Characterizing Search Behavior in Web Archives

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

Web archives are a huge source of information to mine the past. However, tools to explore web archives are still in their infancy, in part due to the reduced knowledge that we have of their users. We contribute to this knowledge by presenting the first search behavior characterization of web archive users. We obtained detailed statistics about the users' sessions, queries, terms and clicks from the analysis of their search logs. The results show that users did not spend much time and effort searching the past. They prefer short sessions, composed of short queries and few clicks. Full-text search is preferred to URL search, but both are frequently used. There is a strong evidence that users prefer the oldest documents over the newest, but mostly search without any temporal restriction. We discuss all these findings and their implications on the design of future 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, Logs, User, Characterization

Keywords

Portuguese Web Archive, Search Behavior

1. INTRODUCTION

The web has a democratic nature, where everyone can publish all kinds of information. News, blogs, wikis, encyclopedias, interviews and public opinions are just a few examples. Part of this information is unique and historically valuable. However, since the web is too dynamic, a large amount of information is lost everyday. Ntoulas et al. discovered that 80% of the web pages are not available after one year [17]. In a few years they are all likely to disappear, creating a knowledge gap for future generations. Most of what has been written today will not persist and, as stated by UNESCO, this constitutes an impoverishment of the heritage of all nations [26].

Several initiatives of national libraries, national archives and consortia of organizations started to archive parts of the web to cope with this problem[1]. Some country code top-level domains and thematic collections are being archived regularly[2]. Other collections related to important events, such as September 11th, are created at particular points in time[3]. In total, billions of web documents are already archived and their number is increasing as time passes. The Internet Archive alone collected 150 billion documents since 1996. The historic interest in the documents is also growing as they age, becoming an unique source of past information for widely diverse areas, such as sociology, history, anthropology, politics or journalism. However, to make historical analysis possible, web archives must turn from mere document repositories into living archives. The development of innovative solutions to search and explore it are required.

Current web archives are built on top of web search engine technology. This seems like the logical solution, since the web is the main focus of both systems. However, web archives enable searching over multiple web snapshots of the past, while web search engines only enable searching over one snapshot of the close present. Users from both systems also have different information needs [4]. Hence, we also expected different search patterns and behaviors, which without a proper response, could degrade results and negatively influence users' satisfaction. We studied the above issues and drew the first profile of how web archive users search. It is based on the quantitative analysis of the Portuguese Web Archive (PWA) search logs [6].

Our results show that users of both types of systems have similar behaviors. In general, web search technology can be adopted to work on web archives. Nonetheless, our identification of the users' specificities provides insights on search behavior, that might contribute to better support the architectural design decisions of future web archives. Examples include optimizing their performance [1] or designing better web interfaces [8].

This paper is organized as follows. In Section 2, we cover the related work. In Section 3, we describe the search environment. 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.

TWAW 2011, March 28, 2011, Hyderabad, India.

[1] see http://www.nla.gov.au/padi/topics/92.html
[2] see http://www.archive.org/
[3] see http://www.loc.gov/minerva/

2. RELATED WORK

2.1 Web Archive User Studies

There are several web archiving initiatives currently harvesting and preserving the web heritage, but very few studies about web archive users. The International Internet Preservation Consortium (IIPC) reported a number of possible user scenarios over a web archive [7]. The scenarios are related to professional scopes and have associated the technical requirements necessary to fulfill them. These requirements include a wide variety of search and data mining applications that have not been developed yet, but could play an important role. However, the hypothetical scenarios did not come directly from web archive users.

The National Library of the Netherlands conducted an usability test on the searching functionalities of its web archive [21]. Fifteen users participated on the test. One of the results was a compiled list of the top ten functionalities that users would like to see implemented. Full-text search was the first one, followed by URL search. Strangely, time was not mentioned in none of the top ten functionalities, despite being present in all the processes of a web archive. The users' choices can be explained by web archives being mostly based on web search engine technology. As a result, web archives offer the same search functionalities. This inevitably constrains the users' behaviors. Another explanation is that Google became the norm, influencing the way users search in other settings.

