Search the Past with the Portuguese Web Archive

Daniel Gomes
daniel.gomes@fccn.pt

David Cruz
david.cruz@fccn.pt

João Miranda
joao.miranda@fccn.pt

Miguel Costa
miguel.costa@fccn.pt

Simão Fontes
simao.fontes@fccn.pt

Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal

ABSTRACT

The web was invented to quickly exchange data between scientists, but it became a crucial communication tool to connect the world. However, the web is extremely ephemeral. Most of the information published online becomes quickly unavailable and is lost forever. There are several initiatives worldwide that struggle to archive information from the web before it vanishes. However, search mechanisms to access this information are still limited and do not satisfy their users who demand performance similar to live-web search engines.

This demo presents the Portuguese Web Archive, which enables search over 1.2 billion files archived from 1996 to 2012. It is the largest full-text searchable web archive publicly available [17]. The software developed to support this service is also publicly available as a free open source project at Google Code, so that it can be reused and enhanced by other web archivists. A short video about the Portuguese Web Archive is available at `vimeo.com/59507267`. The service can be tried live at `archive.pt`.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Distributed systems

Keywords

Web Archiving; Digital Preservation; Temporal Search

1. INTRODUCTION

Human knowledge has been incrementally built for thousands of years. The new generations augment knowledge transmitted by the previous ones. Inventions such as writing, press and, recently, the web, deeply improved this process. However, after a short period of time, the information published on the web becomes unavailable and commonly is lost forever. Ntoulas et al. estimated that only 20% of the pages available today will still be available one year from now [14]. Besides loosing important scientific and historical information, web transience causes that common people are loosing their memories as individuals. Everyday, people take photos and share them directly and exclusively on the web without having the most elementary preservation concerns. As consequence, in the future many people will have difficulties in showing portraits of their ancestors or memories.

The web lacks preservation mechanisms. For centuries, organizations such as archives and libraries, ensured the preservation of information published on printed media for future generations. Since 1996, several web archiving initiatives were created worldwide [9]. Web archives acquire, store, preserve and provide access to information published on the web across time, which also includes contents created before the digital era, that were digitized and published online. These contents include official documents, such as those kept in libraries or museums, but also, commercials, games or pictures that are valuable descriptions of recent history.

Archiving data from the web and preserving it is not enough to make web archives useful for societies. Historical information must be searchable and web users expect a performance similar to the one provided by live-web search engines [15]. Efficient search mechanisms over archived web contents enable the response to numerous everyday life use cases. For instance, web users recovering from broken links, journalists looking for information to document articles, software engineers searching for technical manuals to fix legacy systems, webmasters recovering past versions of their site's pages or historians studying web documents as they do for printed information. However, search engine technology cannot be directly applied to web archives. Search engines provide access to the present and web archives to the past. Search engines process online contents hosted on their original servers. There is no concern with content preservation across time. The information they gathered from the web is meant to be permanently updated to be as fresh as possible. On the other hand, web archives address offline contents, frequently in obsolete formats, that must be preserved and reproduced as close as possible to their original layout. Searching web archives has always a temporal dimension that must be addressed on queries and results.

The Portuguese Web Archive (PWA) aims to preserve web contents of interest to the Portuguese community. It was based on the Archive-access project tools [11], which are used by most web archives worldwide [9]. However, we observed that these tools did not fulfill our users requirements at several levels.

Web archives obtain information from the live web for preservation but when they begin their activities they do not hold historical data so they must obtain it from third-

WWW'13, May 13–May 17, 2013, Rio de Janeiro, Brasil.
ACM .

parties, such as the Internet Archive or site backups. Note that the integration of historical web collections is crucial to populate the growing number of web archives [9]. Full-text search in the Archive-access project is supported by the NutchWAX component, whose core is the Nutch web search engine [4]. However, NutchWAX did not support the indexing of historical collections that contain several versions gathered from the same URL across time. The Archive-access presentation middleware and user interfaces were also inadequate to support temporal search or internationalization. Thus, we researched and developed a new web archive search engine. Its software was shared as a free open source project available at `code.google.com/p/pwa-technologies/`. In December 2012, the service provided public access to 1.2 billion (10^9) files archived from the web from 1996 to 2012. The search service over the PWA is publicly available at `archive.pt`.

2. RELATED WORK

Web archives face many challenges related to scalability and information overload because they accumulate previous documents and indexes, unlike web search engines that drop the old versions when new ones are discovered [2]. Web archives already hold more than 282 billion files and this number continues to grow as new initiatives continue to arise [9]. This data dimension is one order of magnitude larger than the number of documents indexed by the largest web search engine and 150 times more than the content of the Library of Congress. About 89% of the world web archives provide URL search [9], mostly supported by the open source Wayback Machine [16], which returns a list of chronologically ordered versions of that URL. However, this type of search forces the users to know the URLs of the content that contain the required information, some of which may have disappeared many years before.

