FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Characterization of Portuguese Web Searches

Rui Ribeiro

Master in Informatics and Computing Engineering

Supervisor: Sérgio Nunes (PhD)

July 11, 2011

Characterization of Portuguese Web Searches

Rui Ribeiro

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: João Pascoal Faria (PhD)
External Examiner: Daniel Coelho Gomes (PhD)
Supervisor: Sérgio Sobral Nunes (PhD)

5st July, 2011

Abstract

Nowadays the Web can be seen as a worldwide library, being one of the main access points to information. The large amount of information available on websites all around the world raises the need for mechanisms capable of searching and retrieving relevant information for the user. Information retrieval systems arise in this context, as systems capable of searching large amounts of information and retrieving relevant information in the user's perspective. On the Web, search engines are the main information retrieval systems. The search engine returns a list of possible relevant websites for the user, according to his search, trying to fulfill his information need.

The need to know what users search for in a search engine led to the development of methodologies that can answer this problem and provide statistical data for analysis. Many search engines store the information about all queries made in files called *transaction logs*. The information stored in these logs can vary, but most of them contain information about the user, query date and time and the content of the query itself. With the analysis of these logs, it is possible to get information about the number of queries made on the search engine, the mean terms per query, the mean session duration or the most common topics. This analysis is called *Query Log Analysis*. The results of this analysis can be compared with similar studies, to study eventual changes in users' behavior due to factors like: time, culture, language or others. This kind of analysis brings many advantages for the search engine themselves and its users; knowing what users search for, it is possible to improve the search engine's features anticipating and predicting user's behavior.

Analyzing a log from the Portuguese SAPO search engine covering a period of about six months in the year of 2010 some statistics were produced about user's sessions, queries, terms and searched topics. The original log stored about 140 million queries and contained queries about human users and queries made by bots or other automatic processes. As this study should focus only on the analysis of human queries, these queries were removed. Around 30% of the original queries from the dataset were considered as bots and removed. Queries from the same user were delimited by sessions. Two queries from the same user belong to different sessions if they have an inactivity between them of at least 30 minutes. Sessions with more than 100 queries were considered to be made by bots and were removed. Having a majority of users from Portugal (90%) this study present a characterization of the Portuguese community and the similarities and differences with other studies. Users made more queries on the beginning of the week, and on almost all weekends the daily query traffic is below average. The highest hourly traffic is seen between 2:00 p.m. and 5:00 p.m. where the hourly query frequency is 50% above average. Oddly the hourly distributions of human queries and queries made by bots are very similar. The results show that SAPO's users make short duration sessions, with 1 or

2 queries writing between 1 and 3 terms per query. Around 65% of the sessions have a duration lower than 1 minute, and almost 88% last less than 15 minutes. The mean session duration is 5 minutes and 23 seconds. More than half of the sessions only had 1 query and almost 90% of the sessions (85.92%) had up to 5 queries. Only about 5% of the sessions had 10 or more queries. Too common words, *function words*, like adverbs, propositions or pronouns were removed as they have little lexical meaning and are not relevant for the analysis process. Around 90% of the queries have at most 3 terms and only 1% of the queries have 7 or more terms. The mean terms per query is 2.03. Few queries and terms are unique (19.59% and 3.03% respectively). Users rarely use advanced operators (only in 1.5% of the queries) and when modifying or refining a query the number of terms stays unchanged or 1 term is added/removed. Analyzing a random sample of 2,500 queries it was observed that the main topic of interest was *Computers or Internet* accounting for 26.88% of the analyzed queries. The second and third categories were *People, Places or Things* and *Commerce, Travel, Employment or Economy* with 22.64% and 16.56%. These three categories have a similar hourly distribution with the highest query traffic between 10:00 a.m. and 4:00 p.m. and a clear decreasing after 8:00 p.m.. In many categories there is a clear downward tendency as the weekend approaches with some exceptions like *Entertainment or Recreation* which evidences an opossite tendency, rising until Friday dropping on Saturdays and rising again on Sundays.

Resumo

Actualmente, a Web pode ser descrita como uma biblioteca a nível mundial, sendo um dos principais pontos de acesso a informação. A informação dispersa pelos sítios web de todo o mundo faz com que sejam necessários mecanismos capazes de procurar e devolver a informação relevante para o utilizador. Neste âmbito, surgem os sistemas de recuperação de informação como sistemas capazes de pesquisar grandes quantidades de informação e devolver a informação relevante na perspectiva do utilizador. Na Web, os motores de pesquisa são os principais sistemas de recuperação de informação. Tendo por base a pesquisa feita pelo utilizador, o motor de pesquisa devolve uma lista de possíveis páginas relevantes para o utilizador, tentando desta forma satisfazer a sua necessidade de informação.