In a previous publication, we studied the information needs of web archive users [4]. We resort to three instruments to collect quantitative and qualitative data, namely search logs, an online questionnaire and a laboratory study. Our observations were coincident. Users perform mostly navigational searches without a temporal restriction. Other findings show that users prefer full-text over URL search, the oldest documents over the newest and many information needs are expressed as names of people, places or things. Results also show that users from web archives and web search engines have different information needs, which cannot be effectively supported by the same technology.

2.2 Search Log Analysis

Web usage mining focuses on using data mining to analyze search logs or other activity logs to discover interesting patterns. Srivastava et al. pointed five applications for web usage mining: personalization, for adjusting the results according to the user's profile; system improvement, for a fast and efficient use of resources; site modification, for providing feedback on how the site is being used; business intelligence, for knowledge discovery aimed to increase customer sales; and usage characterization to predict users' behavior [25]. We focus on usage characterization. However, our results can be applied to other purposes, such as the efficiency and effectiveness improvements of IR systems [23].

Search logs capture a large and varied amount of interactions between users and search engines. This large number of interactions is less susceptible to bias and enables identifying stronger relationships among data. Additionally, search logs can be analyzed at low cost and in a non intrusive way. Most users are not aware that their interactions are being logged. Users also try to fulfill their real information needs, instead of having tasks assigned by a researcher that can bias their behaviors. 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. Qualitative studies, such as surveys and laboratory studies, can complement log analysis with information that can explain some of the patterns found [14].

Several logs from web search engines were analyzed with the goal of understanding how these systems were used. A common observation across these studies is that most users conduct short sessions with only one or two queries, composed by one or two terms each [11]. When users submit more than one query, they tend to refine the next query by changing one term at a time. Most users only see the first search engine results page (SERP) and rarely use advanced search operators. These discoveries imply that the use of web search engines is different from traditional IR systems, which receive queries three to seven times longer [12]. Queries for special topics (e.g. sex), special types (e.g. question-format) and multimedia formats (e.g. images) are also longer [16]. This shows that the users' behavior varies not only between IR systems, such as search engines, online catalogs and digital libraries, but also depends on the type of information and the way users search. Another aspect that differentiates search behavior is users' demographics (i.e. age, gender, ethnicity, income, educational level) [27].

3. THE SEARCH ENVIRONMENT

The PWA preserves the Portuguese web, which is considered the subset having the most interesting contents for the Portuguese community. Specifically, we define the Portuguese web as all the documents[4] 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. Additionally, the PWA team integrated web collections from several other sources, such as the Internet Archive and the Portuguese National Library. The number of indexed documents have been growing and there are now more than 180 million 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 a large time span (from 1996 to 2009). The experimental version of the PWA has been available as a service to the general public since 2010 at `http://archive.pt/`.

The interaction with the users and the layout of the results is similar to web search engines, such as Google. In a typical session, a user can submit a full-text query and receive a search engine results page (SERP) containing a list of 10 results matching the query. Figure 1 illustrates this case. Each result includes the title of the web page and its crawled date, a snippet of text containing the query terms and the URL. The user can then click on the results to see and navigate in the web pages as they were in the past. If the desired information is not found, the user can repeatedly modify and resubmit the query. In addition, the user can click on the navigation links to explore other SERPs or use the advanced search interface to restrict the query with advanced search operators. These operators can also be added to the query directly in the text box.

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

Figure 1: Search interface after a full-text search.

Figure 2: Search interface after an URL search.

This interface has some specificities. First, the text box is complemented with a date range filter to narrow the results to a time period. Second, each result has an associated link to see all versions throughout time of the respective URL. When clicked, the PWA presents the same search engine versions page (SEVP) as when a user submits that URL on the text box. A table is shown to the user, where each column contains all the versions of a year sorted by date. The user can then click on any version to see it as it was on that date. Figure 2 depicts this interface.

