Adapting search user interfaces to web archives

David Cruz
Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal
david.cruz@fccn.pt

Daniel Gomes
Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal
daniel.gomes@fccn.pt

ABSTRACT

Despite the importance of web archives for the access to historical information published on the Internet, human interaction with web archives systems has not been thoroughly addressed. The web archive search user interface presented on this paper was derived from several rounds of development and usability testing over the Portuguese Web Archive search user interface (available at `archive.pt`). We present our findings gathered while adapting a typical web search user interface to the context of web archive search. We describe how we adapted a typical search user interface to address full-text and URL search over web-archived data, highlighting the unexpected problems detected during usability testing of our interface and current limitations for future work. The obtained results from usability testing showed that the average user satisfaction with our user interface was 70%. The obtained results from anonymous user satisfaction questionnaires yield a 84.3% score. We believe that our work can be applied to improve the quality of the services provided by other web archives.

1. INTRODUCTION

Most web users are not acquainted with web archives and accessing to archived web data is a significantly different user experience than accessing the live web [2, 5, 6]. Current users demand ready-to-use applications and are getting less tolerant to usability barriers. Although there is a significant number of web archives available [9], only preliminary research has been done about the design of user interfaces to gain access to temporal web data. The adoption of inadequate user interfaces to gain access to web archives jeopardizes the return of the investment made to preserve historical web content. User interfaces for web archives must be carefully designed and tested to respond to real-world user requirements and provide functional features specific to the exploitation of Web-archived content.

The Portuguese Web Archive (PWA) began in 2008. It is a public search service with over 1 131 million web files archived since 1996 that aims to preserve web content of interest to the Portuguese community (`archive.pt`). After witnessing the difficulties of our users, we increased our effort on improving their experience and satisfaction while using the PWA.

In this study, we share our experience of adapting a typical live-web search user interface to support web archive interaction. The lessons learned during this work would have been helpful to "kick start" our web archive in 2008. The developed code of our search system is freely available. We believe that our contributions will help other web archivists to improve the user experience and impact of their services.

2. METHODOLOGY AND RESULTS

The applied methodology followed an user-centered design approach [1]. The usability of our search interfaces was tested in collaboration with HCI experts from the Human Computer Interaction and Multimedia research group from the University of Lisbon. We tested several versions of our web archive user interface. Each new version triggered a new round of usability tests. Each testing round consisted of ten tasks presented to six users with no experience of using web archives. Each of the users executed the test individually in the presence of an usability expert. We recorded the screen, audio and participants' facial expressions for later analysis. Participants had to fill a questionnaire before the test to inform about their proficiencies on Internet usage and a post-test satisfaction questionnaire [4]. We finished each test with a debriefing session to further explore users' difficulties and clarify doubts in our observations. We obtained feedback from 21 users with distinct profiles. We analyzed the obtained results using a Likert scale from "1" (strongly unsatisfied) to "7" (strongly satisfied).

The obtained results from laboratory usability testing showed that the average user satisfaction increased from 3.6 (51.4%) on the first version of our interface to 4.9 (70%) on the last version. We also obtained results from anonymous user satisfaction questionnaires filled during dissemination events by users after freely trying the PWA, where we obtained a 5.9 (84.3%) score. This evaluation methodology occurred in an environment less controlled than our laboratory usability testing and used simpler questionnaires to avoid overloading users. On the other hand, the results were obtained from an usage environment closer to reality.
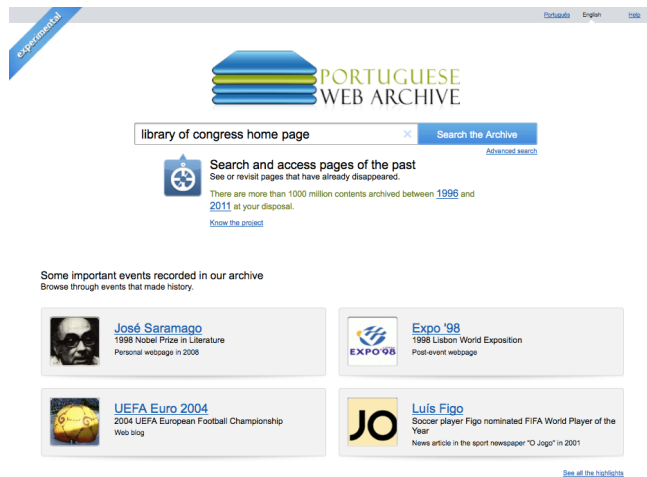
3. ANATOMY OF A WEB ARCHIVE SEARCH USER INTERFACE

Through usability testing on the first versions of the Portuguese Web Archive, we made two determinant observations. The first observation was that searching historical web content was an awkward concept to most web users. The existence of a website (web archive) that provides access to pages that are no longer available on their original websites is a perception that requires technical knowledge about the functioning of the Internet that is far beyond the skills of common web users. The second observation was that obliging the users to choose between a URL and full-

text search interfaces to gain access to web-archived content was ineffective and confusing to them.