A necessidade de saber o que realmente pesquisam os utilizadores num motor de pesquisa e quais os seus comportamentos típicos fez com que se procurassem metodologias que respondessem a estas questões e fornecessem dados estatísticos passíveis de serem analisados. Tipicamente, os motores de pesquisa recolhem informação acerca de todas as pesquisas que são efectuadas. A informação é guardada em ficheiros como registos de transacções. Apesar do tipo de informação contida nestes registos variar, a maioria contém informação acerca do utilizador, data e hora da pesquisa e o conteúdo da pesquisa propriamente dita. Efectuando a análise da informação contida nestes registos, é possível obter informação acerca do número de pesquisas efectuadas no motor de busca, o número médio de palavras por pesquisa, o tempo médio de uma sessão de pesquisa de um utilizador ou mesmo quais os tópicos das pesquisas mais comuns; esta análise é designada por *Query Log Analysis*. Os resultados obtidos podem ser comparados com outros estudos semelhantes de forma a estudar eventuais variações do comportamento do utilizador devido a factores temporais, culturais, linguísticos, entre outros. Este tipo de análise trás muitos benefícios para o próprio motor de pesquisa e para os seus utilizadores. Sabendo o que os utilizadores mais pesquisam poderemos melhorar as funcionalidades do motor de pesquisa antecipando e prevendo os seus comportamentos.

Analisando um registo das pesquisas efectuadas no motor de pesquisa SAPO, cobrindo um período de cerca de 6 meses do ano de 2010, foram produzidas dados estatísticos acerca das sessões dos utilizadores, das suas pesquisas, termos e principais tópicos de interesse. O registo original possuía cerca de 140 milhões de pesquisas, contendo quer pesquisas feitas por humanos como pesquisas feitas por *bots* ou outros processos automáticos. Visto que a análise principal deveria ser focada nas pesquisas feitas por utilizadores humanos estas pesquisas foram removidas. Cerca de 30% das pesquisas da coleccção original foram consideradas como sendo feitas por *bots* e foram removidas. As pesquisas feitas por um mesmo utilizador foram divididas por sessões. Duas pesquisas pertencem a sessões diferentes se tiverem um período de inactividade de pelo menos 30

minutos. Sessões com mais de 100 pesquisas foram consideradas como sendo feitas por *bots* e foram também removidas. Visto que a maioria dos utilizadores era de Portugal (90%) este estudo representa uma caracterização da comunidade portuguesa e mostrar as semelhanças e diferenças com outros estudos. Foram feitas mais pesquisas no início da semana e em quase todos os fins de semana a percentagem de pesquisas está abaixo da média. O período horário com maior percentagem de pesquisas situa-se entre as 14:00 e 17:00 onde a frequência de pesquisas por hora está 50% acima da média. Estranhamente esta distribuição horária é muito semelhante à das pesquisas feitas por *bots*. Os resultados mostram que os utilizadores do SAPO preferem sessões de curta duração, fazendo apenas 1 ou 2 pesquisas contendo entre 1 e 3 termos. Cerca de 65% das sessões têm uma duração inferior a 1 minuto e quase 88% uma duração inferior a 15 minutos. A duração média por sessão é de 5 minutos e 23 segundos. Mais de metade das sessões têm apenas 1 pesquisa e quase 90% das sessões (85,92%) no máximo 5 pesquisas. Apenas cerca de 5% das sessões têm 10 ou mais pesquisas. Palavras muito comuns e sem relevância semântica como advérbios, preposições ou pronomes foram removidas visto que têm pouco contexto léxico e portanto não eram muito relevantes para o processo de análise. Cerca de 90% das pesquisas têm no máximo 3 termos e apenas 1% têm 7 ou mais termos. A média de termos por pesquisa é de 2,03. O número de pesquisas e termos únicos é baixo, 19,59% e 3,03% respectivamente. Os utilizadores do SAPO raramente usam operadores booleanos (apenas em 1,5% das pesquisas totais) e quando modificam uma pesquisa o número de termos existe uma alta probabilidade do número total de termos se manter igual ou de adicionar/remover 1 único termo. Pela análise de uma amostra aleatória de 2.500 pesquisas foi observado que o principal tópico de interesse era *Computadores ou Internet* o qual foi verificado em 26,88% das pesquisas analisadas. O segundo e terceiro tópico mais pesquisado foi *Pessoas, Sítios ou Coisas* e *Comércio, Viagens, Emprego ou Economia* com 22,64% e 16,56% respectivamente. As três categorias mais pesquisadas têm uma distribuição horária semelhante registando a maior frequência de pesquisas entre as 10:00 e 16:00 com um acentuado decréscimo dessa frequência após as 20:00. Em muitas categorias existe uma clara tendência de descida do número de pesquisas à medida que o fim de semana se aproxima com algumas exepções como a categoria de *Entertenimento e Recreação* que evidencia uma tendência oposta, com uma subida do número de pesquisas até a sexta-feira seguido de uma descida aos sábados e subindo de novo aos domingos.

