A Search Log Analysis of a Portuguese Web Search Engine

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

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

Abstract. We present a characterization of the information-seeking behavior of the users of a Portuguese web search engine, based on the analysis of its logs. We obtained detailed statistics about the users' sessions, queries, terms and searched topics over a period of two years. The results show that the users prefer fast and short sessions, composed of short queries and few clicks. The trend is towards a reduction of the number of interactions with the web search engine. We also discuss the specificities and interests of the Portuguese users and their implications on the development of better adapted web search engines.

1 Introduction

Web search engines are one of the most used systems on the Internet. Commercial web search engines, like Google and Yahoo!, receive hundreds of millions of queries per day [1]. Their goal is to satisfy the users' information needs as well as possible. Hence, it is necessary to understand what and how users search, what they expected and the difficulties they face when seeking information.

User studies analyze user behavior through several methods, some quantitative, such as surveys and log mining [1, 2], and others qualitative, such as observations and think-aloud protocols [3, 4]. Qualitative analysis can provide valuable insights on the usability of systems and user satisfaction. However, the time spent experimenting with 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 of qualitative methods is their intrusiveness in the search process. Just the fact that the users are aware of being observed can affect their normal behavior.

On the other hand, 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 between the data. Additionally, analyses of search logs can be cheaper and non intrusive. Previous studies based on search logs show that there are differences between users from different world regions. The users' behavior reflects their distinct language, vocabulary and cultural bindings. For instance, Korean users submit on

[1] see blog.nielsen.com/nielsenwire/online_mobile/nielsen-reports-march-2010-u-s-search-rankings

INForum 2010 - II Simpósio de Informática, Luís S. Barbosa, Miguel P. Correia (eds), 9-10 Setembro, 2010, pp. 525–536

average queries with one term less than U.S. and European users, because their words are often compound nouns [5]. European users also search slightly differently than the U.S. users [6]. For instance, Europeans search more about people and places, while the U.S. users are more focused on e-commerce [1]. Hence, we considered it important to study the specificities of the Portuguese web search engine users and the degree to which the searching technology applied to other countries can be adopted.

This study draws the first profile of Portuguese users. It is based on the 2003 and 2004 search logs from the Tumba! web search engine [7]. As far as we know, this is also the first study about users whose native language is Portuguese, considered the fifth language with more users on the Internet [2]. This profile is slightly outdated since the logs are six years old. Web search engine companies have become increasingly reluctant of releasing their logs because of privacy concerns [8], so logs are now scarce and outdated. However, these logs enable us to directly compare the results with similar studies from the same period. This study is also a baseline for new research over more current logs and might contribute to the development of better adapted web search engines. Examples include optimizing their performance [9] or designing better web interfaces [10].

This paper is organized as follows. In Section 2, we cover the related work. In Section 3, we describe the logs dataset from which we based our study. 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

Several studies performed quantitative analysis of search logs in the past, with the goal of understanding how web search engines 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 [1]. 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 [11]. Queries for special topics (e.g. sex) and multimedia formats (e.g. images) are also longer [2].

A comparison by Spink et al. of the searching behaviors of users from the U.S. and Europe, pointed out a few differences [6]. Statistics for U.S. and European users were collected from the Excite and FAST web search engines, respectively. FAST users were mostly from Germany and Norway. U.S. users submitted on average more terms per query (2.6 vs. 2.3), but less queries per session (2.3 vs. 2.9) and accessed less SERPs (1.7 vs. 2.2). Longer queries may produce better results and this may explain why the U.S. users explored less SERPs. U.S. and European users also searched topics differently. U.S. users searched more about *Commerce, Travel, Employment or Economy*, while European users searched

[2] see `http://www.internetworldstats.com/stats7.htm`

more about *People, Places or Things*. However, Spink et al. only classified the English language queries for the supposed German and Norwegian users, which could have skewed the results.

