Technical Report

Evaluating Web Archive Information Retrieval Systems

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

June 29, 2012

Foundation for National Scientific Computing
1708-001 Lisbon, Portugal

Evaluating Web Archive Information Retrieval Systems

Miguel Costa[1,2] and Mário J. Silva[2]
(miguel.costa@fccn.pt, mjs@di.fc.ul.pt)

[1] Foundation for National Scientific Computing, Lisbon, Portugal
[2] University of Lisbon, Faculty of Sciences, LaSIGE, Lisbon, Portugal

Abstract. The information published on the web is rapidly vanishing along with our collective memory. At least 77 web archives have been developed to cope with the web's transience problem, but despite their technology having achieved a good maturity level, the retrieval effectiveness they provide still presents unsatisfactory results. In this work, we propose an evaluation methodology based on a list of requirements compiled from characterizations of web archives and their users. The methodology includes the design of a test collection and the selection of evaluation measures to support realistic and reproducible experiments. The test collection enabled, for the first time, to measure the effectiveness of state-of-the-art IR technology employed in web archives. Results confirm that the quality of results returned to the users is poor. However, we show how to combine temporal features, along with the regular topical features, to improve the search effectiveness on web archives. The test collection is available to the research community.

1 Introduction

Every day millions of web documents become inaccessible. Some contain unique information that might become as valuable as ancient manuscripts are today. For instance, the speech of a president after winning an election or the reasons that justify the imminent invasion of a foreign country. Together, these documents form a comprehensive picture of our cultural, commercial, scientific and social history, expressed by all kinds of people. It is therefore important to preserve and make these data accessible, not only for historical research [1], but also to improve current technology, such as assessing the trustworthiness of statements [2], detecting web spam [3] or improving web information retrieval (IR) [4].

Recently, UNESCO endorsed the Universal Declaration on Archives[1], which states that *"archives play an essential role in the development of societies by safeguarding and contributing to individual and community memory."* At least 77 initiatives[2] undertaken by national libraries, national archives and consortia of organizations are archiving parts of the web to cope with this problem. In total, more than 181 billion web documents (6.6 PB) are already archived and these numbers, as well as their historic interest, are growing as time passes [5].

[1] see http://www.ica.org/6573/reference-documents/universal-declaration-on-archives.html
[2] see http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

The challenge now is how to make historical analysis possible on all the data that has been accumulated over the years.

Full-text search has become the dominant form of finding web information, as notoriously seen in online search engines. 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 users' queries. However, the poor quality of search results still remains as one of the greatest obstacles in the way of turning web archives into an usable source of information. Users have to spend too much time and effort exploring retrieved documents in order to satisfy their information needs. As the amount of archived data continues to grow, this problem only tends to aggravate.

The quality of search results greatly depends on the availability of suitable evaluation methodologies and test collections. These resources have been a driver of research and innovation in IR, which have been employed throughout the last decades to validate many ideas [6]. They enable 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) facilitate and foster research under a set of controlled conditions. Unfortunately, existing evaluation methodologies and test collections from evaluation campaigns, such as TREC, are not useful for web archives, because they have different task goals and characteristics. For instance, existing collections do not have a temporal dimension, where each document may have several versions throughout time and their relevance depends of the user's period of interest.

In this work we propose an evaluation methodology to measure the effectiveness of web archive information retrieval (WAIR) systems. We believe this methodology along with the test collection created to support it are essential pieces of technology to improve the quality of results. The methodology takes the findings of recent characterizations on web archives and their users in consideration, which is a requirement to provide reliable results tailored for the users' information needs. We demonstrate the methodology's usefulness through an experiment where we measured, for the first time, the search effectiveness of web archives using state-of-the-art methods. Then, we significantly improved the observed effectiveness by exploring temporal features intrinsic to web archives.

The remainder of this paper is organized as follows. In Section 2, we cover the related work. In Section 3, we describe the web archive characteristics that guide the design of the evaluation methodology proposed in Section 4. In Section 5, we present a case study applying the methodology and Section 6, the obtained results. Section 7, finalizes with the conclusions.

2 Related Work

2.1 Web Archives Access

Much of the current effort on web archive development focuses on acquiring, storing, managing and preserving data [7]. However, this is just the beginning.