Acknowledgements

My sincere thanks to Prof. Sérgio Nunes who not only served as my supervisor but also encouraged and challenged me throughout the thesis process. Special thanks to the SAPO company, in particular to their protocol with University of Porto, that provided the dataset and facilities to work on a daily basis. Deepest gratitude to all my friends and family for their support throughout these six months.

Rui Ribeiro

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

IR	Information Retrieval
QLA	Query Log Analysis
TLA	Transaction Log Analysis

Chapter 1

Introduction

The Internet became one of the most useful ways to have access to information. In just a few seconds, it is possible to find what you are searching for, with a high degree of precision. The Internet can be seen as a container, and the Web is a part of it; while the Internet is as a big collection of computer networks, the World Wide Web (WWW) or the Web utilizes that structure to offer content, documents, multimedia, etc. The Web is used daily by many different users across the world, with different information needs. In 2010, almost 70% of the European Union's population used the Internet, and also approximately half of the Portuguese population [Gro10].

Taking into account the large amount of information available on the Web, it is imperative that there are systems capable of retrieving the relevant information to fulfill the users' information needs; those are called Information Retrieval (IR) systems.

> *"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."* [CDMS08]

The process of information seeking can be described as follows: a person is facing a *problem* that requires information for being solved. The representation of the problem in the mind of the user is called *information need*, and it is different from the problem because the user might not comprehend it, in the correct way; the representation of the information need in a natural language, is named *request*. The representation of the information need in a "system" language is called *query* [Miz97]. In the Web context, the main information retrieval systems are Web search engines (e.g. www.google.com); these are websites where the user searches for something, and a list of relevant websites containing information related with the user's query, is presented.

To improve Web search engines, it is important to know what the users search for, in others words, what are their queries. To answer this problem, we need to know what

the users searched for, with which topic was the query related, how much time did the user spent searching and other similar questions. Web Usage Mining techniques cover these problems by analyzing Web log data. Query Log Analysis (QLA) deals with the study of query logs from data registered in a search engine [BYCBGC06]. Jansen [Jan06] defined QLA as *"the use of data collected in a transaction log to investigate particular research questions concerning interactions among Web users, the Web search engine, or the Web content during searching episodes"*. In QLA, search engines' transactional logs are analyzed. These logs store information about queries made on a search engine like the date and time of the query, the terms used or even the IP address of the user who made the query. QLA uses the data in transaction logs to recognize attributes of the search process such as the searcher's actions on a system, the system responses or the evaluation of results by the searcher [Jan06].

Using the statistic information retrieved from the QLA it is possible to increase the existing knowledge regarding how users use Web search engines enabling new features or improvements for the search engine itself. Furthermore, this information is really important to understand the users' behavior along the time.

1.1 Context and Main Goals

SAPO (Servidor de Apontadores Portugueses Online)[1] started in 1995 as a Web directory to respond to Portuguese users' information needs and evolved later into a search engine[2] with an interface that can be seen on Figure 1.1. SAPO was created by seven members of the Computer Science Center of the University of Aveiro [Tel06].

The main goal of this project is to characterize the Portuguese Web searches analyzing SAPO search engine's logs. These logs contain information regarding approximately 140 million records, from a recent period of time (January to July of 2010). Different people have different information needs, and search in different ways. With this analysis, it is possible to have a better understanding about the Portuguese Web searchers community and study their behavior. Having that information, it is possible to make a comparison with similar studies from different Web search engines. Furthermore, it provides solid knowledge for the SAPO search engine developers, giving them a tool to know their users' behavior and present them with new features.

[1] The name corresponds to the Portuguese word for toad.
[2] Available at http://pesquisa.sapo.pt/

Figure 1.1: SAPO Search Engine.

1.2 Document Structure

This document is organized as follows: in Chapter 2 the query log analysis and issues surrounding this topic are explained. In this chapter, some limitations in query log analysis are presented, mainly related with the queries' information collection process. Next, privacy and confidentiality concerns are discussed and some techniques to reduce users' real identity exposure risk are presented, as well as their advantages and disadvantages. In the final sub-chapter an overview of each one of the main QLA phases is made: collection, preparation and analysis, describing what is performed in each of these phases. In the analysis phase, automatic methods that can help in the analysis phase are discussed. In Chapter 3 some related studies are presented. In this chapter it is presented an analysis of six studies and some conclusions are drawn from the comparison of the results of these studies. Chapter 4 describes how the work was done, namely the tools used, how the dataset was organized and all the steps before the analysis process. In Chapter 5 the results of this study are presented. The results are showed for the three main levels of analysis (*Session, Query* and *Term* and a brief analysis of queries made by bots is also presented. Lastly, in Chapter 6 some conclusions are presented.

Chapter 2

Query Log Analysis

2.1 Overview

With the expansion of the Internet, and as more and more people use it, the need to know what and how a person searches, or use some service, is really important. One of the ways to get this information is by analyzing the transaction log files of information retrieval systems. A transaction log file, can be viewed as a file that has recorded the interactions between a user (searching for information) and an IR system [Pet93]. It can also be viewed as a method for automatically capture the type, content and time of the interactions made by a person with a IR system [RB83]. In the Web context, a transaction log is *"an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine"* [Jan06].