3.1 Logs Dataset

Our analysis is based on the logs of the PWA, covering seven months of search interactions, from June to December, 2010. By interactions, we mean all queries and clicks submitted by the users and recorded by the PWA search engine (server side). The seven month span has the advantage of being less likely to be affected by ephemeral trends.

The logs follow the Apache Common Log Format[5]. Each entry corresponds to an interaction with the search engine in the form of a HTTP request. It contains the user's IP address and the user's session identifier. Each entry contains also a timestamp indicating when the interaction occurred and the HTTP request line that came from the client.

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

[5]see `http://httpd.apache.org/docs/2.0/logs.html`

4. METHODOLOGY

The analysis focused on four dimensions: sessions, queries, terms and clicks. We define them in the following way:

- A *session* is a set of interactions by the same user when attempting to satisfy one information need. The session is the level of analysis in determining the success or failure of a search. It is composed by one or more queries and zero or more clicks.

- A *query* is a search request composed by a set of terms. We define an *initial query* as the first query submitted in a session, while all the following queries are defined as *subsequent*. An *identical query* is a query with exactly the same terms as the previous one submitted in the same session. A *unique query* corresponds to one query regardless of the number of times it was logged. The set of unique queries is the set of query variations. An *advanced query* is a query with at least one advanced search operator.

- A *term* is a series of characters bounded by white spaces, such as words, numbers, abbreviations, URLs, symbols or combinations between them. There are also advanced search operators, but they do not count as

terms. We define a *unique term* as one term on the dataset regardless of the number of times it was logged. The set of unique terms is the submitted lexicon.

- A *click* in this context refers to the following of a hyperlink to immediately view a query result (i.e. archived web page). It can be a *SERP click* or a *SEVP click*, depending if the user clicks in a SERP or SEVP.

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

4.1 Log Preparation

We prepared the log fields for analysis through a series of data cleansing steps. All incomplete entries, empty queries and sessions without any query were discarded. Internal queries submitted by the PWA monitoring system, the queries by example displayed on the PWA entry page and sessions conducted by clients identified as web crawlers were also excluded. Additionally, sessions with more than 100 queries were likely to come from crawlers, so they were removed too. This cutoff value of 100 was used in some other studies, thus enabling a more direct comparison with our results [10]. The queries that resulted from navigation clicks to see another SERP were not counted as a new query. These are the same queries parameterized to show more results.

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

4.2 Session Delimitation

Most studies used the users' IP address and/or session identifier to delimit sessions [11]. We used these two parameters to track and delimit user interactions. We also used a time interval t of inactivity to delimit sessions. Two consecutive interactions are included in different sessions if they have an inactivity between them of at least t. Without this gap, we could have sessions of several days, which would hardly represent the reality. Studies diverge on the choice of this interval, from 5 to 120 minutes [11], while others argue that no time boundary is effective in segmenting sessions [13]. We selected the 30 minute interval, because this interval has shown to produce good results, close to the results produced by SVM classifiers that were designed for delimiting sessions [20].

5. LOG ANALYSIS

Statistics were computed from the logged interactions. The first pattern that we detected was that users mostly conducted two types of sessions: with only full-text queries and with only URL queries, in 59.34% and 31.10% of the times, respectively. We defined these as *full-text sessions* and *URL sessions*. In the analysis, we ignored the remaining 9.57% sessions with mixed queries for simplification.

Table 1 shows the general statistics. The users of the PWA performed 6,177 full-text sessions, averaging 2.23 queries per session. The number of query terms per query was 2.84, with 6.42 characters per term. The users saw 1.44 SERPs per query and clicked 1.06 times on their hyperlinks to view a result. They hardly clicked to see all versions of a result. This only happened in 0.06 times per query. Overall, these results mean that for each query, the users saw mostly the first and sometimes the next SERP, where they clicked once.

	full-text	URL
Sessions	6,177	3,237
Queries	13,770	4,986
Terms	39,132	-
SERPs	19,812	-
Clicks on SERPs	14,664	-
Clicks on SEVPs	-	3,861
Queries per Session	2.23	1.54
Terms per Query	2.84	-
SERPs per Query	1.44	-
Clicks on SERP per Query	1.06	-
Clicks on SEVP per Query	-	1.56
Characters per Term	6.42	27.27
Initial Queries	44.86%	64.92%
Subsequent Queries	55.14%	35.08%
- Modified	44.53%	-
- Identical	20.35%	21.44%
- Terms Swapped	3.75%	-
- New	31.37%	78.56%
Unique Queries	68.82%	73.95%
Unique Terms	26.66%	-
Queries never repeated	54.38%	59.99%
Terms never repeated	13.88%	-

Table 1: General statistics.