The National Library of the Netherlands conducted a usability test on the searching functionalities of its web archive [15] and derived a list of the top 10 functions that users would like to see implemented. Full-text search was the first ranked, followed by URL search. At least 67% of web archives support full-text search for a part of their collections [9].

Research on the design of user interfaces to search historical web collections is giving its first steps. It has mainly being focused on the exploitation of the temporal perspective of data through the introduction of new user interface elements such as timelines or information clusters [1, 12]. However, the presented research does not address the specific requirements of searching historical web collections. At most, the original data for this research was obtained from curated online news archives, which is not representative of the heterogeneity of data addressed by a web archive. Hearst's book presents a comprehensive analysis of user interface design to search the live web but searching over historical web collections is not addressed [10]. For instance, the date of publication is mandatory on any news article, but very hard to identify on a typical web page [3, 5]. Note that the date of publication is crucial to enable users to explore the temporal perspective of data. This demo will show how these challenges were overcome in practice to create the PWA search service.

3. HISTORICAL WEB GRAPH COMPUTATION

Millions of contents archived across time can match a query from the user. Hence, ranking the contents is essential to provide relevant search results. The ranking of web search results has been thoroughly studied in Information Retrieval. However, searching collections composed by contents harvested form the web across time raises new challenges.

Link-based ranking algorithms assume that a link created by an author to a URL attributes importance to it. Therefore, the number of links received by a URL is an indicative of its importance.

The anchors associated to links provide a short description for the linked content. These descriptions have an additional value because they were created by content readers. Link-based algorithms contribute to rank web search results through the analysis of links and anchor texts of the web graph. Therefore, applying this kind of algorithms to historical web collections to improve search results sounds promising. However, link-based algorithms cannot be applied directly to web archives. Each node in a live-web graph is uniquely identified by a URL, but in a web archive each URL is associated to multiple content versions acquired along time (see Figure 1(a)). Thus, the nodes of the graph derived from a historical web collection must be identified by content versions (URL + timestamp) and not just by URLs. Computing the web graph ignoring this fact leads to the following problems:

Links to the future: Figure 1(b) shows that $URLx$ in 1996 links to a version of $URLy$ in 2004. The version of $URLy$ in 2004 can have a completely different content than when the author of $URLx$ created the link;

Link overweighting: Figure 1(c) shows that $URLy$ is overweighted, because it persisted for a long time and not because it received many links from different authors.

The solution adopted in the PWA to enable the application of link-based ranking algorithms was to pre-process the web graph to generate temporal layers between the URL versions. First, we identify each content's version by the pair <URL, timestamp>. We used the timestamp of content acquisition, but other timestamps could also be used, such as the creation or publication date. For each link found on the graph from a $URLx$ to $URLy$, we linked each version of $URLx$ to the closest archived version of $URLy$ within a time window. The time window size varies according to the timestamp of the referrer URL. The time window is defined as $]-\infty, timestamp+t]$, where t varies according to the timespan of the collection. We set t to one month because the distribution of the crawl dates suggested that the collection was generated from monthly web crawls. For the daily crawls, we set t to one day. The time window limitation eliminates links to versions in a far future. Limiting links to the closest version of the referenced URL eliminates the problem of link overweighting. The proposed pre-processement algorithm also reduces the number of links in the web graph. An evaluation of ranking algorithms to support temporal search has been undertaken [7]. This demonstration aims to expose how the chosen algorithms perform in practice on a live service.

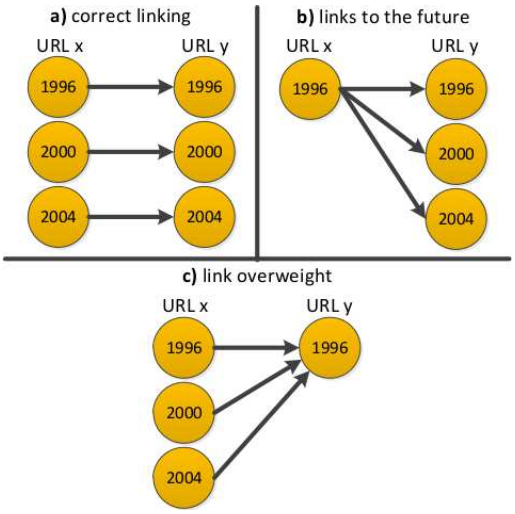

Figure 1: Problems with the computation of link graphs generated from historical web collections.

