-chosen web search tasks. In *Proc. of the 40th Hawaii International Conference on System Sciences*, 83–91. [73](#)
- SALAHELDEEN, H. & NELSON, M. (2012). Losing my revolution: how many resources shared on social media have been lost? *Theory and Practice of Digital Libraries*, 125–137. [1](#)
- SALTON, G. (1986). *Introduction to modern information retrieval*. McGraw-Hill Computer Science Series. [161](#), [162](#)
- SALTON, G. & BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval. **24**, 513–523. [29](#)

REFERENCES

- SANDERSON, M. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 162–169. [38](#)
- SCHÖLKOPF, B. & SMOLA, A. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press. [33](#)
- SHIOZAKI, R. & EISENSCHITZ, T. (2009). Role and justification of web archiving by national libraries: a questionnaire survey. *Journal of Librarianship and Information Science*, **41**, 90–107. [15](#)
- SILVERSTEIN, C., MARAIS, H., HENZINGER, M. & MORICZ, M. (1999). Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, vol. 33, 6–12. [89, 91](#)
- SLOVENIAN WA (2011). Historical Archives of Ljubljana. <http://www.zal-1j.si/>, Accessed on March 2011. [158](#)
- SNOW, R., O'CONNOR, B., JURAFSKY, D. & NG, A. (2008). Cheap and fast – but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 254–263. [41](#)
- SONG, F. & CROFT, W. (1999). A general language model for information retrieval. In *Proc. of the 8th International Conference on Information and Knowledge Management*, 316–321. [29](#)
- SPINK, A., OZMUTLU, S., OZMUTLU, H.C. & JANSEN, B.J. (2002). U.S. versus European web searching trends. *SIGIR Forum*, **36**, 32–38. [98](#)
- SRIVASTAVA, J., COOLEY, R., DESHPANDE, M. & TAN, P. (2000). Web usage mining: discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, **1**, 23. [27](#)

REFERENCES

- TAHMASEBI, N., GOSSEN, G. & RISSE, T. (2012). Which words do you remember? Temporal properties of language use in digital archives. In *Proc. of the 2nd International Conference on Theory and Practice of Digital Libraries*, 32–37. [3](#), [124](#)
- TAO, T. & ZHAI, C. (2007). An exploration of proximity measures in information retrieval. In *Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 295–302. [29](#), [162](#), [163](#)
- TASMANIAN WA (2011). Our Digital Island, State Library of Tasmania. <http://odi.statelibrary.tas.gov.au/>, Accessed on March 2011. [158](#)
- TAYLOR, M., GUIVER, J., ROBERTSON, S. & MINKA, T. (2008). SoftRank: optimizing non-smooth rank metrics. In *Proc. of the International Conference on Web Search and Web Data Mining*, 77–86. [37](#)
- TEEVAN, J., ALVARADO, C., ACKERMAN, M. & KARGER, D. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proc. of the SIGCHI Conference on Human factors in Computing Systems*, 422–429. [27](#), [67](#)
- TEEVAN, J., DUMAIS, S., LIEBLING, D. & HUGHES, R. (2009). Changing how people view changes on the web. In *Proc. of the 22nd Annual ACM Symposium on User Interface Software and Technology*, 237–246. [81](#)
- THOMAS, A., MEYER, E.T., DOUGHERTY, M., VAN DEN HEUVEL, C., MADSEN, C. & WYATT, S. (2010). Researcher engagement with web archives: challenges and opportunities for investment. Tech. rep., Joint Information Systems Committee (JISC). [25](#)
- TOFEL, B. (2007). 'Wayback' for accessing web archives. In *Proc. of the 7th International Web Archiving Workshop*. [22](#)
- UK Gov WA (2011). The National Archives. <http://www.nationalarchives.gov.uk/webarchive/>, Accessed on March 2011. [158](#)