Session duration	% full-text sessions	% URL sessions
$[0, 1[$	59.93%	81.19%
$[1, 5[$	23.07%	12.42%
$[5, 10[$	6.22%	2.97%
$[10, 15[$	2.77%	1.95%
$[15, 30[$	4.95%	1.02%
$[30, 60[$	2.19%	0.46%
$[60, 120[$	0.73%	0.00%
$[120, 180[$	0.10%	0.00%
$[180, 240[$	0.05%	0.00%
$[240, \infty[$	0.00%	0.00%

Table 2: Session duration (minutes).

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

5.1 Session Level Analysis

5.1.1 Session duration

The duration of a session is measured from the time the first query is submitted until the last time the user interacted with the PWA. We ignore if the user spent more session time viewing the archived web pages after the last interaction or used part of the time doing parallel tasks [19]. We assigned a 0 minutes duration to sessions composed by only one query.

The large majority of sessions ended quickly as shown in Table 2. Around 60% of the full-text sessions lasted less than 1 minute and 89% less than 10 minutes. Only around 3% of the sessions had a longer than an half hour duration. Each session took in average 4 minutes and 8 seconds. URL sessions took even less time than full-text sessions. In average, each session took 1 minute and 14 seconds. Around 81% of the sessions lasted less than 1 minute and only 6% took longer than 5 minutes.

# queries	% full-text sessions	% URL sessions
1	64.98%	72.10%
2	12.53%	15.57%
3	7.48%	6.21%
4	5.00%	3.06%
5	2.72%	1.11%
6	1.65%	0.56%
7	1.12%	0.74%
8	0.68%	0.28%
9	1.26%	0.19%
≥10	0.58%	0.09%

Table 3: Number of queries per session.

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

Table 4: Number of terms changed per modified full-text query.

5.1.2 Query distribution

Table 3 shows that the majority of the users only submitted one query. Around 85% of the full-text sessions had up to 3 queries and less than 3% had 10 or more queries. This last number can represent highly motivated users searching for special topics (e.g. sex) [16].

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

5.2 Query Level Analysis

5.2.1 Modified queries

Sometimes users submit sequences of queries as a way to refine or reformulate the search in a trial and error approach. We consider that two sequential queries submitted on the same session have the same information need if they share at least one term. In this case, we called the second query a modified query. We ignored the stopwords (too common terms) in this analysis. Thus, a modified query could be a specialization of the query (adding terms), a generalization (removing terms) or both at the same time.

We counted 44.53% of modified queries from all subsequent full-text queries. Looking at Table 4, we see that around 71% of the modified queries are the result of a zero or one change on the number of terms. A zero length change means that the users modified some terms, but their number remained the same. Users tend to add more terms in the modified queries rather than to remove them. We counted around 42% versus 25%. PWA's users tend to go from broad to narrow queries, such as in web search engines [3, 12, 22].

advanced operator	% advanced queries	% total queries
NOT	3.62%	0.94%
PHRASE	78.10%	20.20%
SITE	12.81%	3.31%
TYPE	5.48%	1.42%
total	100.00%	25.86%

Table 5: Advanced operators per full-text query.