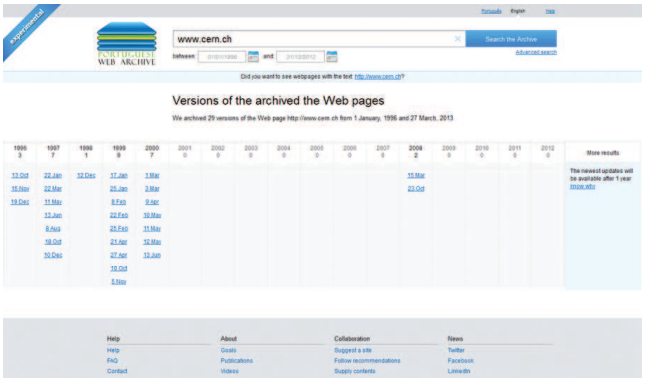

Figure 2: Result page for a URL search on the Portuguese Web Archive (history page).

4. USER INTERFACE DESIGN

Most web archives were created and are maintained by libraries [9]. The digital library user interfaces aim to provide several search choices to the users through a stricter model of interaction based of faceted search (e.g. title, author, abstract). This leads to complex user interfaces composed by several UI elements that require strong contextualization and decisions by the users to provide relevant search results. On its turn, the typical search engine interface is simpler and more familiar to users [8] which diminishes the learning curve. The downside is that it usually does not consider the temporal dimension.

The choice of the search UI for the PWA was conditioned by the adopted data acquisition policy. Digital libraries host carefully curated collections with rich curated meta-data to enable faceted search interfaces, but the PWA broadly archives large amounts of web data as search engines do. Thus, we decided to use a typical web search engine interface as baseline. A web archive UI must additionally address temporal search restrictions (e.g. definition of date interval), versioning of URLs on search results (e.g. version comparison) and reproduction of archived contents with

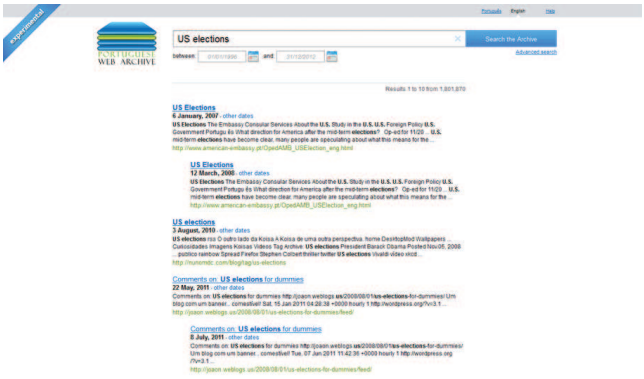

Figure 3: Result page for a full-text search on the Portuguese Web Archive.

meta-data for temporal contextualization (e.g. present crawl date). The main difference with traditional search engines' interface is that web archives have to address the temporal dimension of the searched data.

During the development of our UI we performed several iterations of laboratory usability tests to identify interaction problems and validate changes. The usability testing was used to verify both whole pages and specific components, with each interface change triggering a new round of tests.

The PWA interface is available at `archive.pt` and it is composed by:

Archived content view: presents the archived content along with the original URL and crawl date. It supports link navigation within the archive;

URL search results list (Figure 2): presents the history of crawled versions from a given URL in a yearly grid. Each date links to the archived content view;

Full-text search results list (Figure 3): for each search result presents the title that links to the archived content, its crawl date, a *other dates* link to the history page of that URL and a snippet of the content containing the query terms;

Search form: (Figures 2 and 3) the search form is present on the top of the URL and full-text lists. It is composed by a text search box that receives the query terms and two datepickers to restrict the crawl dates of the contents to be searched;

Advanced search form: enables users to refine search by defining phrase, term negation, results sorting, crawl dates, file format, site restriction and number of presented results.

Most users preferred to use the search box to do temporal restrictions on queries instead of using the available elements (textfields for dates, datepicker) to change the interval for their queries. Even so, 35% of the queries had the search date interval changed.

Only 20% of the users answered that they knew what a web archive was. The users compared the behavior of their favorite search engine with our web archive and expected the same response speed and search results quality due to the UI similarity. They did not understand the difference between

searching the live web and historical web collections. Using a web archive to access pages that are no longer available on the live web is a confusing concept to most users and requires technical knowledge about the functioning of the Internet.

The PWA supports full-text and URL search. Our first UI versions were composed by two distinct search forms: one for full-text search and another one for URL search. This approach failed because users did not understand the difference between search types. They inserted full-text queries on the URL search form, and vice-versa. The fact may be justified by the users' tendency to fill the first text field that looks like a searchbox [13]. The solution was to present a single textfield that receives any query. If the query term is composed by a URL, the corresponding history page is presented to the user. If the query terms include a URL but also other terms, it does a full-text search with the query terms but also presents a suggestion link to the history page of the queried URL. Otherwise, it performs a full-text search for the query terms. The URL queries are expanded to find results crawled with different URLs, that are likely to present reference the same content, for example, with and without "www." prefix, trailing "/" or "index.html" string.