The evolution of behaviors throughout time hardly changed. The longitudinal study of Spink et al. over Excite's users of 1997, 1999 and 2001, showed that the only significant difference was an increase, from 28.6% in 1997 to 50.5% in 2001, in the number of queries where only one SERP was viewed [12]. Another finding was that search topics shifted from entertainment and sex to commerce and people. The authors stated that e-commerce queries coincided with changes on the information distribution of the web. According to Lawrence and Giles, in 1999 about 83% of the web servers contained commercial content [13].

Jansen's analysis of the Altavista logs of 1998 and 2002, indicated that users evolved to longer queries, longer sessions, more modified queries and less results seen [14]. It seems that users became more persistent when searching for information. However, the results from 1998 could have been affected by the inactivity timeout of 5 minutes selected to delimit sessions, which is shorter than in posterior studies. Jansen and Spink analyzed users from FAST in 2001 and 2002 [15]. They detected a move toward a great simplicity in searching. Single query sessions and single term queries increased. The percentage of users refining queries decreased. On the other hand, users saw more SERPs. The frequency of topics searched by the European users also changed. There was a large increase from 22.5% to 41.5% of queries searching about *People, Places or Things*. On the other hand, there was a decline of more than 5% for queries searching about *Computers or Internet* and *Sex or Pornography*.

There were other important discoveries. Beitzel et al. showed that the volume of queries varies along the hours of the day and the days of the week [16]. The topics searched are also more or less popular at different times of the day. Hölscher and Strube discovered that expert users submit more complex queries and have more flexible searching strategies than the newbies [4]. Ozmutlu et al. pointed out that users may search for multiple topics in a single session [17].

3 Tumba!'s Logs Dataset

Tumba! was a search engine for the Portuguese web, which was available as a public service from 2002 to 2006 [7]. At the time, the Portuguese web was considered the subset of web pages satisfying one of the following conditions: (1) hosted on a site under a .PT domain; (2) hosted on a site under other domain (except .BR), written in Portuguese and with at least one incoming link from a web page hosted under a .PT domain. The interaction with the users and the layout of the results was similar to other web search engines, such as Google.

Our analysis is based on the Tumba!'s logs, covering two full years of search interactions, 2003 and 2004. By interactions, we mean all queries and clicks submitted by the users and recorded by the web search engine. This large time range has several advantages. First, we can see how the users evolved over the years. Second, it is less likely to be affected by ephemeral trends. Third, we can identify seasonal search patterns, for instance, during the Christmas season.

The logs follow the Apache Common Log Format (see `http://httpd.apache.org/docs/2.0/logs.html`). 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 (id). However, Tumba! did not register the session id in the log. Each entry contains also a timestamp indicating when the interaction occurred, the HTTP request line that came from the client and two parameters of the data sent back by the server, which are the HTTP response' status code and size. Figure 1 presents three entries of this log, which are purely illustrative.

```
213.22.91.10 - [03/Feb/2004:23:15:27 +0000] "GET ?q=lisbon&lang=pt HTTP/1.1" 200 19978
213.22.91.10 - [03/Feb/2004:23:15:31 +0000] "GET ?q=lisbon&lang=pt&start=10 HTTP/1.1" 200 21419
213.22.91.10 - [03/Feb/2004:23:15:33 +0000] "GET ?q=lisbon&lang=pt&start=10&
  click=pt.wikipedia.org/wiki/Lisbon&rank=12 HTTP/1.1" 200 18409
```

Fig. 1: Log entries format.

We never used the log data to match a real identity. However, we had to check that these logs did indeed correspond to Portuguese users. We counted 90% of Tumba!'s users with IP addresses assigned to Portugal and near 98% of the interactions were submitted through the Portuguese language interface. This strongly indicates that the users were mostly Portuguese.

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 that belong to 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 and 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 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 are not counted 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 in a SERP to immediately view a query result (i.e. web page).

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

4.1 Log preparation

Abnormal sessions and queries could skew the results of the study. Thus, 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 Tumba! watchdog 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 the queries submitted by users. This cutoff value of 100 was also used in some other studies, thus enabling a more direct comparison with our results [15, 14]. 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 since capitalization was ignored. Extra white spaces were removed and since the search engine did not perform stemming, all variations of a query term were considered as different terms. The set of query terms also includes punctuation, misspellings and mistakes.