The data must be accessible. Recently, 82% of the European web archives considered the improvement of access tools a high priority [8]. 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. A recent survey reported that 89% of the world-wide web archives support this type of access [5]. However, this type of 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, for instance by category or theme, which was shown to be provided by 79% of web archives. Full-text search has become the dominant form of information access, specially in web search systems, such as Google, which has 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 [9] and the most used when supported [10]. Even with the high computational resources required for this purpose, 67% of world-wide web archives surveyed support full-text search for at least a part of their collections [5]. In another survey of European web archives this percentage is 70% [8]. As a result, in this work we focus on full-text search.

The large majority of web archives that support full-text search are based on the Lucene search engine[3] or extensions of Lucene to handle web archive extensions, such as NutchWAX[4]. The results of these web archives are visibly poor and frequently mentioned by the stakeholders as unsatisfactory [5]. Cohen et al. showed that the out-of-the-box Lucene produces low quality results, with a MAP (Mean Average Precision) of 0.154, which is less than half when compared with the best systems participating in the TREC Terabyte track [11].

2.2 IR Evaluations

IR evaluations straddle a spectrum between two opposite, but complementary views: a user-centered and a system-centered [12]. The goal of user-centered evaluations is to quantify whether people can use a system to retrieve relevant documents. These evaluations provide rich qualitative data about the users' interactions with the system, for instance, from experiments with users in a laboratory [13] or in their natural environment (in-situ) [14]. The goal of system-centered evaluations is to quantify whether a system retrieves relevant documents, independently of how well users interact with it. The most popular example is the Cranfield paradigm established in the 1960s by Cleverton. This paradigm defines the creation of test collections for evaluating retrieval results composed by three parts: (1) a **corpus** representative of the items (often documents) that will be encountered in a real search environment; (2) a set of **topics** describing users' information needs; and (3) **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. Our proposed methodology follows the Cranfield paradigm, but it is an extension for web archives.

[3] see `http://lucene.apache.org/`

[4] see `http://archive-access.sourceforge.net/projects/nutch/`

3 Web Archive Characteristics

In this Section we outline the characteristics and considerations that a representative test collection should have to provide an accurate simulation of the users' search behavior in an operational web archive.

3.1 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, we found some tendencies that a corpus should include:

selective and broad national crawls. 80% of the 42 world-wide web archive initiatives surveyed, exclusively hold content related to their country, region or institution [5]. All initiatives performed selective crawling, for instance, focusing in one sub-domain or topic. These collections are narrower, but deeper, trying to crawl all about the topic. 26% of the initiatives also performed broad crawling, including all documents hosted under a country code top-level domain or geographical location. These collections are wider, but shallower. In another survey of European web archives, 71% of them operate selective crawls and 23% broad domain crawls [8].

a 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 genre of document also influences the number of versions. For instance, newspapers have a higher change rate, while scientific articles tend to be static for long periods.

a diverse set of media types. The characterization of web collections shows that all 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 that comprise over 95% of all web contents [15].

a volume of data between 1TB and 100TB. 81% of web archive collections have a volume of data smaller than 100TB [5]. The predominant volume of data is between 1TB and 10TB (31%) or between 10TB and 100TB (31%).

between 100 million and 1 billion documents. 78% of web archive collections contain less than 1 billion documents (i.e. files) [5]. The predominant number of documents is between 100 million and 1 billion (43%).

a large temporal 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, 58% of the web archives contain collections up to 7 years old [5]. The corpus should have a large time span to not bias future WAIR technology to a specific period when some design patterns and technologies prevailed.

3.2 Search Topics

The evaluation of an information system, such as a web archive, must take into account the characteristics and needs of the user community. Characterizations of web archive users exhibit some tendencies that topics should include:

generic use cases. Despite some professional categories being more prone to use a web archive, such as historians, the user population is generic. There are numerous everyday life use cases that web archives can fulfill as exemplified by Ras and Bussel [9] and as log analysis have shown [16].

navigational and informational queries. The information needs of web archive users are mostly navigational, i.e. users intend to see how a web page or site was in the past or how it evolved throughout time [16]. 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.

1/3 of queries restricted by date range. Despite all users' 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 these queries [10].

queries without temporal clues. Only 3% of queries have expressions that could indicate a temporal dependent intent, such as *Euro 2004* [10].

short queries, each with 1 to 3 terms. A typical full-text session is composed by 1 or 2 queries, having each 1 to 3 terms [10]. 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.