5.2.2 Identical and New queries

A variety of reasons can lead users to repeat queries, such as a refresh of the SERP or SEVP, a back-button click or the submission of the same query more than once due to a network or search engine delay. When analyzing the full-text queries, we counted 20.35% of identical queries, where each query has exactly the same terms as the previous one made in the same session (see Table 1). We also counted the subsequent queries with the same terms, but written in a different order. For instance, a query *Web Archive* followed by a query *Archive Web*. Only a small number of subsequent queries, 3.75%, had the order of the terms swapped. Besides the modified and identical queries, the users also submitted 31.37% of subsequent queries with only new terms. This indicates that at most this percentage of subsequent queries were the result of a new information need.

We divided the subsequent URL queries in identical and new. 78.56% of the subsequent queries were new. The remaining 21.44% were the result of the same URL submission (see Table 1).

5.2.3 Advanced queries

In the PWA, users could use four advanced search operators: NOT, to exclude all results with a term in their text (e.g. *-web*); PHRASE, to match all results with a phrase in their text (e.g. *"web archive"*); SITE, to match all results from a domain name (e.g. *site:wikipedia.org*); TYPE, to match all results from a media type (e.g. *type:PDF*).

Table 5 presents the percentages of advanced queries (i.e. with at least one advanced search operator). It shows that 25.86% of the queries included operators. This is a significantly higher percentage when compared with studies over web search engines [9, 12, 22]. The reason is the PHRASE operator, which represents 78.10% of the choices. The PWA suggested a URL within quotes for each URL submitted, to inform the users that they could match the URL in the text. However, even when ignoring the URLs within quotes, the percentages are roughly the same. The second most used operator was the SITE, occurring in 12.81% of the advanced queries. The TYPE and NOT operators were insignificantly used when compared to the total number of queries.

5.2.4 Term distribution

The distribution of the terms per full-text query listed in Table 6 shows that the majority of the queries had 1 or 2 terms. This is also visible by the 2.84 average of terms per query (see Table 1). Around 87% of the queries had up to 5 terms and only 3% had 10 or more terms. These results indicate that the users tend to submit short queries. These values are useful, for instance, to optimize index structures [15] or to determine the adequate length of the input text boxes on the user interface [8].

# terms	% full-text queries
1	35.77%
2	24.99%
3	15.14%
4	7.54%
5	3.55%
6	4.47%
7	2.40%
8	1.92%
9	1.46%
≥10	2.77%

Table 6: Number of terms per query.

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

Table 7: SERPs viewed per query.

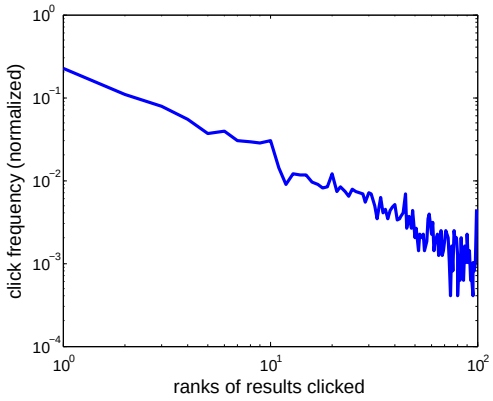

Figure 3: Distribution of ranks clicked on SERPs.

5.2.5 SERPs

The users saw on average about 1.44 SERPs per full-text query. All users saw the first SERP as expected, since the PWA always returned it after a query. Then, the users followed the natural order of the SERPs, but in a sharp decline (see Table 7). For instance, the second SERP was viewed in 14.44% of the queries. This indicates that prefetching the second SERP would not significantly improve web archive performance. On the other hand, the close percentages of the following SERPs indicate that prefetching them can bring improvements as shown in other studies [5].

5.2.6 Clicks on SERPs

About 66% of the clicks occurred on the first SERP from almost a click per query. The users clicked on 1.06 times per query to access an archived web page listed on the SERPs. We observed that users clicked on the rank of results following a power law distribution, with a 0.88 correlation (see Figure 3). These results are similar to web search engine studies, which also present a discontinuity in the last ranking position of each SERP (multiple of 10) [2].