Search logs capture a large amount of interactions between users and search engines, and are less susceptible to bias, much because they capture the normal user behavior. Unlike qualitative methods (e.g. observations), there isn't anyone observing the user interacting with the search engine [CS10].

The analysis of these logs is named Transaction Log Analysis (TLA). TLA can be viewed as *"the study of electronically recorded interactions between on-line information retrieval systems and the persons who search for information found in those systems"* [Pet93]. Also it is the detailed and systematic examination of each search command or query by a user and the following database result or output [BBD+98]. One of the sub-categorizations of TLA is Query Log Analysis (also known as Search Log Analysis), meaning the analysis of search engine logs [JST08].

With QLA, it is possible to gain a clearer understanding of the interactions between searcher, content and systems. This opens a way to achieve some stated objective, such as improved system design, advanced searching assistance, or identified user information-searching behavior. Using this methodology, it is possible to examine search episodes in

order to isolate trends and typical interactions between searchers and the system [Jan06].

This is, already, an important research area which directly impacts pay-per-click marketing, Web-site-optimization strategies, and Web and Intranet search engine design [JS06]. Despite of collecting large quantity of clickstream data, few companies use this information effectively [SDP06].

2.2 Limitations

Query logs capture explicit description of users' information needs. These logs, capture the interactions that follow a user's query and derivate traces that further characterize the user and its interests [MT07]. Many researches criticized TLA as a research methodology. They state that transaction logs do not record the users' perceptions of the search, and therefore cannot measure the underlying information need of the searchers of the searchers or the satisfaction with the obtained results [Jan06]. Kurth comments [Kur93] that transaction logs can only deal with the actions the user takes, not their perceptions, emotions or background skills. Transaction logs are mainly a server-side data collection, therefore some users' interactions are masked from these logging mechanisms (e.g. click on back or print button) [Jan06]. Limitations and faults were also pointed to measures and metrics used in QLA. People can be logged on to the Web but not using it; it is difficult to associate an IP to an individual (many people could use the same IP address); the fact that a page was downloaded does not mean that anyone actually wanted it [NHLW99].

Some applications were developed to remedy the server-side data collection limitations. Velayathan and Yamada presented their work [VY07] in which connections between user interest and user behavior were explored and offered an alternative method for evaluating Web pages by incorporating client side logs. Kelly [Kel04] used a software package that tracks a person's computer activities. Jansen developed an application [Jan06] to be used with transaction logs and other IR studies. This application simply logs interaction with the IR system, along with other applications, using Dynamic Data Exchange (DDE), and outputting the data to a text file. This application can log a wide range of user interactions, including interactions with the browser toolbar, the system clipboard, scrolling results, among others.

Many of these client-side applications have obvious improvements over the typical server-side methods; despite these improvements, they come with some disadvantages (e.g. privacy concerns) that can discourage their use.

2.3 Privacy and Confidentiality Concerns

Query log analysis poses an obvious tradeoff: the various advantages of log analysis bring at the same time privacy concerns namely in the aspects of users' confidentiality; a proper balance between these two factors must be achieved. One of the biggest challenges of QLA is sharing information without compromising user privacy. When it is possible to associate the searcher with a real identity, log analysis assaults one of the most basic principles, a person's privacy. Cases like the AOL scandal [Tim06], raised new questions about if log data can be anonymized and shared; thus, nowadays public logs are scarce and outdated.

Certain approaches try to resolve these problems, but some of them have impact in the usefulness of the data. Eytan Adar [Ada07], stated that queries that are highly specific to an individual are of seldom occurrence; a possible solution would be to store queries only from a minimum number of occurrences, hence reducing the risk of exposure at the cost of raising the difficulty in identifying new queries. Murray and Teevan [MT07], showed that the meaning of privacy is misleading, mostly because our understanding of privacy has shifted. Different nations have very different notions about what protections an individual deserves. The browser plug-in TrackMeNot[1] has a way to protect user's privacy. This software sends large quantities of pseudo-random queries from a user's browser to mask that user's real query.

The technology alone cannot solve the problems associated with privacy and most of the techniques presented have side effects on the usefulness of the data [MT07].

2.4 Query Log Analysis Phases

Despite the fact that the decisions made in a QLA process may vary, there are some common steps. Jansen, enumerated three major steps in a TLA [Jan06]:

- **Collection**: the process of collecting the interaction data for a given period in a transaction log;
- **Preparation**: the process of cleaning and preparing the transaction log data for analysis;
- **Analysis**: the process of analyzing the prepared data.

For a Web search engine, the main goal of this process is to collect data about the transactions made between the users and the search engine. The type of information to be collected must be defined so that proper analysis of that information can be done later.

