Understanding the Information Needs of Web Archive Users

Miguel Costa [1,2]
miguel.costa@fccn.pt

Mário J. Silva [2]
mjs@di.fc.ul.pt

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

ABSTRACT

A complete characterization of web archive users must respond to three questions: why, what and how do users search? This study focuses on the first two: what are the user intents and which topics are most interesting to them? Answers to these questions are essential for guiding the development of web archives towards better user satisfaction. We used three instruments to collect quantitative and qualitative data, namely, search logs, an online questionnaire and a laboratory study. The obtained results are coincident. Users perform mostly navigational searches and do not restrict searches by date. Other findings show that users prefer full-text over URL search and the oldest documents over the newest. We discuss all these findings and their implications in the design of search engines for web archives.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.7 [**Digital Libraries**]: User issues

General Terms

Web, Archive, Preservation, User, Characterization

Keywords

Portuguese Web Archive, Information needs

1. INTRODUCTION

Web archiving initiatives have been working for several years in harvesting and preserving the countries' web heritage (see `http://www.nla.gov.au/padi/topics/92.html`). They have archived billions of web documents, many containing unique and valuable information from the past. The Internet Archive alone collected 150 billion documents since 1996. On the other hand, the web archive retrieval technology is still in its early stages. Most of these data are only searchable by URL, which the users have to remember. The few web archives that provide full-text search are based on web search engine technology, which ignores the temporal dimension of collections. This leads us to question whether this technology can support 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, this is not always easy, since some times users only have a vague idea of what they want the system to do. We faced this problem when we started developing the access functionalities for the Portuguese Web Archive (PWA) [8]. People had a great difficulty in suggesting anything without seeing the system working. Showing similar systems from other countries, helped them to understand the concept of the project. Nevertheless, without real information needs over past documents and subjects they could remember and explore, the responses continued to be too vague. The only feedback we received was 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 we implemented them. However, full-text and URL search are not an end in themselves. They are mechanisms to obtain some kind of information, such as a subject written in the past.

With the public release of the PWA experimental version, it was finally possible to collect valuable feedback from the 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 [13, 7]. We expect the same behavior in full-text search over web archive collections. Third, it structures the elaboration of a representative information retrieval (IR) evaluation over web archives [5]. Being IR mostly an empirical discipline, joint evaluation initiatives are undeniably important to foster IR research.

We used three methods to collect data from users, namely, search logs, an online questionnaire to be answered by the users while they were searching and a laboratory study. All experiments were conducted on the PWA, which contains nearly 150 million web documents accessible by full-text and URL search. As far as we know, this is the largest web archive collection searchable by full-text and over such large time span. The documents range between 1996 and 2009.

Results show that users from web archives and web search engines have different information needs, so they cannot be supported by the same technology. Results also show that

10th International Web Archiving Workshop, Vienna, Austria.
.

the few search functionalities where time is present are infrequently used. However, when used, they are mostly employed for picking the oldest documents. This discovery can be used in the ranking of results. We discuss all findings and the implications on the development of future web archives. We also draw the first profile of web archive users.

This paper is organized as follows. In Section 2, we cover the related work. In Section 3, we describe the PWA user interface. The methodology of analysis is explained in Section 4 and the results are detailed in Section 5. Section 6 finalizes with the discussion of results and conclusions.

2. RELATED WORK

2.1 Web Archive Users

Although there are several web archiving initiatives currently harvesting and preserving the web heritage, there are very few studies about web archive users. The web archiving user survey from the National Library of the Netherlands is the most comprehensive study [16]. Still, only fifteen users participated in it. The study compiled a list of the top ten functionalities that users would like to see implemented. Full-text is the first one, followed by URL search. In none of the top ten functionalities is time mentioned. However, being time present in all the processes and foreseen solutions over a web archive, shouldn't the past web be searchable in both time and text dimensions? The users' choices can be explained by the fact that web archives are mostly based on web search engine technology and as result, web archives offer the same searching functionalities without the time dimension [5]. This inevitably constrains the users' behaviors. Another explanation is that Google became the norm to the users, influencing the way users search in other IR systems. We realized this in our preliminary experiments conducted on the PWA [5]. The experiments also revealed that users sometimes select a date range filter incorporated in the interface to narrow the search to a specific period. This filter exists in most web archives and in some cases serves to disambiguate queries.