5.2.7 Query frequency distribution

We ranked the full-text unique queries by their decreasing frequency and verified that their distribution fits the power law with a 0.96 correlation. This finding was also observed in web search engines [1, 5]. It means that a small number of queries was submitted many times, while a large number of queries were submitted just a few times. Figure 4 depicts the cumulative distribution of queries. For instance, by caching around 27% of the most frequent queries, the PWA could respond to 50% of the total query volume.

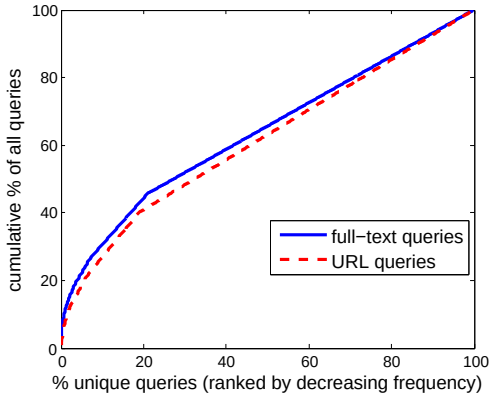

Figure 4: Cumulative distributions of queries.

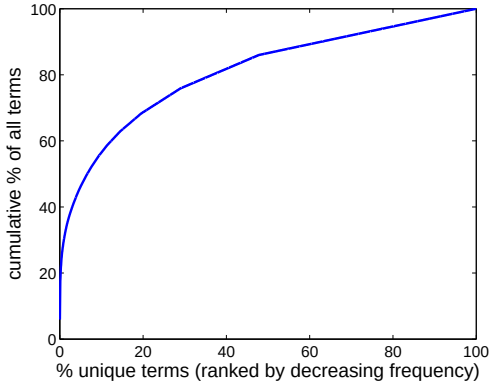

Figure 5: Cumulative distribution of full-text terms.

We also ranked the URL unique queries by their decreasing frequency and verified that their distribution, once again, fits the power law with a 0.96 correlation. By caching around 32% of the most frequent URL queries, the PWA could respond to 50% of the queries. Although satisfactory, the percentage of queries cached are much superior than in previous studies [3]. This is likely due to the small number of sessions analyzed, which leads to a reduced repetition.

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

5.3 Term Level Analysis

5.3.1 Term frequency distribution

Analogous to the query frequency distribution, we ranked the full-text unique terms by their decreasing frequency. Their distribution fits the power law with a 0.97 correlation. As depicted in Figure 5, the cumulative distribution shows that it is necessary to cache just around 6% of the most frequent terms to handle 50% of the queries. Much less RAM is necessary to cache terms than queries for a similar hit rate. These results are consistent with others presented for web search engines [1, 3]. However, caching the terms instead of the queries, adds extra processing over the posting lists of the inverted index, to evaluate the documents matching the query. A proper trade-off must be found.

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

Table 8: Queries restricted by date.

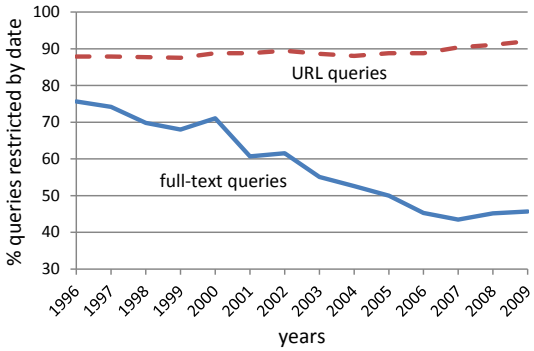

Figure 6: Distribution of years included in queries restricted by date.

5.4 Temporal Level Analysis

5.4.1 *Queries restricted by date*

The users restricted by the end date 23.55% of the full-text queries, while only 1.64% by the start date. The start and end dates were both changed in 12.98% of the queries. The same pattern exists in URL queries as shown in Table 8, where the start date was changed almost only when the end date also was. This indicates that users are more interested in old documents. The idea is reinforced by the distribution of the years included in the full-text queries restricted by date. As it can be seen in Figure 6, the older the years, the more likely they are of being included in queries. However, the URL queries have an almost constant rate.