[1] http://cs.nyu.edu/trackmenot/

2.4.1 Collection

One of the earlier decisions that researches should do is to decide what type of content to collect from a given interaction of a searcher with the search engine. This decision depends on various aspects, like what needs to be investigated, what resources are available, what is the frequency of data collection, and so on [JST08]. The taxonomy of user-system interactions has many states like *view results* of the search made, *selection* of some page link, *execute* by searching for some query and others [JM05]. The decisions made in this phase are very important because the collection of the right data will allow researchers to make deductions regarding searchers behaviors.

The use of transactional logs is a good way for collecting data in a unobtrusive way, in other words the normal behavior of the searcher is not changed; furthermore this method does not interfere with the information retrieving process. Most of the transaction logs are primarily a server-side data collection, with known limitations (discussed in Chapter 2.2). Despite different transaction logs record different types of data, there is common information in all of them. The majority of transaction logs include information about:

- **User Identification**: the IP address of the searcher computer;
- **Date/Time**: the date and time of the interaction with the search engine;
- **Search Query**: the query terms entered by the user.

Other common fields recorded are information about *Language* and *Page Viewed* [Jan06]. Figure 2.1 shows the log format used in Tumba!'s study. This log follows the Apache Common Log Format [Fou].

```
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
```

Figure 2.1: Tumba!'s log entries format [CS10].

2.4.2 Preparation

Before the collected data can be analyzed it is necessary to *clean* the information contained in the log, because not all of the information is relevant and some can be misleading for the analysis process. In this phase we should remove abnormal data that introduces bias to the results of the study. This information can be of two levels: query level and session level.

In terms of queries, information about incomplete queries and empty queries should be removed. Also, because there are queries made by non-human (e.g. Web crawlers[2])

[2]Programs used by search engines to find out and download Web documents

a maximum number of queries per session should be defined. Some studies used 100 queries as this value [CS10, JSP05, JS05]. This number is almost 50 times greater than the reported mean search session [JSS00], and provides a good threshold to remove some non-human searches. Still, distinguishing between human and non-human searchers can not be done accurately [JS05].

Most of the studies use a gap to delimit the sessions. This means that if a user inserts a new query within the defined gap, the query is part of the same session; otherwise the new query is part of a new session. The value for this gap differs between studies, and it is possible to find studies that use values from 5 minutes [SMHM99] to 30 minutes [CS10].

The preparation phase has an important role to make the information on the log more relevant and accurate.

2.4.3 Analysis

The core of the QLA process is obviously the analysis phase. At this stage, the data in the log is analyzed with the main purpose of providing results for some metrics and compare them with similar studies. This process is, in many studies, done by using text-processing scripts or relational databases, given that the logs are usually stored in ASCII text files. There are few studies that actually give a precise guide on how to do these analysis; Jansen presented a stepwise methodology [Jan06], using relational databases, to conduct log analyzes. This analysis can bring improvements for search engines in terms of performance [BYGJ+08] or even design [Hea09]. With QLA it is possible to know the searchers' behavior, so the indexing methods can be improved and new features be provided by the search engine.

In a way to standardize QLA studies, Jansen provided a common language [JP01], which defined metrics and levels of analysis. The analysis should focus on three levels: the **session**, the **query** and the **term**. The **session** represents the entire sequence of queries, entered by the same user within a limited duration, to address one or more information needs [JS06]. At session level it is common to make analysis of the session's duration and number of queries per searcher. Session's duration is the total time between the user's first query until the time he leaves the search engine. A **query** is composed by a string of zero or more characters inserted into the search engine. The first query made by a particular searcher is named *initial query*; a subsequent query made by the same searcher, and identical to his previous queries is referred as a *repeated query*. A subsequent query by the same searcher that is different than any of the searcher's previous queries is a *modified query*. The number of *unique queries* is the amount of distinct queries in the dataset regardless the number of time they were logged. At query level, it is interesting to determine query length, the use of complex operators or query frequency. A **term** is defined as a string of characters separated by some delimiter such as a space,

a colon, or a period. The researcher should decide what delimiter to use. The analysis at this level provides results for term frequency or topical analysis [JP01]. Some studies also provide information about *clicks*, being possible to have statistical information about the search engine results pages seen by users and frequency data about clicks [CS10].

To overcome certain limitations of typical QLA methods, automatic topic discovery methods should help researchers when used in conjunction with search engines' logs. Gravano et al. proposed a categorization scheme for queries based on their geographical locality [GHL03]. By defining queries as *global*, their best matches are broad, global pages, not localized pages with a limited geographical scope; queries defined as *local* often include a location name or implicitly request "localized" results (e.g. the query "houses for sale"). For example, a Web page with general information about wildflowers could be considered as a global page, likely to be of interest to a geographically broad audience; in contrast, a Web page with information about houses for sale in a specific city could be seen as a local page, likely to be of interest only to a an audience in a relatively narrow region. Depending on the query character of local or global, this query is best answered by Web pages of the same geographical type. Automatic methods are very interesting from the point of view of discovering and categorizing topics. The information in the logs provides implicit feedback that is very valuable; the terms in queries can be used to describe the topic that users were trying to find. For example, if many users reach a document using certain keywords, then it is very likely that the information in this document can be summarized by those words. Having this in mind, Poblete et al. [PBY08] created a method based on frequent query patterns that shows clear results of improvement in the quality of results given by the IR system. Beitzel and Lewis [BJF+05] showed an approach for mining vast amounts of unlabeled data in search logs. Combining manual matching and supervised learning allowed to classify a larger proportion of queries than other techniques. The idea of supervised learning consists of training a classifier on the manual classifications to enable the classification of new queries in the respective categories.