The search user interface designed for the Portuguese Web Archive explores the users' familiarity with traditional search engines by offering similar layout and familiarity and enhance it with specific functions and contextual information required from web archives.

3.1 Search homepage

Figure 1: Interface for the home page of the web archive search user interface deployed on the Portuguese Web Archive.

Figure 1 presents the home page for the web archive search service. This homepage presents a search box without any temporal controls and some highlights of archived pages. The highlights are fundamental anchors that allow users to explore curated content, especially if the users never used a web archive before and have no clue about what they can find in it. The users observe examples that illustrate the type of information that they may find and progressively gain awareness about the potential of searching a web archive, thus reducing the cognitive effort of first-use. The publication of selected archived pages on the home page improved the overall user satisfaction with the service. Watching archived pages with historical value that have already disappeared triggered feelings of nostalgia which increased the positive perceptions, reflected through comments, about the provided service and general usefulness of web archives. Unlike most live-web search engines, our web archive search home page also includes a fat footer. The objective was to provide additional links to information that clarifies users about the context of web archiving and web archive search. For instance, links to: texts and videos about the project, a form to suggest sites to be archived, news or help.

We designed our interaction model to support both types of queries and shift the burden of detecting the query type to the system. Users only have to fill one search box and our system detects the type of query and presents the results in an interface tailored for that query type. When the query is composed exclusively by a URL, the corresponding version history results are returned. The results also include the versions from different URLs that are likely to reference the same content (e.g. www.site.pt, site.pt, site.pt/index.html).

If the query is exclusively composed by text, the system returns full-text search results. If the query includes text and a URL, the system returns the full-text search results and suggests a link to the versions history of the URL. By doing so, the web archive interface becomes similar to live-web search engines, which users are already acquainted with, and guides them from familiar ground to the new context of searching historical web content.

3.2 Full-text search results

Figure 2 shows the user interface in full-text search mode. It is comprised by a typical search field for the query and a list of search results. But also, date input fields and datepickers to restrict the temporal interval of the queries. The two datepickers define lower and upper limits of the page archive dates to be searched. The results are shown on a results page similar to traditional live-web search engines. What differs is that we give greater emphasis to the archival date of each result. We tried several layouts and found that the position where users better recognized the dates was bellow the result title. Even so, some users ignored the unusual display of the date of archival within the search results. The interface for full-text search allows users to sort the results by relevance or archive date through the "sort:" operator or the advanced search interface. The advanced search also provides additional fields to allow more specific queries where users can restrict for: words, phrases, excluded words, file type, website or number of displayed results.

3.3 URL history search results

Searching for a URL or clicking on the "other dates" links on the full-text results page directs the users to the history view of that URL. The results are presented on a grid layout where each column group the several archived versions of a specific year, starting from the oldest year, supported by the Archive, up to the most recent year.

Each column then lists the available versions for that year, starting from the oldest. The users have an overall view of the versions available for a given URL. Clicking on the date link opens the correspondent version of the archived page. The grid layout approach was well understood by users. The versions from the current year are unavailable because the PWA only provides access to the archived pages one year after their archival so that the accesses to archived content do not concur with the original live-web sites (embargo policy). However, we display the current year column with a notice that explains the embargo policy to the users.