5.4.2 *Clicks on temporal versions*

Documents tend to have just a few years with archived versions, thus segmenting the number of clicks per year would likely bias the results. Instead, we computed for all URL queries, the percentage of clicks in each year y_i with at least one version. We measured as $\frac{clicks(y_i)}{times(y_i)}$, where the denominator represents the number of times the year y_i was displayed to the user, and the numerator the number of clicks in y_i. For instance, the first year y_1 is 1997 if there is no archived versions for that URL in 1996. Otherwise, y_1 is 1996.

In Figure 7 it is visible that users clicked much more on the first year with archived versions than on the remaining years. The first year was clicked in 55% of the times, while all the others were clicked at most 20%. With exception of the eighth year, the first three years had the higher percentages. This shows a preference for the older documents.

5.4.3 *Implicit temporal queries*

We counted the number of queries with temporal expressions, since they represent a temporal dependent intent. We started by experimenting named-entity recognition tools for Portuguese. However, queries are not grammatical, so the tools presented a small precision. Instead, we used a simple match of all the queries with years, months and day patterns. Then, we classified a random subset of 1,000 queries to validate our detection patterns. Surprisingly, they worked very well. The patterns achieved a precision, recall and accuracy, of 89%, 100% and 98%, respectively. The patterns created some false positives, but unexpectedly no false negatives. This was mostly, because there were no temporal expressions in the logs without date patterns (e.g. last decade).

All matches were manually validated, from which we excluded the false positives. In the end, we counted 3.49% of queries with temporal expressions. Almost all are related with past events, such as *world cup 2006*. This is a small percentage in line with the 1.5% of temporal expressions found in the logs of the AOL web search engine [18].

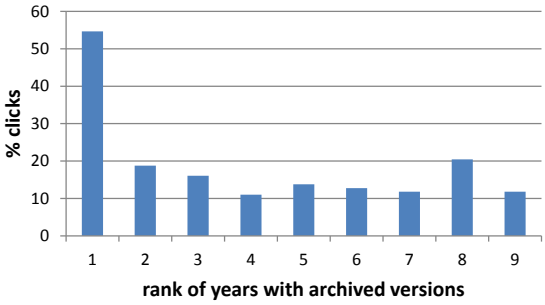

Figure 7: Clicks on years with archived versions (from oldest to newest).

6. DISCUSSION AND CONCLUSIONS

Search patterns from users of web archives and web search engines are contrasted in Table 9. Web archive users submit more single query sessions, which reflects in a smaller number of queries per session. In a nutshell, web archive users iterate less. This can be explained by most of the information needs of web archive users being navigational, contrary to the needs of web search engine users [4]. Web archive users search for known-items using names, titles and URLs, some within quotes, that give good clues of the desired information. Another explanation is that web archive users submit longer queries, which could lead to better results.

On the other hand, the single term queries, the SERPs viewed per query and the topic most seen, are in conformity with web search engine results [3, 10, 11, 12, 16]. The classification of searched topics in web archives followed a different taxonomy, so they are not directly comparable. Still, Commerce is the most searched topic for navigational queries and People for informational queries [4].

Overall, the search patterns of the users of both types of systems show no evidence precluding the adoption of web search engine technology for web archive search. This was a surprise to us, because users from both systems have different information needs. For instance, users said they wanted to see the evolution of a page throughout time, but they tend to click on one or two versions of each URL. All information needs of the users are focused on the past, but most of the user queries are not restricted by date, neither contain temporal expressions. Users search as in web search engines. This behavior may be the consequence of we having offered a similar interface, leading them to search in a similar way.

IR system	web search engine			web archive
world region	U.S.	Europe	Portugal	Portugal
name	**Excite** [11, 24]	**FAST** [11, 24]	**Tumba!** [3]	**PWA** [4] & this study
single query session	55%-60%	53%-59%	41%-50%	65%
queries per session	2.3	2.9	2.5 - 2.9	2.2
single term queries	20% - 30%	25% - 35%	40%	36%
terms per query	2.6	2.3	2.2	2.84
advanced queries	11% - 20%	2% - 10%	11% - 13%	26%
SERPs viewed per query	1.7	2.2	1.4	1.4
topic most seen	Commerce, Travel	People, Places	Commerce, Travel	Commerce & People

Table 9: General comparison between users from web search engines and web archives.