4.2 Session delimitation

Previous studies used the users' IP address and/or session identifier (id) to delimit sessions. However, the Tumba! logs did not capture the session id. Hence, we cannot tell if the requests with the same IP came from different computers behind the same proxy server.

Most studies also used a time interval t of inactivity to delimit sessions. Two consecutive interactions are included on different sessions if they have an inactivity between them of at least t. This gap serves to separate two information needs of the same user, asked at different times. 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 minutes [18] to 30 minutes [19], while others argue that no time boundary is effective in segmenting sessions [20]. We selected the 30 minute interval, since 95% of the sessions are shorter and because 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 [21].

5 Log Analysis

The logged interactions and their parameters were statistically accounted. The users of the Portuguese web search engine Tumba! performed 254,728 and 133,827 searching sessions in 2003 and 2004, respectively. Analyzing the averages, the users submitted 2.94 queries per session in 2003 and 2.49 in 2004. In these two years, the average number of query terms was around 2.2, with nearly 7 characters per term. The users saw 1.4 SERPs per query and clicked around 0.7 times on their hyperlinks to view a result. This means that for each query, the users saw mostly the first and sometimes the next SERP, where they clicked at most once. Table 1 shows these general statistics. The results remained almost constant during the two years, except for a decrease in the number of queries per session. Next, we will detail our analysis and explain the remaining results.

Dataset	1 year-2003	1 year-2004
Sessions	254,728	133,827
Queries	749,914	333,871
Terms	1,630,392	738,576
SERPs	1,087,369	474,157
Clicks	584,161	240,961
Queries per Session	2.94	2.49
Terms per Query	2.17	2.21
SERPs per Query	1.45	1.42
Clicks per Query	0.78	0.72
Characters per Term	6.99	6.80
Initial Queries	33.97%	40.08%
Subsequent Queries	66.03%	59.92%
- Modified	32.80%	33.48%
- Identical	29.35%	32.71%
- Terms Swapped	0.26%	0.29%
- New	37.59%	33.52%
Unique Queries	44.03%	48.52%
Unique Terms	8.00%	10.33%
Queries never repeated	30.02%	34.04%
Terms never repeated	3.72%	4.77%

Table 1: General statistics.

Session duration	year 2003 % sessions	year 2004 % sessions	Δ
$[0, 1[$	43.18%	53.31%	+10.13%
$[1, 5[$	25.89%	21.67%	-4.22%
$[5, 10[$	10.69%	8.91%	-1.78%
$[10, 15[$	5.88%	4.80%	-1.08%
$[15, 30[$	8.89%	7.29%	-1.60%
$[30, 60[$	4.47%	3.33%	-1.14%
$[60, 120[$	0.93%	0.64%	-0.29%
$[120, 180[$	0.07%	0.04%	-0.03%
$[180, 240[$	0.01%	0.01%	0.00%
$[240, \infty[$	0.00%	0.00%	0.00%

Table 2: Session duration (minutes).

# queries	year 2003 % sessions	year 2004 % sessions	Δ
1	40.73%	49.52%	+8.79%
2	22.10%	21.10%	-1.00%
3	12.71%	10.86%	-1.85%
4	7.76%	6.09%	-1.67%
5	4.97%	3.84%	-1.13%
6	3.24%	2.43%	-0.81%
7	2.24%	1.61%	-0.63%
8	1.50%	1.17%	-0.33%
9	1.09%	0.78%	-0.31%
$\geq$10	3.67%	2.61%	-1.06%

Table 3: Number of queries per session.

5.1 Session Level Analysis

Session duration The duration of a session is measured as the time between the first query submitted until the last time the user interacted with the search engine. We ignore if the user spent more session time viewing web pages clicked from the SERP or used part of the time doing parallel tasks [17].