Automatic topic detection methods could provide a great help categorizing users' queries; the large amount of data in search logs, demands the use of such automatic methods.

Chapter 3

Characterization of Web Searches

There are many studies about log analysis of Web search engines. In all of them, the analysis is made at different levels and the results are compared with other similar studies. Table 3.1 summarizes the most common variables of analysis in these three levels. Researchers tend to analyze the logs at the three main levels: query, session and term. Table 3.2 summarizes the information gathered in some of these studies.

Table 3.1: Analysis Variables.

Query Level	**Session Level**	**Term Level**
Number of Queries	Number of Sessions	Number of Terms
Unique Queries	Queries per Session	Unique Terms
Initial Queries	Session Duration	Characters per Term
Subsequent Queries		Term Frequency Distribution
Modified		Topical Analysis
Identical		
New		
Terms Swapped		
Advanced Queries		
Terms per Query		
Query Frequency Distribution		

To present the results and conclusions of the studies made in this area, some of them where selected as a way of comparison. The information was collected from different studies: (1) a 1998 study of the AltaVista Web search engine [SMHM99], (2) a 1999 study of the Excite Web search engine [WSJS01], (3) a 2001 study of the AlltheWeb.com Web search engine [JS05], (4) a 2002 study of Altavista Web search engine [JSP05], (5) and (6) 2003 and 2004 studies of Tumba! Web search engine [CS10]. Following Jansen and Spink's view [JS06], we can group the studies from the geographical perspective of the Web search engine; there is an European and a US grouping. Thus, studies of

Table 3.2: Web Search Engine Studies.

Search Engine	Data Collection	Queries	Terms	Sessions
Excite [JSS00]	16 September 1997	1,025,908	1,277,763	211,063
Fireball [HS00]	1-31 July 1998	16,252,902	Not Reported	Not Reported
AltaVista [SMHM99]	2 August-13 September 1998	993,208,259	Not Reported	285,474,117
Excite [WSJS01]	1 December 1999	1,025,910	1,500,500	325,711
BWIE [CVn01]	3-18 May 2000	71,810	116,953	83,232
AlltheWeb.com [JS05]	6 February 2001	451,551	1,350,619	153,297
Excite [SJWS02]	30 April 2001	1,025,910	1,538,120	262,025
AlltheWeb.com [JS05]	28 May 2002	957,303	2,225,141	345,093
AltaVista [JSP05]	8 September 2002	1,073,388	1,073,388	369,350
Tumba! [CS10]	January-December 2003	749,914	1,630,392	254,728
AOL [BJC+07]	1 week December 2003	Several hundred million	Not Reported	Not Reported
Tumba! [CS10]	January-December 2004	338,871	738,576	133,827
AOL [BJC+07]	September 2004 - February 2005	Several billion	Not Reported	Not Reported

AlltheWeb.com and Tumba! search engines complete the European side; from the US side the remaining search engines: Excite and AltaVista. AlltheWeb.com study was considered at the time of the study a major and predominantly European Web search engine [JS05]. With this grouping it is possible to make deductions about the behaviors of users from different regions of the world.

The time range of these studies is from 1998 to 2004, which is a wide range to observe changes in the users' behavior; studies from the same search engine in different time periods make a greater contribution to this aspect. The results presented in each one of the studies, shows a big fluctuation in their values; as a better way of comparison between them, percentages are used. Following the common analysis levels of most studies [JSP05, JS05, JS06, CS10, WSJS01], Table 3.3 summarizes the results gathered. In the preparation phase, researches have to make some decisions in order to prepare the log for the analysis. Transaction logs contain searches from both human users and agents [JSP05]. Attempting to consider only the human users' searches, in most studies is used a cutoff value, defining the maximum number of queries per session that a human searcher may have done. This value is 100 in all of the presented studies (not reported on the Excite's study). Another decision that must also be done is about the session delimitation, defining what time interval should be used to say that two information needs of the same user belong to a different session, trying not to skew the results with ambiguous session times [CS10]. Most of the studies presented do not use this delimitation, they just measure the time from the first submitted query until the user left the search engine; this can possibly lead to abnormal session times: the mean duration on the 2002 study of AltaVista (no. 4) was 58 minutes and 10 seconds, but with a standard deviation of about 3 hours, also the 2001 study of AlltheWeb.com (no. 3) had a mean session duration of 2 hours and 22 minutes but with a standard deviation of almost 5 hours.