The studies referred on the last four tendencies were only conducted on the Portuguese Web Archive (PWA). However, we believe that they are general, because these results also show that users from the PWA and a Portuguese web search engine have a similar search behavior [17]. Thus, the differences between both systems do not affect the way users search in them. In turn, the results compiled about web search engine users across the U.S. and Europe, including Portugal, were also similar [18, 17]. Thus, the users' distinct language, vocabulary and culture have a small impact in the users' search behavior. In conclusion, despite some nuances, it seems that users from both types of systems and different countries, have similar search behaviors.

3.3 Relevance Criteria

A document d collected at n periods has n archived versions $\{v^d_{t_1}, ..., v^d_{t_n}\}$. 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 <URL, timestamp>. Second, the relevance is bi-dimensional. Each version has associated a temporal relevance along with a topical relevance.

Topical relevance The navigational and informational queries issued to web archives, represent different information needs that require different criteria when judging topical relevance. A navigational query intends to find an archived document d for some end. Hence, any version v^d_t of document d has a high topical relevance and any version v^x_t of a document x is measured according to how well

x matches d. For instance, a version v_t^x may be partially relevant if x is a good alternative to d. For informational queries, the topical relevance of a version v_t^x is measured according to how well it describes the searched topic in detail. Hence, since all versions v_t^x of a document x may be different, they all may have different topical relevance. Knowing this, we can propagate the topical relevance between versions of the same document. Only one version v_t^x of each document x needs to be assessed for navigational queries. All the other versions v_t^x receive the same relevance degree. The same is valid for informational queries, but only when the content of versions v_t^x is very similar (e.g. near-duplicates).

Temporal relevance The relevance of archived versions depends also on the period of interest of users. Users explicitly express a date range that acts as a filter and excludes all versions with timestamps outside this range. This is the users' expected behavior, so we assume that the filtered versions are temporally non-relevant. All the others are considered equally relevant in the temporal dimension, because in web archives: (1) there is not a preference by a period within the date range (e.g. older or newer). This is different from some applications, such as news search engines, where recent and updated information is preferred [19]; (2) highly relevant documents for a topic may exist throughout all search period, despite being known that there are periods that tend to concentrate more relevant documents [20].

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

- topical relevance degree for a navigational topic.
- topical relevance degree for an informational topic if their content is very similar (e.g. near-duplicates).
- temporal relevance degree for a topic if t_i and t_j are both within or without the search interval.

In case of these two versions $v_{t_i}^d$ and $v_{t_j}^d$ have the same topical and temporal relevance for a topic u, we define them as **redundant** for u.

4 Evaluation Methodology

Our proposed methodology depicted in Figure 1 extends the Cranfield paradigm to support the ad hoc retrieval task for web archives. The methodology has the following steps:

1. characterization of web archives along with their collections and users. With the knowledge compiled in the previous Section, we are able to build a representative test collection to draw valid conclusions.
2. selection of a representative corpus of the documents that will be encountered in a real search environment. The corpus must fit the tendencies observed in world-wide web archives, such as their collections' size and temporal span.

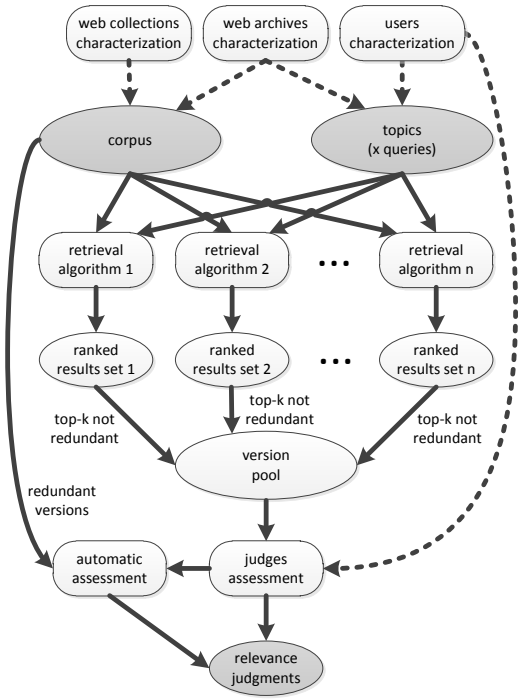

Fig. 1: Methodology for building a WAIR test collection.

3. selection of topics based on the users' information needs and search behavior. 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 to match and rank document versions for each topic. These algorithms should contemplate topical and temporal features to explore both search dimensions.
5. aggregation of all top-k versions returned by each retrieval algorithm for each topic, ignoring the redundant 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. In the end, we obtain the version pool.
6. manual assessment of all versions in the pool by a set of judges according to the user's information need defined in each topic. The set of judges assign a topical degree of relevance taking into account the goals 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 version $v_{t_i}^d$ manually assessed. Each version $v_{t_j}^d$ of d receives the same topical relevance degree given to $v_{t_i}^d$ if their implicit relevance can be inferred (i.e. if they are redundant).