The International Internet Preservation Consortium (IIPC) reported a number of possible user scenarios over a web archive [9]. It describes the technical requirements necessary to satisfy the hypothetical goals of web archive users. It highlights several information needs, some derived from professional scopes. However, these did not come directly from users. Our work tries to identify and aggregate the information needs from real users using the experimental version of the PWA.

2.2 Users' Information Needs

Users' information needs have been investigated in different IR systems, specially web search engines that are the most studied. There exists a consensus among researchers about the taxonomy proposed by Broder [3] and refined by Rose and Levinson [17]. Broder classified web search engine queries into three broad classes according to the user goal, which can be: (1) **navigational** - to reach a web page or site in mind; (2) **informational** - to collect information about a topic, usually from multiple pages without a specific one in mind; (3) **transactional** - to perform a web-mediated activity (e.g. shopping, downloading a file, finding a map) [3]. 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, was 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 the Broder taxonomy of web search, creating sub-classes for the informational and transactional categories [17]. 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. presented an integrated view of them [10].

Different IR systems and environments have users with different information needs. For instance, Church and Smyth used diary studies to explore information needs of mobile users [4]. 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.

3. THE PWA USER INTERFACE

The PWA preserves the Portuguese web, which is considered the web with most interest for the Portuguese community. Specifically, we define the Portuguese web as all the documents[1] satisfying one the following rules: (1) hosted on a site under a .PT domain; (2) hosted on a site under another domain, but embedded in a document under the .PT domain; (3) suggested by the users and manually validated by the PWA team.

The experimental version of the PWA is a public service since April 2010. It is accessible from `http://archive.pt`, where the users may choose between a Portuguese and an English language interface. Currently, it provides nearly 150 million documents searchable by full-text and URL, and complemented with a date range filter to narrow the results to a time period. Other web archives, such as Padicat (see `http://www.padicat.cat`) and Pandora (see `http://pandora.nla.gov.au`), provide similar access. However, in our interface both full-text and URL queries are submitted from the same text box. The PWA interprets the type of query and presents the results accordingly.

When the PWA receives a full-text search, it returns a results page containing a list of 10 results matching the query. Figure 1 illustrates a typical session, where the interaction with the user and the layout of the results is similar to web search engines, such as Google. The results are ranked by relevance to the query, determined by the PWA ranking model. Each result includes the title of the web page and its crawled date, a snippet of text containing the query terms and the URL of the web page. 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 not found, the user can repeatedly modify and resubmit the query. In addition, the user can click on the navigation links to explore other result pages or use the advanced search interface to restrict the query with advanced operators. Figure 2 shows the available operators, such as the restriction by format and the sort by

[1]The terms document and file are used interchangeably in this study. For instance, it can be a web page, an image, a PDF file.

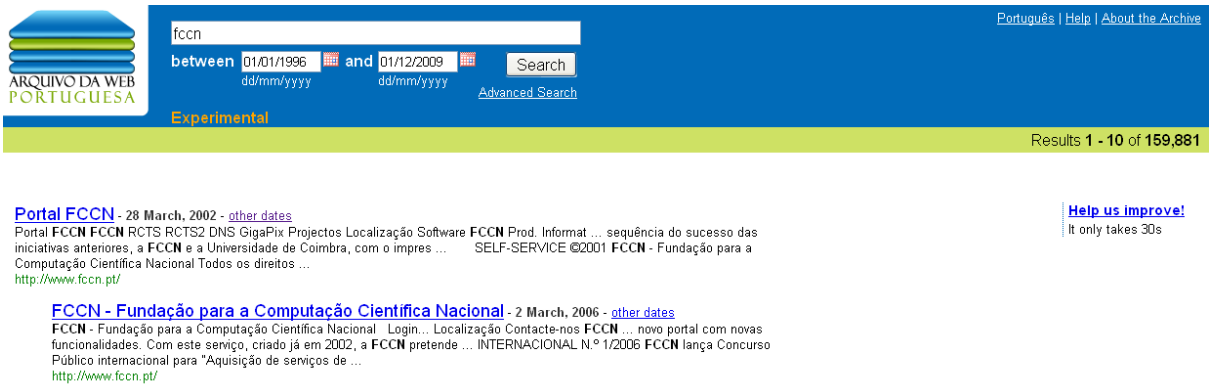

Figure 1: Search interface after full-text search.