We can see on Table 2 that sessions ended quickly and the tendency is to end even faster. There was an increase, from 43.18% in 2003 to 53.31% in 2004, of the sessions with less than 1 minute. The average duration of the sessions also decreased, from 6 minutes and 31 seconds in 2003 to exactly 5 minutes in 2004. Around 80% of the sessions lasted less than 10 minutes and only less than 1% had a duration longer than one hour.

Query distribution Table 3 shows that the majority of the users did not go beyond their second query. Information retrieval is an iterative process, but the users hardly iterated. Around 90% of the sessions had up to 5 queries and only less than 4% had 10 or more queries. This last number can represent highly motivated users searching for special topics (e.g. sex) [2]. The results also show that there was an increase of almost 9% on sessions with only one query from 2003 to 2004. This is the main reason why the averages of the queries per session and the session duration decreased between the two years.

5.2 Query Level Analysis

Modified queries Sequences of queries are sometimes a way for users to refine or reformulate the search in a trial and error approach. A modified query is

# terms changed	year 2003 % modified queries	year 2004 % modified queries	Δ
≤-5	0.32%	0.40%	+0.08%
-4	0.48%	0.55%	+0.07%
-3	1.48%	1.71%	+0.23%
-2	5.01%	5.41%	+0.40%
-1	15.58%	15.77%	+0.19%
0	30.23%	28.97%	-1.26%
+1	36.52%	35.56%	-0.96%
+2	7.39%	8.23%	+0.84%
+3	1.97%	2.18%	+0.21%
+4	0.64%	0.79%	+0.15%
≥+5	0.37%	0.44%	+0.07%

Table 4: Number of terms changed per modified query.

defined as a subsequent query pertaining to the same information need and it is assumed that two queries have the same information need if they share at least one term. 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 32.80% in 2003 and 33.48% in 2004 of modified queries from all subsequent queries (see Table 1). Looking to Table 4, we see that more than 80% 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 47% versus 23%. As other users, Tumba!'s users tend to go from broad to narrow queries [11, 22, 18].

Identical and New queries Sometimes the users repeat queries. This can happen for a variety of reasons, such as a refresh of the SERP, a back-button click or the submission of the same query more than once due to a network or search engine delay. We counted 29.35% in 2003 and 32.71% in 2004 of identical queries (see Table 1), where each query is exactly the same as the previous one made in the same session. We also counted the subsequent queries with the same terms, but written in a different order. For instance, a query *Lisbon Portugal* followed by a query *Portugal Lisbon*. Only a small number of subsequent queries, 0.26% in 2003 and 0.29% in 2004, had the order of the terms swapped. Besides the modified and identical queries, the users also submitted in the same session, 37.59% in 2003 and 33.52% in 2004, of subsequent queries with only new terms (see Table 1). This indicates that at most, around of one third of the subsequent queries are the result of a new information need.

Advanced queries An advanced query is a query with at least one advanced operator. In Tumba!, the users could use three advanced operators: NOT, to exclude all results with a term in their text (e.g. *-Lisbon*); PHRASE, to match all results with a phrase in their text (e.g. *"cities of Portugal"*); SITE, to match all results from a domain name (e.g. *site:wikipedia.org*).

Table 5 contains the percentages of advanced queries. It shows that only, 12.79% in 2003 and 11.40% in 2004, of the queries included advanced operators.

advanced operator	year 2003		year 2004		Δ
	% adv. queries	% total queries	% adv. queries	% total queries	% adv. queries
NOT	2.91%	0.37%	3.60%	0.41%	+0.69%
PHRASE	43.15%	5.52%	49.93%	5.69%	+6.78%
SITE	53.95%	6.90%	46.46%	5.30%	-7.49%
total	100.00%	12.79%	100.00%	11.40%	

Table 5: Advanced operators per query.

SERP viewed	year 2003 % queries	year 2004 % queries	Δ
1	100.00%	100.00%	0.00%
2	16.76%	14.38%	-2.38%
3	8.56%	7.41%	-1.15%
4	5.20%	4.52%	-0.68%
5	3.57%	3.10%	-0.47%
6	2.54%	2.25%	-0.29%
7	1.93%	1.73%	-0.20%
8	1.51%	1.40%	-0.11%
9	1.25%	1.18%	-0.07%
≥10	5.74%	6.71%	+0.97%

Table 6: SERPs viewed per query.