4.1 Evaluation Metrics

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. The measures should consider the relevant versions ranked ahead of the non-relevant and the dependency between redundant versions. It is certainly unpleasant for the user to see multiple versions of the same document in a results page. If this is the case, the typical web archive user interface has associated to each result a link to show all versions of the respective document in a chronological view.

We have two choices to model this dependency. The first, is to design or adopt a measure, such as α-nDCG, that penalizes the relevance of redundant versions [21]. The second, is to use a standard measure after ignoring the redundant versions. We 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 criteria, than for a bi-criteria where relevance is traded-off with diversity. Notice that, search result diversification is an NP-hard optimization problem [22]. As a drawback, the WAIR systems should collapse these redundant versions before presenting the results to the users, as they already do.

Concluding, we promote diversity in search results by ignoring easily identifiable redundant versions and then, applying a standard evaluation measure. Any measure that can make use of these 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. Otherwise, if the number of WAIR systems is small, it is likely that a significant number of relevant versions will not be found beyond rank k and the results biased.

5 Test Collection Construction

In this Section we present a case study to empirically validate the proposed evaluation methodology. We detail the design of a test collection for the Portuguese Web Archive (PWA).

5.1 Corpus Selection

Our corpus is composed by six collections of the Portuguese web, broadly considered the web subset of interest to the Portuguese. Since the goal is to create a corpus representative of the documents that are encountered in a real search environment, we only included collections indexed and searchable through the public access given by the PWA at `http://archive.pt`. The corpus' main characteristics are detailed in Table 1, showing a significant heterogeneity in age, size and type. They are the result of different types of crawling that amount to around 256 million documents, which consist of 6.2 TB compressed in ARC format (8.9 TB uncompressed) [23]. This corpus contains some of the first documents published in the Portuguese web in 1996 and go until 2009. It includes all type of

#	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 Portuguese National Library
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 1: Web collections that compose the corpus.

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. We consider this corpus sufficiently comprehensive and representative, but not too large to discourage its use. The ClueWeb09[5] 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 our 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.

5.2 Search Topics Selection

We randomly sampled queries from the PWA's query log that fit the general search patterns presented in Section 3. From these queries we created 50 navigational topics, where one third have temporal 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 [24]. We tried to select topics of different difficulties 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. We 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 these capture the real and diverse users' 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 identified by a number 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 date range of interest to the user. These two fields are the ones submitted to the WAIR system. The description specifies the user's 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:

[5] see `http://lemurproject.org/clueweb09/`

```
<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.

5.3 Retrieval

WAIR system The corpus was indexed by the PWA IR system, which was 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 topic's query: (1) versions are topically matched by comparing with the query's terms; (2) matched versions are temporally filtered according to the topic's search period; (3) the topically matched and temporally not filtered versions are ranked by topical and temporal similarity.

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. We implemented 9 models. The first was the Lucene's term-weighting function[6], which is computed over 5 fields (anchor, content, title, hostname, url) with different weights. The second was a small variation of this function used in NutchWAX, with a different normalization by field length. These two models can be considered the state-of-the-art of IR in web archives, since most of the IR technology is based on the Lucene search engine and NutchWAX. As a baseline and third model, we selected the Okapi BM25 [25].

We implemented two time-aware models that: (1) score higher on documents with more versions; (2) score higher on documents with a larger time span between the first and last archived versions. Both are defined by the same function:

$$f(v_t^d) = \frac{\log_{10}(x)}{\log_{10}(y)} = \log_y(x) \tag{1}$$

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. y is the maximum possible x for normalization. Each of these functions, $f1$, was linearly combined with the NutchWAX's term-weighting function, $f2$, using three different weights (0.1, 0.25, 0.5). That is, functions f_1 and f_2 were combined in three models: (1) $0.1{*}f_1 + 0.9{*}f_2$; (2) $0.25{*}f_1 + 0.75{*}f_2$; and (3) $0.5{*}f_1 + 0.5{*}f_2$. All functions were normalized to a value between 0 and 1. We generally denote these linearly combined models by TVersions and TSpan.

[6] see `http://lucene.apache.org/java/2_9_0/api/all/org/apache/lucene/search/Similarity.html`