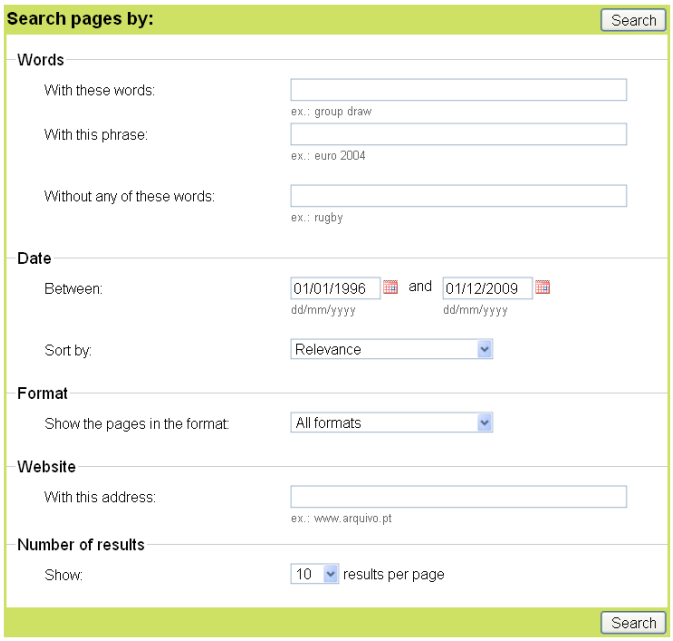

Figure 2: Advanced Search Interface.

one of the three criteria: relevance, newest first or oldest first. These advanced operators can also be added to the query directly in the search box of the main interface.

Each result has also an associated link to see all versions of the respective page. When clicked, the PWA presents the same results page as when a user submits that URL. Figure 3 depicts the interface, which is similar to the one of the Wayback Machine (see `http://www.archive.org/web/web.php`). A table is shown to the user, where each column contains all 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. METHODOLOGY

User study methods can be classified into three groups: (1) client-side [6] or server-side [11] log analysis of the users interactions with the system; (2) surveys based on interviews [19] or questionnaires [1] conducted on users; (3) experiments with users in a laboratory [2] or in their natural environment (in-situ) [14]; All methods have pros and cons, so we experimented one of each group as complementary ways of analysis. Next, we synthesize the chosen methods.

4.1 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. Another advantage of this method is its unobtrusiveness, i.e. non intrusiveness in the users' normal behavior. Most users are not aware 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.

Contextual information must be collected using other methods. A possibility is to ask users directly, showing online interactive questionnaires when the users are performing or concluding a critical function. This allows the users to enter fresh opinions on the systems' 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 systems usability and users 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

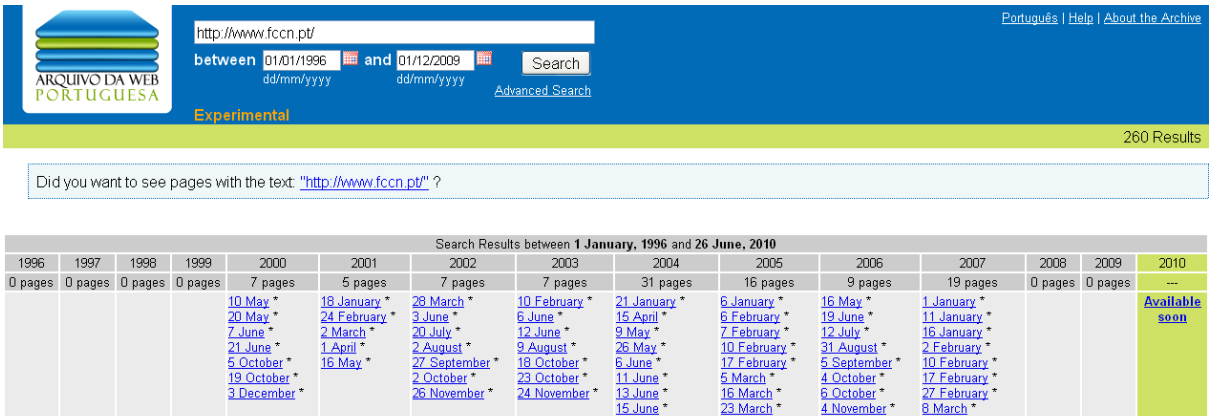

Figure 3: Search interface after URL search.