REFERENCES

- UK WA (2011). British Library. <http://www.webarchive.org.uk/ukwa/>, Accessed on March 2011. 158
- UNESCO (2003). Charter on the preservation of digital heritage. Adopted at the 32nd session of the General Conference of UNESCO, http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf. 1, 53
- UNITED STATES SECURITIES AND EXCHANGE COMMISSION (2010). Form 10-K, Google Inc. <http://www.sec.gov/Archives/edgar/data/1288776/000119312511032930/d10k.htm>. 50
- UNIVERSITY OF MICHIGAN WA (2011). Bentley Historical Library. <http://bentley.umich.edu/uarphome/webarchives/webarchive.php>, Accessed on March 2011. 158
- VAN DE SOMPTEL, H., NELSON, M.L., SANDERSON, R., BALAKIREVA, L.L., AINSWORTH, S. & SHANKAR, H. (2009). Memento: time travel for the web. *CoRR*, **abs/0911.1112**. 20
- VON AHN, L. & DABBISH, L. (2004). Labeling images with a computer game. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 319–326. 40
- VON AHN, L. & DABBISH, L. (2008). Designing games with a purpose. *Communications of the ACM*, **51**, 58–67. 40
- VOORHEES, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, **36**, 697–716. 38, 39
- VOORHEES, E. (2009). Topic set size redux. In *Proc. of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 806–807. 112
- VOORHEES, E. & HARMAN, D. (1999). Overview of the eighth text retrieval conference (TREC-8). In *Proc. of the 8th Text REtrieval Conference*, vol. 8, 1–24. 39

REFERENCES

- VOORHEES, E. & HARMAN, D. (2005). *TREC: experiment and evaluation in information retrieval*. MIT Press. 38, 101, 102
- WA OF ČAČAK (2011). Public Library Čačak. <http://digital.cacak-dis.rs/english/web-archive-of-cacak/>, Accessed on March 2011. 158
- WA PACIFIC ISLANDS (2011). Web Archiving Project for the Pacific Islands at the University of Hawaii at Manoa Library. <http://library.manoa.hawaii.edu/research/archiveit/>, Accessed on March 2011. 158
- WA SINGAPORE (2011). National Library Board Singapore. <http://was.nl.sg/>, Accessed on March 2011. 158
- WA SWITZERLAND (2011). Swiss National Library. http://www.nb.admin.ch/nb_professionnel/01693/index.html?lang=en, Accessed on March 2011. 158
- WA TAIWAN (2011). National Central Library of Taiwan. <http://webarchive.ncl.edu.tw/nclwa98Front/>, Accessed on March 2011. 158
- WARP (2011). Web Archiving Project, National Diet Library. <http://warp.da.ndl.go.jp/search/>, Accessed on March 2011. 158
- WEBARCHIV (2011). National Library of the Czech Republic. <http://en.webarchiv.cz/>, Accessed on March 2011. 158
- WEBER, I. & CASTILLO, C. (2010). The demographics of web search. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 523–530. 28
- WEB@RCHIVE (2011). Austrian National Library. <http://www.onb.ac.at/ev/about/webarchive.htm>, Accessed on March 2011. 158
- WEIKUM, G., NTARMOS, N., SPANIOL, M., TRIANTAFILLOU, P., BENCZUR, A.A., KIRKPATRICK, S., RIGAUX, P. & WILLIAMSON, M. (2011). Longitudinal analytics on web archive data: it's about time! In *Proc. of the 5th Conference on Innovative Data Systems Research*, 199–202. 4, 20

REFERENCES

- XU, J. & LI, H. (2007). AdaRank: a boosting algorithm for information retrieval. In *Proc. of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 391–398. [34](#), [138](#)
- XUE, G., YANG, Q., ZENG, H., YU, Y. & CHEN, Z. (2005). Exploiting the hierarchical structure for link analysis. In *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 186–193. [30](#)
- YAMAMOTO, Y., TEZUKA, T., JATOWT, A. & TANAKA, K. (2007). Honto? Search: estimating trustworthiness of web information by search results aggregation and temporal analysis. *Advances in Data and Web Management*, 253–264. [1](#)
- YEH, J., LIN, J., KE, H. & YANG, W. (2007). Learning to rank for information retrieval using genetic programming. In *Proc. of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*. [33](#)
- YU, P.S., LI, X. & LIU, B. (2004). On the temporal dimension of search. In *Proc. of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, 448–449. [32](#)
- ZARAGOZA, H., CRASWELL, N., TAYLOR, M., SARIA, S. & ROBERTSON, S. (2004). Microsoft Cambridge at TREC-13: web and hard tracks. In *Proc. of the 13th Text REtrieval Conference*. [29](#)
- ZOBEL, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 307–314. [39](#)
- ZOBEL, J. & MOFFAT, A. (2006). Inverted files for text search engines. *ACM Computing Surveys (CSUR)*, **38**, 6. [16](#)