3.4 Reproduction of archived content

The interface currently in production that reproduces archived content presents a banner on the top of the archived page with the original URL and the archived date. Having an interface element always visible presents consistent hints that archived pages are different and behave differently from live webpages. However, we observed usability problems related to the reproduction of archived pages that deserve further research in future work. Users frequently lost perception about if they were navigating through archived content reproduced by a web archive or the live-web. One reason for this fact was that when the users scrolled-down the archived page, they lost visual contact with the top banner. On the other hand, the banner interfered with layout of some
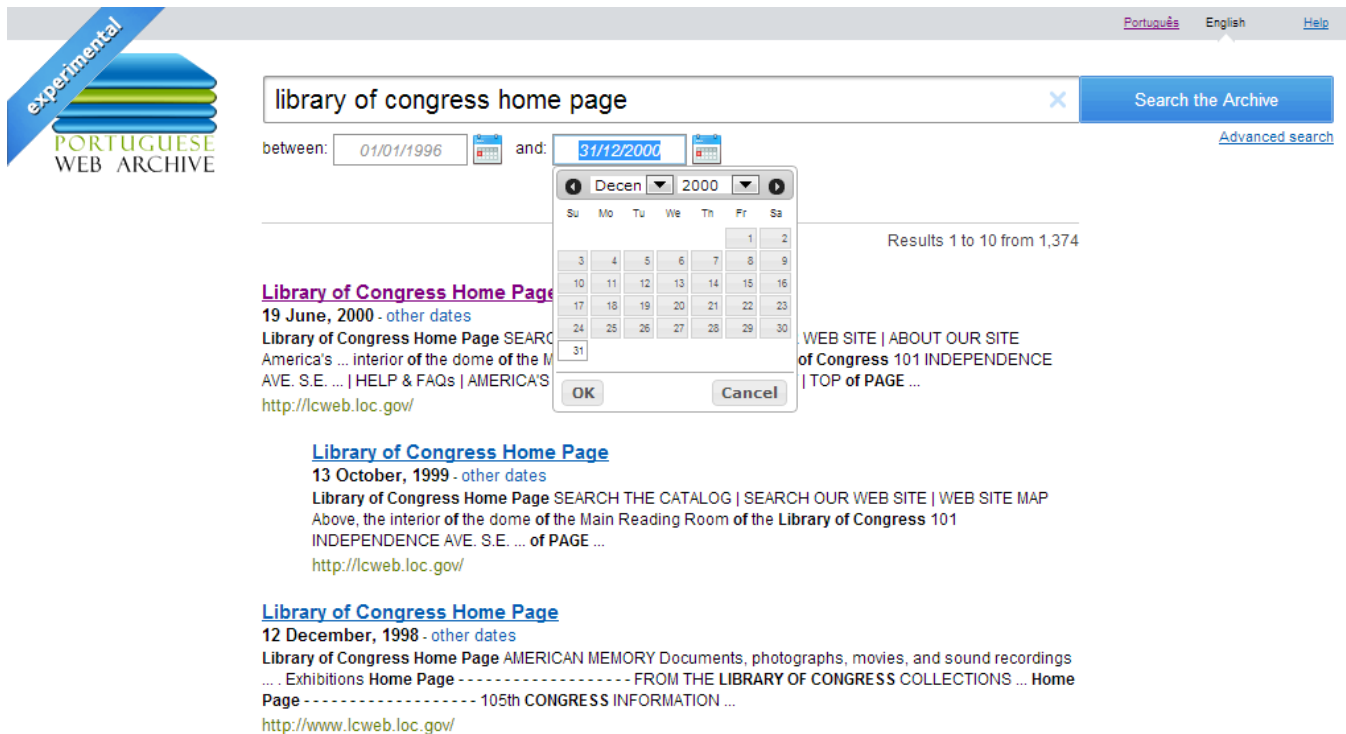
Figure 2: Interface for the web archive full-text search deployed on the Portuguese Web Archive.

archived pages by appearing on top of important content or links, for example in pages containing framesets or when the layout used absolute positioning in CSS.

Figure 3 presents the future interface design for reproducing an archived content. We took special care during the requirement analysis for this interface because it is unique to web archive search. Live-web search engines do not have to reproduce pages nor provide historical features, at most they provide a simple "cache" function that displays the textual content of the last version of an indexed pages without further concerns about maintaining its original layout. The presented design was derived through brainstorms using KJ-method [7] and several rounds of usability testing using low-fidelity paper prototypes [3] with a think-aloud protocol [8]. Notice that, unlike the previously presented interfaces, this new interface design has not yet been deployed to production on the PWA.

The archived page is reproduced in an internal frame so that the original layout is isolated and without interferences from the additional features on the page. We concluded that the interface for viewing archived pages should provide contextual information about the page (URL, date, help), features for sharing by e-mail and the main social networks (Twitter and Facebook), and for saving a copy of the archived page as image, PDF or compressed file. A sidebar enables the users to switch between versions of the archived page without having to return to the history page. To maximize the viewport devoted to the archived page, those contextual and navigational interfaces can be collapsed to a narrow bar above the archived page with the minimal information needed: The PWA logo that links to the Homepage, the URL, the date of the version presented and a button to expand the interface. Contrary to the old interface, the new one always show contextual information about the archived page, even if the archived pages are scrolled down by the users.