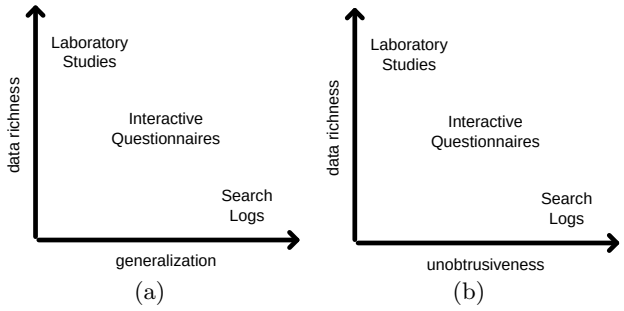

Figure 4: Data collecting methods used.

data to generalize results and rich data to explain them. Figures 4(a) and 4(b) represent the relation between the three chosen 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(a) and the degree of unobtrusiveness in Figure 4(b), where search logs surpass all others. Next, we detail the experiments.

4.2 Experiment # 1: Search logs

4.2.1 Procedure

We 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 the sessions with more than 100 queries were also excluded. Sessions with many queries were likely to come from web crawlers and we were only interested in queries submitted by human users.

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, we used the users' IP address and session identifier to delimit sessions [11]. We also used a time interval t of inactivity. Two consecutive interactions are included on different sessions if they have an inactivity between them of at least t. Without this interval, we could have sessions of several days, which would hardly represent the reality. We 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 SVM classifiers used for delimiting sessions [15].

After delimiting the sessions, we followed the Rose and Levinson idea of developing a tool for assisting session classification [17]. Using this tool, we manually analyzed the queries and clicks of 400 sessions 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. The taxonomy of the topics was based on the Jansen et al.'s studies [11].

4.2.2 Participants

The PWA contains all kind of contents from the Portuguese web. Moreover, the PWA is a public service, so we 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.

We never used the log data to match a real identity. However, we checked the location of the users' IP addresses. We counted 81% of PWA users with IP addresses assigned to Portugal and near 94% of the interactions were submitted through the Portuguese language interface.

4.3 Experiment # 2: Interactive Questionnaire

4.3.1 Procedure

Our 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, our 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 1 shows this message: *Help us improve! It only takes 30s.*

The questionnaire, presented in Appendix A, was designed based on existing guidelines described by Jansen et al. [12]. It was implemented using the Google Forms framework (see `http://docs.google.com`), 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 we suggest or new ones that we did not envision. The second, focuses in determining if the need is restricted to a specific date range. The third, asks 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. We chose to restrict the number of questions to five, without demographic or experience related questions, because the participation rate on this type of experiences tends to be low. Increasing the number of questions, specially open-questions, would further reduce this rate.

We performed two pre-evaluation studies with five users each 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, we manually validated all responses. To guarantee that the same user had not submitted the questionnaire multiple times, we checked the users' IP addresses and session identifiers.

4.3.2 Participants

Of the six users that opened this questionnaire through the searching interface, no one answered it. This indicates problems in the design adopted to captivate users and in the questionnaire itself. We detected that users 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.

Due to lack of responses, we asked people to experiment the PWA and then to answer the questionnaire. We disseminated this request 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. We 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.

4.4 Experiment # 3: Laboratory study

4.4.1 Procedure

The experiment was conducted by the LaSIGE Human-Computer Interaction and Multimedia Research Team (see `http://hcim.di.fc.ul.pt/`) 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. We will not discuss these usability tests, since they are out of scope of this paper. A new paper will detail them. However, 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 [18]. 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 software. 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 Appendix A. The questionnaire was anonymous. Sixth, the researchers thanked the participant's help.

4.4.2 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 professions, interests and academic degrees. We believe that this diversity reflects the population of potential 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!.

5. RESULTS

All information needs of web archive users focus on past data and match a class from the taxonomy proposed by Broder [3]. As result, we aggregated options 1 and 2 from the first question (Q1) presented in Table 1. Both options refer a web page or site in mind, so we considered them navegational. Option 3 match the informational need and option 4 the transactional. We will not discuss the other options, since the results show that they are not likely real or statistically significant in frequency.