Hence, new types of interfaces must be experimented, such as the temporal distribution of documents matching a query or timelines, which could create a richer perception of time for the user and eventually trigger different search behaviors.

Nevertheless, the identification of the users' specificities might contribute to the development of better adapted web archives. We observe a strong preference in searching and seeing the oldest documents over the newest. This finding can be used in ranking results, when no other temporal data is given. The ranking should also be tuned for navigational queries when the query type is unknown. Queries, terms, clicked ranks and seen archived pages follow a power law distribution. This means that all have a small fraction that is repeated many times and can be explored to increase the performance of web archives.

The PWA is still experimental and has a much smaller user base than commercial web search engines. Still, we believe that the obtained results are general, but studies over larger datasets and from other web archives are necessary to confirm this. Our future work will use these results to improve the architecture and retrieval algorithms of the PWA.

7. ACKNOWLEDGMENTS

This work could not be done without the help and infrastructure of the PWA team. We thank Michel da Corte for her review of the paper and FCT (Portuguese research funding agency) for its Multiannual Funding Programme.

8. REFERENCES

[1] R. Baeza-Yates, A. Gionis, F. P. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. Design trade-offs for search engine caching. *ACM Transactions on the Web*, 2(4):1–28, 2008.

[2] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *Proc. of the 3rd Latin American Web Congress*, page 242, 2005.

[3] M. Costa and M. J. Silva. A search log analysis of a Portuguese web search engine. In *Proc. of the 2nd INForum - Simpósio de Informática*, pages 525–536, 2010.

[4] M. Costa and M. J. Silva. Understanding the information needs of web archive users. In *Proc. of the 10th International Web Archiving Workshop*, pages 9–16, 2010.

[5] T. Fagni, R. Perego, F. Silvestri, and S. Orlando. Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. *ACM Transactions on Information Systems*, 24(1):51–78, 2006.

[6] 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.

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

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

[9] C. Hölscher and G. Strube. Web search behavior of Internet experts and newbies. *Computer networks*, 33(1-6):337–346, 2000.

[10] B. Jansen and A. Spink. An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management*, 41(2):361–381, 2005.

[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. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000.

[13] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. of the 17th ACM Conference on Information and Knowledge Management*, pages 699–708, 2008.

[14] D. Kelly. *Methods for evaluating interactive information retrieval systems with users*, volume 3 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc., 2009.

[15] C. Lucchese, S. Orlando, R. Perego, and F. Silvestri. Mining query logs to optimize index partitioning in parallel web search engines. In *Proc. of the 2nd International Conference on Scalable Information Systems*, pages 1–9, 2007.

[16] K. Markey. Twenty-five years of end-user searching, Part 1: Research findings. *American Society for Information Science and Technology*, 58(8):1071–1081, 2007.

[17] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proc. of the 13th International Conference on World Wide Web*, pages 1–12, 2004.

[18] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. In *Proc. of the Advances in Information Retrieval, 30th European Conference on IR Research*, pages 580–584, 2008.

[19] S. Ozmutlu, H. Ozmutlu, and A. Spink. Multitasking Web searching and implications for design. *American Society for Information Science and Technology*, 40(1):416–421, 2003.

[20] 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.

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

[22] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, volume 33, pages 6–12, 1999.

[23] F. Silvestri. *Mining query logs: Turning search usage data into knowledge*, volume 4 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc., 2010.

[24] A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen. U.S. versus European Web searching trends. *SIGIR Forum*, 36(2):32–38, 2002.

[25] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):23, 2000.

[26] UNESCO. Charter on the Preservation of Digital Heritage. Adopted at the 32nd session of the General Conference of UNESCO, October 17, 2003. `http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf`.

[27] I. Weber and C. Castillo. The demographics of web search. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 523–530, 2010.