4. DATEPICKER TUNNING

The UI element that required the most tweaks was the datepicker. Standard datepickers are conceptually simple, only presenting a grid of the days of the month and left/right arrows to view previous/next months. However, web archives collect data that can span through decades. For example, the Portuguese Web Archive hold pages archived from 1996 to 2012. Thus, traversing this date range using a standard datepicker would require 203 clicks. After several design iterations, we concluded that a web archive datepicker should use drop-down lists to allow a quicker selection of month and year of the time span of the search (see figure 2).

We observed that for tasks with implicit days (e.g., "Movies released during June 2000"), users only specified the month and year but did not specify the day. Then, they either dismissed the datepicker by clicking outside (doing so closed it without saving the date) or became confused hesitating on how to proceed next. For the users, choosing the month was sufficient to communicate their temporal intent to the datepicker and got flustered because they had to do the extra work of choosing and clicking on a specific day. This unsatisfactory user interaction was overcome by adding a "OK" confirmation button and a "Cancel" button to dismiss the datepicker. With these buttons, users gained a strong visual anchor to decide unambiguously how to submit a new date or close the datepicker without any change to the current date. When the users click the "OK" button without selecting a specific day, the context of the datepicker determines its next state. If the user is defining the lower limit for

Figure 3: Interface design for reproducing archived pages.

the search interval through the left datepicker, the first day of the month is selected. If the user is defining the upper limit for the search interval through the right datepicker, the last day of the month is selected.

We also observed that some participants on the usability tests clicked first on the day before adjusting the month or year. The default behavior for the datepicker was to close immediately after the day was selected without leaving the opportunity for further adjustments. This user behavior depends on the date format the users were most familiar with. For example, for the date 24 December 1996, the users interacted with the datepicker according to their mental model of the date (day, month, year) and not to the visual organization of the information presented through the datepicker (month, year, day).

5. CONCLUSIONS

Search user interfaces for web archives must be similar to live-web search engines to facilitate the adoption of web archives by new users. This study presented the design and lessons learned while developing the Portuguese Web Archive (PWA) search user interface. The PWA follows the typical pattern of a live-web search user interface but enhances it with features to manipulate historical web content. Several aspects had to be carefully addressed in the design of the web archive search user interface such as how to handle different query types, how to present results that span across time or how to make users notice temporal information associated to archived pages. The results obtained from laboratory usability testing showed that the average user satisfaction increased from 51% on the first version of our interface to 70% on the last one that is currently in production at `archive.pt`.

Our main conclusions were that the home page for a web archive search service must present contextual information, such as examples of archived pages, that enable new users to gain awareness about what is and the potential of searching a web archive. Web archives should gracefully combine full-text with URL search. Users became unaware of their query misspellings, therefore web archives must provide query suggestion mechanisms. Standard datepickers

are not adequate to be used in web archives and needed adjustments to be successfully applied as user interface elements to define the time scope of searches. By presenting the adjustments that we made to our interface and explaining their rationale, we expect to raise awareness about the importance of user interfaces to the success of web archives as useful services for modern societies. All the code and interface resources are freely available to be reused and improved at `code.google.com/p/pwa-technologies/`.

6. REFERENCES

[1] C. Abras, D. Maloney-Krichmar, and J. Preece. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, 37(4):445–56, 2004.

[2] P. Aravind, V. Arce, and P. Roessler. Qualitative Assessment of the Internet Archive's Wayback Machine. Technical report, University of California, Berkeley, May 2002.

[3] A. Coyette, S. Kieffer, and J. Vanderdonckt. Multi-fidelity prototyping of user interfaces. In *Human-Computer Interaction–INTERACT 2007*, pages 150–164. Springer, 2007.

[4] J. R. Lewis. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact.*, 7:57–78, January 1995.

[5] J. Niu. Functionalities of web archives. *D-Lib Magazine*, 18(3/4), 2012.

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

[7] J. Spool. The KJ-technique: A group process for establishing priorities. *User Interface Engineering*, May 2004.

[8] M. W. Van Someren, Y. F. Barnard, J. A. Sandberg, et al. *The think aloud method: A practical guide to modelling cognitive processes*. Academic Press London, 1994.

[9] Wikipedia. List of web archiving initiatives — wikipedia, the free encyclopedia, 2013. [Online; accessed 15-April-2013].