5.1 Experiment # 1: Search logs

Searching for a known page or site was the most frequent need. It 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. We counted 16.12% of navigational sessions containing only URL queries. Surprisingly, the URL queries represent 20.96% of all queries submitted.

Collecting information about a subject written in the past, was the second most frequent need. A total of 37.83% of the sessions were initiated due to this informational need. Downloading an old file, i.e. the transactional need, orig-

Q1	Which of the following phrases describe better what you were doing?	Need	Exp. #1	Exp. #2	Exp. #3
1	Seeing how a web page or site, that I know, was in the past (e.g. my homepage).	Navigational	47.70%	31.58%	47.62%
2	Seeing the evolution over time of a web page or site (e.g. the Google.pt page).	Navigational	9.21%	21.05%	33.33%
3	Collecting information about a subject written in the past (e.g. Iraq war).	Informational	37.83%	31.58%	14.29%
4	Downloading an old file (e.g. music, video, image or software).	Transactional	5.26%	10.53%	4.76%
5	Recovering a web page or site that disappeared (e.g. to recover my Blog).	Transactional	0%	5.26%	0%
6	Seeing the evolution over time of the popularity of a subject (e.g. crisis).	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 1: Distribution of information needs for the three experiments.

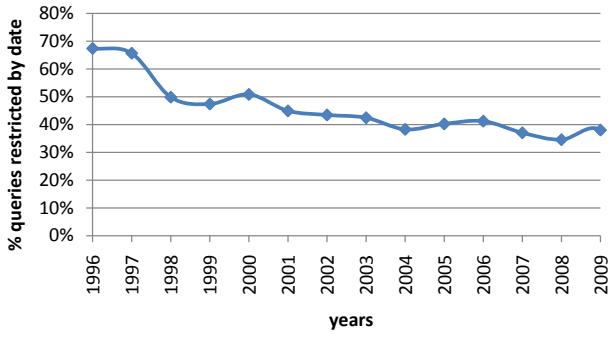

Figure 5: Distribution of years included in queries restricted by date.

(a)

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

(b)

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 2: Distribution of topics per (a) navigational and (b) informational needs.

inated 5.26% of the sessions. In this case, users searched mostly for images, but also searched for software, music, TV commercial jingles and bit torrent files.

The PWA users only restricted queries by date in 15.79% of the sessions, as shown in Table 1. Analyzing all the logs, we discovered that 11.02% of the queries were restricted with a start date, while 27.44% were restricted with an end date. This indicates that users are more interested in old documents. This idea is reinforced by the distribution of the years included in the queries restricted by date. As it can be seen in Figure 5, the older the years, the more likely they are of being included in queries. Another indication is that the option of sorting results by date was used in 0.5% of the queries to present the newest first, while it was used 3% to present the oldest first.

Finally, we separately classified the searched topics for the navigational and informational needs. For the navigational, we classified the sessions according to the topics to which the sites are mostly about. Table 2(a) 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, we classified the sessions according to the topics of the information searched. Table 2(b) shows that *People* was the most searched topic, corresponding to 36.52% of the sessions. Unexpectedly, 14.78% were about *Health* and 9.57% about *Entertainment*.

5.2 Experiment # 2: Interactive Questionnaire

Options 1 and 3 from the first question (Q1) presented in Table 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%. Option 4, i.e. the transactional need, corresponds to 10.53% of the participants choices. The second question (Q2) whether users searched between dates, almost divided the answers. Around 47% answered *Yes*.

We 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 comparison side-by-side 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.

We 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 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 use-case suggested was the creation of trustability profiles, based on the companies and employers background published on the past web.

5.3 Experiment # 3: Laboratory study

Table 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.

6. DISCUSSION AND CONCLUSIONS

All experiments indicate similar tendencies, despite the percentage variations. We believe these variations are mostly due to the small number of participants in experiments 2 and 3. Our 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 in Table 3 attest this. This changing of needs should be reflected in the retrieval technology, such as the ranking of results.

2. Most users do not restrict searches by date. They do not seem to have this need. However, 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.

3. Near 21% of the submitted queries had only a URL. These URLs represent web pages that the users want to see. Hence, they can be used as seeds for both bulk or selective harvesting approaches. These numbers also show that URL queries are common and should be supported. Nevertheless, users prefer full-text search.