Table 3.3: Data collected from Web search engine studies [1].

Study no.	**1 [SMHM99]**	**2 [WSJS01]**	**3 [JS05]**	**4 [JSP05]**	**5 [CS10]**	**6 [CS10]**
	AltaVista	Excite	AlltheWeb.com	AltaVista	Tumba!	Tumba!
Data Collection	August 2 to September 13 ,1998	December 1 ,1999	February 6, 2001	8 September, 2002	January to December ,2003	January to December ,2004
Sessions	285,474,117	325,711	153,297	369,350	254,728	133,827
Queries	993,208,159	1,025,910	451,551	1,073,388	749,914	333,871
Boolean Queries	20.4%	8%	1%	20.0%	12.79%	11.40%
Terms	NR	1,500,500	1,350,619	1,073,388	1,630,392	738,576
Unique	NR	61.6%	13%	9.5%	8.00%	10.33%
Queries Per Session Cutoff	100	NR	100	100	100	100
Session Delimitation	5m	NR	None	None	30m	30m
Mean Terms Per Query	2.35	2.4	2.4	2.92	2.17	2.21
Terms Per Query						
1 Term	25.8%	29.8%	25%	20.4%	39.30%	39.98%
2 Terms	26.0%	33.8%	36%	30.8%	29.00%	26.87%
3+ Terms	27.6%	36.4%	39%	48.5%	31.70%	33.15%
Mean Queries Per Session	2.02	1.9	3.0	2.91	2.94	2.49
Session Length						
1 Query	77.6%	60.4%	53%	47.6%	40.73%	49.52%
2 Queries	13.5%	19.8%	18%	20.4%	22.10%	21.10%
3+ Queries	6.9%	19.8%	29%	32.0%	37.13%	29.38%
Mean Session Duration	NR	NR	2h22min	58m10s	6m31s	5m
Session Duration						
< 5min	NR	NR	26.2%	71.6%	69.07%	74.98%
5-10min	NR	NR	6.2%	6.1%	10.69%	8.91%
> 10min	NR	NR	67.6%	22.3%	20.24%	16.11%
Mean Pages Viewed Per Query	1.39	1.6	2.2	NR	1.45	1.42
Result Pages Viewed						
1 Page	85.2%	42.7%	83%	72.8%	68.11%	76.66%
2 Pages	7.5%	21.2%	10%	13.0%	16.76%	14.38%
3+ Pages	7.3%	36.1%	7%	14.1%	15.13%	8.96%

[1] NR- Not Reported in the analyzed study

At *session level*, the results are balanced across studies. The number of mean queries per session fluctuates between the minimum and maximum values of 1.9 and 3.0 queries (Excite and AlltheWeb.com studies). There is a notable percentage of users that only did one query per session, but no remarkable changes in this values along the years and across studies. The use of 5 minutes cutoff in the 1998 study of AltaVista probably over estimates the number of sessions with only one query (77.6%) [JS06]. The mean session duration has a great variation between studies: the 2001 study of AlltheWeb.com and 2002 of AltaVista show high values (2h22m and 58m10s) unlike Tumba!'s 2003 and 2004 studies (6m31s and 5m). Only Tumba!'s studies used a session duration cutoff value, so it is possible that other studies have some long sessions that do not represent the reality and affect this value, also denoted by their high standard deviation values. The values presented also show that session durations tend to be short. In Tumba!'s studies, around 80% of the sessions lasted less than 10 minutes and only less than 1% lasted over than one hour [CS10]. The 2002 study of AltaVista also showed that 81% of the sessions took less than 15 minutes, and nearly 72% fewer than 5 minutes [JSP05]. The results from AlltheWeb.com could be skewed by long sessions, although 52% of the sessions were less than 15 minutes [JS05]. In the studies presented there seems to be a tendency for short duration sessions.

At *query level*, there are also no significant changes between the studies. The number of mean terms per query is very similar between studies, having the minimum of 2.17 in 2003 study of Tumba! and the maximum value of 2.92 in the 2002 study of AltaVista. These results show that users tend to submit short queries, but there is a tendency for the query's length to slowly increase. The wide time range of the two AltaVista studies show that there was a notable increase in queries with 3 or more terms and a decrease of 1 term queries. The two year study of Tumba!'s search engine does not have so significant changes, showing short increases of 1 term queries and on queries with more than 3 terms. The percentage of users modifying queries increased significantly in the AltaVista search engine's studies from 20.4% in 1998 to 52.4% in 2002 [JSP05]. In the Excite study, this value was 39.6%. In 2003 Tumba!'s study 32.80% of the subsequent queries were modified and 33.48% in 2004 [CS10]. The use of boolean operators in queries (like "+" or "-") varies greatly between studies. Although the use of this operators seems to be fairly low, there are significant changes between the US search engines and the European ones. With the exception of the Excite search engine, US search engines have higher percentage values of boolean queries. In AltaVista's studies, the percentage of boolean queries held stable between 1998 and 2002 at around 20%. European users seem to be less familiar with the use of advanced operators, as shown by Tumba! and AlltheWeb.com's studies. These findings are consistent with other studies [JS05, JS06]. When modifying a query, Portuguese users tend to add a term [CS10], unlike other studies that show that is common to maintain the same number of terms when changing a

query [JSS00, SMHM99, WSJS01].