The addition of a query spellchecker had great impact on the perceived quality of the web archive. Users frequently mistyped queries and blamed the web archive for poor search results, often failing to spot their own mistypes. After the introduction of the spellchecker, there were fewer negative user comments. The adopted spellchecker is based on Hunspell [6]. The presented changes increased the overall user satisfaction from 51% on the first version of the UI to 71% on the last one.

5. DEMONSTRATION

This demonstration will present a searchable web archive and how research results from several areas were applied in practice to develop this service. We will demonstrate how searching archived web data differs from searching the live web, requiring new ranking algorithms and fine-grain tunings on typical search interfaces.

The Internet Archive preserves information archived from the .PT domain. A collection of these files archived between 1996 and 2007 was integrated on the PWA and became full-text searchable. Web users can witness the difference between the user experience provided by the PWA full-text search service and the Internet Archive Wayback Machine, which is limited to URL search. The PWA is intended to be used by international users. Thus, its search interface is available in Portuguese (`arquivo.pt`) and English (`archive.pt`). Despite the majority of the archived content being in Portuguese, it also preserves content in other languages such as English, Spanish or French. The PWA has been used as an information source by international researchers, that studied archived content written in languages that they did not know, by combining our search service with automatic translation tools, such as Google Translate.

The PWA also provides access via an OpenSearch API that has been used to facilitate the development of new web applications that search archived data, such as mash-ups that combine past and present information about politicians. Nonetheless, the simplest way to demonstrate the OpenSearch feature is to create a browser search box that directly searches the PWA. This simple feature is useful to professionals that frequently need to search for past web content, such as journalists or humanities researchers.

There are at least 77 web archiving initiatives established worldwide [17]. We believe that this demonstration will be interesting to web archivists, but also to organizations concerned with historical preservation that aim to start their own web preservation initiatives, such as national archives or libraries. A short video about the PWA is available at `vimeo.com/59507267` but the service can be tried live at `archive.pt`.

6. ACKNOWLEDGMENTS

We thank the Internet Archive for supplying us the information archived from the .PT domain and for its continuous and tireless efforts to preserve and grant access to human knowledge. We acknowledge Marco de Sá and Rui Lopes for their precious collaboration in the user interface design.

7. REFERENCES

[1] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proc. of the 18th ACM Conference on Information and Knowledge Management*, pages 97–106, 2009.

[2] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011)*, pages 1–8, 2011.

[3] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Trend detection through temporal link analysis. *American Society for Information Science and Technology*, 55(14):1270–1281, 2004.

[4] M. Cafarella and D. Cutting. Building Nutch: Open Source Search. *Queue*, 2(2):54–61, 2004.

[5] L. Clausen. Concerning etags and datestamps. In *Proc. of the 4th International Web Archiving Workshop*, volume 16, 2004.

[6] M. Costa, J. Miranda, D. Cruz, and D. Gomes. Query Suggestion for Web Archive Search. Technical report, Foundation for National Scientific Computing, 01 2012.

[7] M. Costa and M. J. Silva. Evaluating web archive search systems. In *Proceedings of the 13th international conference on Web Information Systems Engineering*, WISE'12, pages 440–454, Berlin, Heidelberg, 2012. Springer-Verlag.

[8] C. De Rosa, J. Cantrell, J. Hawk, and A. Wilson. *College Students' Perceptions of Libraries and Information Resources: A Report to the OCLC Membership.* OCLC, 2006.

[9] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. In *International Conference on Theory and Practice of Digital Libraries 2011*, Berlin, Germany, September 2011.

[10] M. Hearst. *Search User Interfaces.* Cambridge University Press, 2009.

[11] Internet Archive. Nutchwax - Home Page. `http://archive-access.sourceforge.net/`, March 2008.

[12] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching Through Time in the New York Times. In *Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval*, pages 41–44, 2010.

[13] J. Nielsen and H. Loranger. *Prioritizing Web Usability.* New Riders, 2006.

[14] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.

[15] M. Ras and S. van Bussel. Web archiving user survey. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek), 2007.

[16] B. Tofel. 'Wayback' for Accessing Web Archives. In *7th International Web Archiving Workshop (IWAW07)*, Viena, Austria, September 2007.

[17] Wikipedia. List of web archiving initiatives — wikipedia, the free encyclopedia, 2012. [Online; accessed 22-June-2012].