4. Nearly half of the informational needs are focused on names of people, places or things. Many navigational queries only contain companies or institutions names. Named entity recognition can be a valuable technique to identify the best pages referring those names.

5. Users preferentially search the oldest documents. This discovery can be used in the ranking of results, when no other temporal data is given. This also indicates, as expected, that the importance of web archives tend to increase as the data ages.

6. Web archives fail in supporting some important needs. The most commonly sought was seeing and exploring the evolution of a web page or site. Tools to support fast comparisons between pages and sites should be researched. Another need that is not supported, but that was significantly mentioned, is image search.

This study provides the first general picture of why and what web archive users search. We believe that the obtained results are general, but studies over other web archives are necessary to confirm this. Our future work will focus on building a test collection composed by: a corpus representative of the documents encountered in a web archive; a set of topics simulating the users' information needs; and relevance judgments indicating which documents are relevant and nonrelevant for each topic. The results from this study are essential to guide the creation of this collection, which in turn will serve to improve the quality of web archive results.

7. ACKNOWLEDGMENTS

This work could not be done without the help and infrastructure of the Portuguese Web Archive team [8]. We thank the Human-Computer Interaction and Multimedia Research Team for conducting the laboratory study and Michel da Corte for her textual review of the paper. We also thank FCT (Portuguese research funding agency) for its LASIGE multi-annual support and POSC/EU for co-funding this work.

8. REFERENCES

[1] A. Aula, N. Jhaveri, and M. Käki. Information search and re-access strategies of experienced web users. In *Proc. of the 14th International Conference on World Wide Web*, pages 583–592, 2005.

[2] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *Proc. of the 28th International Conference on Human Factors in Computing Systems*, pages 35–44, 2010.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[4] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proc. of the 13th International Conference on Intelligent User Interfaces*, pages 247–256, 2009.

[5] M. Costa and M. J. Silva. Towards information retrieval evaluation over web archives. In *Proc. of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 37–40, 2009.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, 2005.

[7] X. Geng, T. Liu, T. Qin, A. Arnold, H. Li, and H. Shum. Query dependent ranking using k-nearest neighbor. In *Proc. of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–122, 2008.

[8] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the Portuguese web archive initiative. In *Proc. of the 8th International Web Archiving Workshop*, 2008.

[9] A. T. W. Group. Use cases for access to Internet Archives. Technical report, Internet Preservation Consortium (IIPC), 2006.

[10] B. Jansen, D. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, 44(3):1251–1266, 2008.

[11] B. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[12] B. Jansen, A. Spink, and I. Taksa. *Handbook of Research on Web Log Analysis: Surveys as a Complementary Method for Web Log Analysis*. Information Science Reference, 2008.

Study	Information Need		
	Informational	Transactional	Navigational
Broder user survey [3]	39%	36%	24.5%
Broder log analysis [3]	48%	30%	20%
Rose et al. 1st log analysis [17]	60.9%	24.3%	14.7%
Rose et al. 2nd log analysis [17]	61.3%	27%	11.7%
Rose et al. 3rd log analysis [17]	61.5%	25%	13.5%
Jansen et al. log analysis [10]	65%	20%	15%
Experiment #1 log analysis	37.83%	5.26%	56.91%
Experiment #2 questionnaire	31.58%	10.53%	52.63%
Experiment #3 laboratory study	14.29%	4.76%	80.95%

Table 3: Distribution of information needs of several studies.

[13] I. Kang and G. Kim. Query type classification for web document retrieval. In *Proc. of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, 2003.

[14] M. Kellar, C. Watters, and M. Shepherd. A field study characterizing Web-based information-seeking tasks. *American Society for Information Science and Technology*, 58(7):999–1018, 2007.

[15] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proc. of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 239–248, 2005.

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

[17] D. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of the 13th International Conference on World Wide Web*, pages 13–19, 2004.

[18] D. Russell and C. Grimes. Assigned tasks are not the same as self-chosen Web search tasks. In *Proc. of the 40th Hawaii International Conference on System Sciences*, pages 83–91, 2007.

[19] J. Teevan, C. Alvarado, M. Ackerman, and D. Karger. 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*, pages 422–429, 2004.

APPENDIX

A. SURVEY ABOUT THE SEARCH OF THE PORTUGUESE WEB ARCHIVE

```
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:
```