The small use of advanced operators is in accordance with previous studies [4, 18, 11, 3]. The SITE and PHRASE operators divided the preferences, being used in more than 43% of the advanced queries each. The NOT operator was used in less than 4% of the advanced queries and was insignificantly used when compared to the total number of queries. Overall, it seems that the users were unfamiliar with the advanced operators. Eastman and Jansen suggest that the low presence of advanced operators is due to the little or no benefit they provide [23].

SERPs The users saw on average about 1.4 SERPs per query on the two analyzed years. Table 6 presents the SERPs viewed per query. All users saw the first SERP as expected, since the search engine always returned it after a query. Then, the users followed the natural order of the SERPs, but in a sharp decline. For instance, the second SERP was viewed in 16.76% of the queries. This indicates that prefetching of SERPs would not significantly improve web search engine performance. The results also show slight decreases on all the SERPs viewed. For instance, from 2003 to 2004 the second SERP was viewed less 2.38%. Moreover, the percentage of sessions where the users viewed only the first SERP increased from 68.11% in 2003 to 76.66% in 2004.

Clicks The users clicked around 0.7 times per query to access a web page listed on the SERPs. We analyzed the results they clicked and observed that its distribution fits the power law, with a 0.98 correlation (see Figure 2). This is similar to other studies, which also present a discontinuity in the last ranking position of each SERP (multiple of 10) [24]. The results almost did not vary between the two years. More than 70% of the clicks occurred on the first SERP. This is a good indicator of the ranking quality of the Tumba! web search engine.

Term distribution The distribution of the terms per query listed in Table 7 shows that the length of the queries varied little from 2003 to 2004. The majority of the queries had 1 or 2 terms. This is also visible by the 2.2 average of terms per query (see Table 1). More than 93% of the queries had up to 4 terms and more than 99% up to 7 terms. These results indicate that the users tend to submit short queries, with each term having in average 6.99 characters in 2003 and 6.80 in 2004. These values are useful, for instance, to optimize index structures [25] or to determine the adequate length of the input text boxes on the interface.

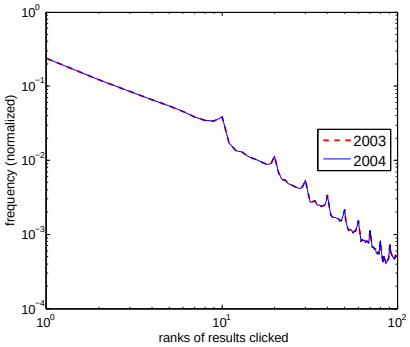

Fig. 2: Distribution of ranks clicked.

# terms	year 2003 % queries	year 2004 % queries	Δ
1	39.30%	39.98%	+0.68%
2	29.00%	26.87%	-2.13%
3	18.66%	18.84%	+0.18%
4	7.04%	7.37%	+0.33%
5	3.27%	3.66%	+0.39%
6	1.41%	1.66%	+0.25%
7	0.68%	0.85%	+0.17%
8	0.29%	0.37%	+0.08%
9	0.15%	0.18%	+0.03%
≥10	0.21%	0.22%	+0.01%

Table 7: Number of terms per query.

Query frequency distribution We ranked the unique queries by their decreasing frequency and verified that its distribution fits the power law as in other studies [9], with a 0.95 correlation. This means that a small number of queries were submitted many times, while a large number of queries were submitted just a few times. The 36,000 and 22,000 most frequent queries in 2003 and 2004, respectively, represent 50% of the total volume of submitted queries. However, they are only 11% in 2003 and 13.56% in 2004, of all the unique queries. Figure 3 depicts the 2003 and 2004 cumulative distributions of queries. We can see that, by caching a little more than 10% of the most frequent queries, the Tumba! search engine could respond to 50% of the query requests.