```
1. Imagine that to find the page of:
     José Saramago, Nobel Prize-Winning Writer in 1998.
2. You submit the query:
     josé 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):
```

Fig. 2: Form to assess navigational topics.

5.4 Relevance Assessment

Manual Assessment Three people, including the topics creator, assessed each of the 2 065 <URL, timestamp, topic> triplets aggregated in the version pool. They followed strict guidelines and document versions were presented in a random order for assessors to not be able to know from which algorithm they were retrieved and at which rank they were shown. Figure 2 shows the form used for collecting the relevance assessments for the navigational topics. We used a three-level scale of relevance from which the judges could choose.

The usefulness of the test collection depends heavily on the reliability of relevance judgments. Hence, we analyzed their level of consensus. The inter agreement between judges measured by Fleiss's 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 agreement, lending confidence to the judgment quality. These inter agreement values are inline with the ones of TREC judges [26].

Automatic Assessment The relevance assessments is the most time-consuming part of creating a test collection. Hence, we took advantage of the characteristics described in Section 3.3 to automatically assess 267 822 versions following the seventh step of the proposed methodology. This is a huge increase when compared with the number of versions manually assessed.

For each version manually assessed, we used the PWA IR system to find all the redundant versions of the same document within the search period specified in that topic. Then, we assigned the same topical relevance degree to all of them.

6 Experiments & Results

Table 2 presents the results of the ranking models described above. The bold entries indicate the best result for 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.

We can now determine, for the first time, how effective is the IR technology typically used in web archives. For instance, the Lucene and NutchWAX's results

Metric	time-unaware			time-aware	
	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 $p<0.05$ against NutchWAX

Table 2: Results for the tested ranking models.

achieved a S@1 value of 0.28 and 0.32, which is less than half of the best results achieved in the 2004 Web Track, that was a S@1 of 0.65 [27]. Despite these values are not directly comparable due to the different test collections, a great deal of work is still necessary to reach the S@1 value of 0.84 provided by Google [28].

An interesting 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's models, for a statistical significance level of 0.01 using a 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 tested with a multi-version corpus as the one we built.

6.1 Topic difficulty

Figure 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 some topics, such as the 11 and 14, which present very poor results. There are several reasons for this. One is that the query terms did not match the searched document. For instance, the query used was *Dona Maria Segunda (second) Theatre*, but the text versions and link references only contained the terms *Dona Maria II Theatre*. A query expansion is necessary to improve the results in these cases.

6.2 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 many relevant documents and

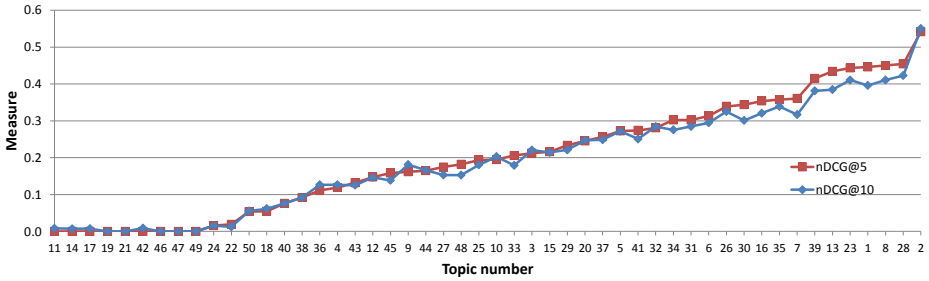

Fig. 3: Navigational topics sorted by the average of the 9 tested ranking models.

is biased toward that system. Nevertheless, 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 [29].

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

7 Conclusions & Future Work

Billions of past web documents containing our history are currently archived. However, their access is still in an early stage, preventing users from unfolding the full potential of web archives. Other IR domains showed us that the quality of search results depends greatly on the availability of suitable resources for evaluation. These resources are missing for WAIR systems, which explains why no evaluation had ever been conducted. In this work we describe the methodology employed in a test collection based evaluation for WAIR systems. We faced several challenges from the lack of knowledge about the users to the relevance assessment of archived versions of documents. In the end, we were able to measure, for the first time, the effectiveness of state-of-the-art WAIR technology. As anticipated, the results were poor, which proves the need of a common evaluation framework to foster research in web archives. We expect that our research may lead to a novel IR task in a major evaluation campaign, such as TREC.