5.3 Term Level Analysis

Term frequency distribution Analogous to the query frequency distribution, we ranked the unique terms by their decreasing frequency. Its distribution also fits the power law, now with a 0.98 correlation. As depicted in Figure 4, the cumulative distribution shows that it is necessary to cache just around 1% of the most frequent terms, 1,200, 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 the ones presented by Baeza-Yates et al. [9]. However, caching the terms instead of the queries, adds the extra processing over the posting lists to evaluate the results matching the query. A proper tradeoff must be found.

5.4 Topical analysis

Table 8 lists the 20 most frequent queries and terms searched in 2003 and 2004. *Sexo* (sex) is the most searched query in both years, while *emprego* (job) is the second. Then, in 2003, like in 2004, the top 20 contains many queries related with sex and some queries related to the University of Lisbon, such as *mestrados* (master degrees), since Tumba! was also the University's site search engine. Other interests were games, the *totoloto* lottery, maps, chat, postcards, mp3 and music, the Benfica team and the Euro 2004 soccer event, humor and jokes, the eMule

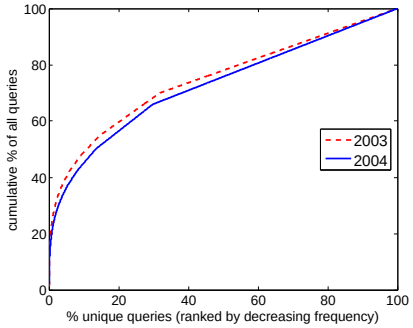

Fig. 3: Cumulative distributions of queries.

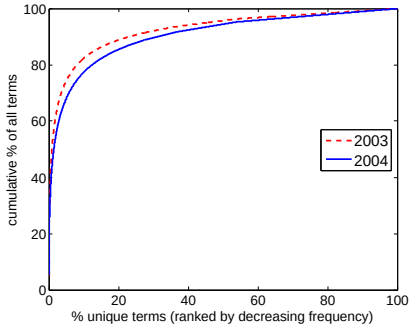

Fig. 4: Cumulative distribution of terms.

P2P program, taxes and photos. As a rough profile, we could say that the main concerns of the Portuguese users are sex, job and entertainment.

With the purpose of determining the types of information that people search for, we manually classified for each year, a random sample of 1,000 queries under the eleven general categories defined by Spink et al [12]. We chose this taxonomy for reasons of comparability with previous studies. To reduce bias, the same queries were independently classified by two evaluators, which then resolved their discrepancies. Table 9 shows the query percentages by topic categories. The most searched category was *Commerce, Travel, Employment or Economy*, with 22.40% of the queries in 2003 and 20.30% in 2004. This broad topic decreased 2.10%. On the other hand, the second most searched category, *People, Places or Things*, increased 2.90%, from 14.80% to 17.70%. Despite the sexual related queries being on the top of the most searched queries, the *Sex or Pornography* topic counts only with 4.90% of the queries in 2003 and 5.80% in 2004. Noteworthy, is also the 3.60% decrease of the *Entertainment or Recreation* topic.

6 Discussion and Conclusion

The Portuguese users, like other users, did not spend much time and effort on individual web searches. The Portuguese users submitted short sessions with short queries and few clicks, did not see beyond the first SERP and rarely used advanced operators. From 2003 to 2004 the average session duration and queries per session decreased. Sessions with less than one minute increased 10% and sessions with only one query increased 9%. The sessions where only the first SERP was viewed increased 8%. These results are in accordance with other results about the European users, which indicate that searching is moving toward a greater simplicity [15]. An analysis over an extended period is necessary to confirm this tendency. However, if verified, web search engines will be receiving less data while they cope with providing the same good results.

The Portuguese users have some peculiarities. The most common modification from users of other studies is to maintain the same number of terms when changing a query [11, 22, 18]. The Portuguese users on the other hand, tend to

rank	year 2003				year 2004			
	query	queries	term	terms	query	queries	term	terms
1	sexo	1.26%	sexo	1.03%	sexo	2.04%	sexo	1.39%
2	emprego	0.29%	portugal	0.40%	emprego	0.24%	portugal	0.42%
3	isep	0.14%	fotos	0.35%	emule	0.22%	fotos	0.33%
4	jogos	0.12%	lisboa	0.32%	jogos	0.15%	lisboa	0.28%
5	totoloto	0.10%	emprego	0.26%	chat	0.13%	jogos	0.24%
6	escola	0.10%	escola	0.23%	pornografia	0.13%	imagens	0.22%
7	mestrados	0.10%	porto	0.22%	totoloto	0.13%	2004	0.21%
8	pornografia	0.09%	jogos	0.21%	f******	0.12%	emprego	0.20%
9	porno	0.09%	imagens	0.17%	porno	0.10%	emule	0.20%
10	mapas	0.08%	mapa	0.17%	cadastro comercial	0.10%	download	0.19%
11	cadi	0.08%	trabalho	0.17%	anedotas	0.10%	porto	0.18%
12	chat	0.08%	gratis	0.16%	irs	0.09%	escola	0.18%
13	postais	0.07%	download	0.16%	travestis	0.08%	mapa	0.17%
14	mp3	0.07%	cursos	0.15%	mestrados	0.08%	lei	0.17%
15	benfica	0.07%	portuguesa	0.15%	euro 2004	0.08%	escolas	0.16%
16	humor	0.07%	universidade	0.15%	tumba	0.07%	gratis	0.16%
17	contos eroticos	0.07%	formação	0.14%	google	0.07%	trabalho	0.14%
18	acompanhantes	0.07%	2003	0.14%	contos eroticos	0.07%	comercial	0.14%
19	anedotas	0.07%	musica	0.14%	horarios	0.07%	portuguesa	0.14%
20	sexo gratis	0.07%	ensino	0.14%	sexo gratis	0.07%	cursos	0.14%

Table 8: The 20 most frequent searched queries and terms. Characters **** hide expletives.

	Categories	year 2003	year 2004	Δ
		% queries	% queries	
1	Commerce, Travel, Employment or Economy	22.40%	20.30%	-2.10%
2	People, Places or Things	14.80%	17.70%	2.90%
3	Health or Sciences	10.50%	11.80%	1.30%
4	Education or Humanities	7.20%	10.50%	3.30%
5	Society, Culture, Ethnicity or Religion	5.60%	6.10%	0.50%
6	Computers or Internet	6.40%	5.90%	-0.50%
7	Sex or Pornography	4.90%	5.80%	0.90%
8	Entertainment or Recreation	8.70%	5.10%	-3.60%
9	Government	7.00%	4.20%	-2.80%
10	Performing or Fine arts	1.60%	1.60%	0.00%
11	Unknown or Other	11.20%	11.30%	0.10%

Table 9: Topic categories of the queries.

refine the query by adding a term. For the other searching aspects, we can make the rough generalization that the European users submit less information to satisfy an information need than the U.S. users, but compensate by seeing more SERPs [6, 1]. The Portuguese users, which are also European, submit even less information and see even less SERPs than the other users. Table 10 summarizes these findings. We can speculate that these differences are due to the cultural differences of users, which are less tolerant and give up more easily. Other hypothesis is that the differences are due to the superior results' quality or superior interface that enabled the users to find the information sooner. The analysis over the search logs is insufficient to respond this question.

An important finding of this study is that the specificities of the Portuguese users do not preclude the general adoption of searching technology used by the U.S. and European users. On the other hand, the identification of these specificities can contribute to the development of better adapted web search engines. For instance, our results show that caching around 1% of the most frequent query terms enables response to 50% of the queries and caching the last query of a user in a session enables response to near a third of the queries.

world region search engine	U.S. Excite	Europe FAST	Portugal Tumba!
single term queries	20% - 30%	25% - 35%	40%
terms per query	2.6	2.3	2.2
queries per session	2.3	2.9	2.49 - 2.94
advanced queries	11% - 20%	2% - 10%	11% - 13%
SERPs viewed	1.7	2.2	1.4
topic most seen	Commerce, Travel	People, Places	Commerce, Travel

Table 10: General comparisons between users.