We 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 they existed, the more likely it is of being important. We achieved statistically significant improvements in both models over the state-of-the-art IR typically used in web archives. This is just the first step in leveraging temporal information to improve WAIR systems. There is a large margin for progress that could also be applied to web search engines. In the future, we intend to create a dataset for *learning-to-rank* experiments from our test collection, to combine temporal evidences implicitly hidden in the corpus and query matches. The automatic

assessment obtained with our methodology provide a fast mean of generating vast amounts of labeled data for machine learning optimization.

References

1. Kitsuregawa, M., Tamura, T., Toyoda, M., Kaji, N.: 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. (2008) 1–8
2. Yamamoto, Y., Tezuka, T., Jatowt, A., Tanaka, K.: Honto? search: Estimating trustworthiness of web information by search results aggregation and temporal analysis. Advances in Data and Web Management (2007) 253–264
3. Chung, Y., Toyoda, M., Kitsuregawa, M.: 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. (2009) 9–16
4. Elsas, J., Dumais, S.: Leveraging temporal dynamics of document content in relevance ranking. In: Proc. of the 3rd ACM Inter. Conference on Web Search and Data Mining. (2010) 1–10
5. Gomes, D., Miranda, J., Costa, M.: A survey on web archiving initiatives. In: Proc. of the International Conference on Theory and Practice of Digital Libraries. (2011)
6. Voorhees, E., Harman, D.: TREC: Experiment and evaluation in information retrieval. MIT Press (2005)
7. Masanès, J.: Web Archiving. Springer-Verlag New York Inc. (2006)
8. Foundation, I.M.: Web archiving in Europe. Technical report, CommerceNet Labs (2010)
9. Ras, M., van Bussel, S.: Web archiving user survey. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek) (2007)
10. Costa, M., Silva, M.J.: Characterizing search behavior in web archives. In: Proc. of the 1st International Temporal Web Analytics Workshop. (2011)
11. Cohen, D., Amitay, E., Carmel, D.: Lucene and Juru at Trec 2007: 1-million queries track. In: Proc. of the 16th Text REtrieval Conference. (2007)
12. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Volume 3 of Foundations and Trends in Information Retrieval. Now Publishers Inc. (2009)
13. Aula, A., Khan, R.M., Guan, Z.: How does search behavior change as search becomes more difficult? In: Proc. of the 28th International Conference on Human Factors in Computing Systems. (2010) 35–44
14. Kellar, M., Watters, C., Shepherd, M.: A field study characterizing Web-based information-seeking tasks. American Society for Information Science and Technology **58**(7) (2007) 999–1018
15. Baeza-Yates, R., Castillo, C., Efthimiadis, E.: Characterization of national web domains. ACM Transactions on Internet Technology **7**(2) (2007)
16. Costa, M., Silva, M.J.: Understanding the information needs of web archive users. In: Proc. of the 10th International Web Archiving Workshop. (2010) 9–16
17. Costa, M., Silva, M.J.: A search log analysis of a Portuguese web search engine. In: Proc. of the 2nd INForum - Simpósio de Informática. (2010) 525–536
18. Jansen, B., Spink, A.: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing and Management **42**(1) (2006) 248–263
19. Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., Buchner, K., Liao, C., Diaz, F.: Towards recency ranking in web search. In: Proc. of the 3rd ACM International Conference on Web Search and Data Mining. (2010) 11–20
20. Jones, R., Diaz, F.: Temporal profiles of queries. ACM Transactions on Information Systems (TOIS) **25**(3) (2007)
21. Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proc. of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. (2008) 659–666
22. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proc. of the 2nd ACM International Conference on Web Search and Data Mining. (2009) 5–14
23. Burner, M., Kahle, B.: The Archive File Format
24. Voorhees, E.: Topic set size redux. In: Proc. of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. (2009) 806–807
25. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework. Volume 3 of Foundations and Trends in Information Retrieval. Now Publishers Inc. (2009)
26. Al-Maskari, A., Sanderson, M., Clough, P.: Relevance judgments between trec and non-trec assessors. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. (2008) 683–684
27. Craswell, N., Hawking, D.: Overview of the TREC-2004 Web Track. NIST Special Publication (2005) 500–261
28. Lewandowski, D.: The retrieval effectiveness of search engines on navigational queries. In: Aslib Proceedings. Volume 63. (2011) 354–363
29. Blanco, R., Halpin, H., Herzig, D., Mika, P., Pound, J., Thompson, H., Tran Duc, T.: Repeatable and reliable search system evaluation using crowdsourcing. In: Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information. (2011) 923–932