References

1. Jansen, B., Spink, A.: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing and Management **42**(1) (2006) 248–263
2. Markey, K.: Twenty-five years of end-user searching, Part 1: Research findings. American Society for Information Science and Technology **58**(8) (2007) 1071–1081
3. Aula, A., Khan, R.M., Guan, Z.: How does search behavior change as search becomes more difficult? In: Proc. of the 28th International Conference on Human Factors in Computing Systems. (2010) 35–44
4. Hölscher, C., Strube, G.: Web search behavior of Internet experts and newbies. Computer networks **33**(1-6) (2000) 337–346
5. Park, S., Ho Lee, J., Jin Bae, H.: End user searching: A Web log analysis of NAVER, a Korean Web search engine. Library and Information Science Research **27**(2) (2005) 203–221
6. Spink, A., Ozmutlu, S., Ozmutlu, H.C., Jansen, B.J.: U.S. versus European Web searching trends. SIGIR Forum **36**(2) (2002) 32–38
7. Costa, M.: Sidra: a flexible web search system. Master's thesis, University of Lisbon, Faculty of Sciences (November 2004) Also available as Technical Report DI/FCUL TR 4-17.
8. Barbaro, M., Zeller, T.: A face is exposed for AOL searcher No. 4417749. New York Times **9** (2006) `http://www.nytimes.com/2006/08/09/technology/09aol.html`.
9. Baeza-Yates, R., Gionis, A., Junqueira, F.P., Murdock, V., Plachouras, V., Silvestri, F.: Design trade-offs for search engine caching. ACM Transactions on the Web **2**(4) (2008) 1–28
10. Hearst, M.: Search User Interfaces. Cambridge University Press (2009)
11. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the Web. Information Processing and Management **36**(2) (2000) 207–227
12. Spink, A., Jansen, B., Wolfram, D., Saracevic, T.: From e-sex to e-commerce: Web search changes. IEEE Computer **35**(3) (2002) 107–109
13. Lawrence, S., Giles, C.: Accessibility of information on the web. Intelligence **11**(1) (2000) 32–39
14. Jansen, B., Spink, A., Pedersen, J.: A temporal comparison of AltaVista Web searching. American Society for Information Science and Technology **56**(6) (2005) 559–570
15. Jansen, B., Spink, A.: An analysis of Web searching by European AlltheWeb.com users. Information Processing and Management **41**(2) (2005) 361–381
16. Beitzel, S., Jensen, E., Chowdhury, A., Frieder, O., Grossman, D.: Temporal analysis of a very large topically categorized web query log. American Society for Information Science and Technology **58**(2) (2007) 166–178
17. Ozmutlu, S., Ozmutlu, H., Spink, A.: Multitasking Web searching and implications for design. American Society for Information Science and Technology **40**(1) (2003) 416–421
18. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. In: ACM SIGIR Forum. Volume 33. (1999) 6–12
19. Cacheda, F., Vina, A.: Understanding how people use search engines: a statistical analysis for e-business. E-work and e-commerce: novel solutions and practices for a global networked economy (2001) 319
20. Jones, R., Klinkner, K.L.: 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, New York, NY, USA, ACM (2008) 699–708
21. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: Proc. of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. (2005) 239–248
22. Spink, A., Wolfram, D., Jansen, M., Saracevic, T.: Searching the web: The public and their queries. American Society for Information Science and Technology **52**(3) (2001) 226–234
23. Eastman, C., Jansen, B.: Coverage, relevance, and ranking: The impact of query operators on Web search engine results. ACM Transactions on Information Systems (TOIS) **21**(4) (2003) 383–411
24. Baeza-Yates, R., Hurtado, C., Mendoza, M., Dupret, G.: Modeling user search behavior. In: Proc. of the 3rd Latin American Web Congress. (2005) 242
25. Lucchese, C., Orlando, S., Perego, R., Silvestri, F.: Mining query logs to optimize index partitioning in parallel web search engines. In: Proc. of the 2nd International Conference on Scalable Information Systems. (2007) 1–9