In terms of *result pages* seen by users, it is clear that the users rarely go beyond the second page. The value of mean pages viewed per query has a minimum of 1.6 and a maximum of 2.2 among the studies. Jansen and Spink [JS06] stated that the tendency is to view fewer pages over time; this affirmation is consistent with the results showed. With the exception of the Excite study, every other study shows a percentage of users seeing only the first page greater than 65%. Excite users seem to be more persistent, with an abnormal percentage of 36.1% seeing 3 or more result pages; although other Excite studies (1998 and 2001) show that there was also a tendency for users to view fewer page results [JS06]. Again in the session cutoff in the 1998 study of AltaVista could skew the results and increase the percentage of users seeing only one page. US users seem to see fewer result pages than the European ones.

In *term analysis* and *topic classification* there are some differences between studies. Excite's study shows an abnormal percentage of unique terms, 61.6%, unlike the other studies where the maximum percentage verified was 12.79% in the 2003 study of the Tumba! search engine. This value is unusual, but normal for the Excite search engine, as Spink et al. [SJWS02] showed in a comparison of Excite logs from different time periods; the logs contained large amounts of terms that either are never repeated or used with low frequency like personal names, spelling errors, non-English terms and Web-specific terms such as URLs. In Tumba!'s study, Costa and Silva stated that caching only 1% of the most frequent terms it would be possible to handle 50% of the queries [CS10]. To evaluate the main topics of interest, in most studies a random sample of queries is selected for analysis; queries are then classified under eleven general categories defined by Spink et al. [SJWS02]. Selecting three studies, Table 3.4 shows the five most frequent topic categories for each of them. The topics *Commerce, Travel, Employment or Economy* and *People, Places or Things* occupy the first two places in later studies. These findings match other studies stating that there is a decrease in topics like *Computers, Internet or Technology Items* and *Sex or Pornography* opposing to an increase in topics like *People, Places or Things* and *Commerce, Travel, Employment or Economy* that account for about 50% of the queries [JS05, SJWS02, JS06]. In a study comparing data collected from Excite's search engine from 1997, 1999 and 2001, Spink et al. detected a shift in search topics. Categories like *Entertainment or Recreation* and *Health or Sciences* moved down the ranking. *Commerce, Travel, Employment or Economy* and *People, Places or Things* moved up. In 1997, about one in six queries was about sex; by 2001 this was down to one in twelve [SJWS02].

There are other studies in this field with interesting results and conclusions that are worth mentioning. Pass et al. [PCT06], presented a paper based on pictures, as shown in Figure 3.1, showing information collected from AOL's search engine, and giving information about query space, user sessions, user behavior, operational requirements, content

Table 3.4: Ranked Topic Classification.

Excite 1997 (2,414 queries)	**% Queries**
Entertainment or Recreation	19.9%
Sex, Pornography or Preferences	16.8%
Commerce, Travel, Employment or Economy	13.3%
Computers or Internet or Technology items	12.5%
Health or Sciences	9.5%
AltaVista 2002 (2,603 queries)	**% Queries**
People, Places or Things	49.27%
Commerce, Travel, Employment or Economy	12.52%
Computers or Internet or Technology items	12.40%
Health or Sciences	7.49%
Education or Humanities	5.07%
Tumba! 2004 (1,000 queries)	**% Queries**
Commerce, Travel, Employment or Economy	20.30%
People, Places or Things	17.70%
Health or Sciences	11.80%
Education or Humanities	10.50%
Society, Culture, Ethnicity or Religion	6.10%

space and user demographics. The query topics are diverse with queries about entertainment, shopping and porn occupying the first three places. Queries about personal finance are mostly done between 8h00 and 12h00; queries about music are mostly done between 1h00 and 4h00. In the universe of all queries, 28% are reformulations of a previous query.

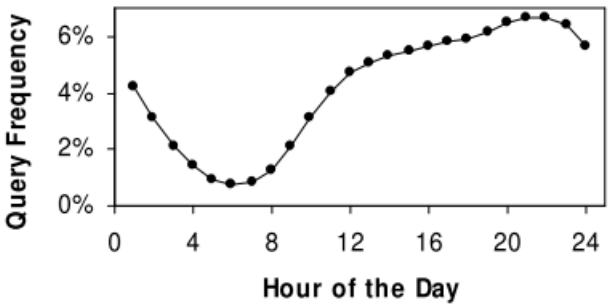

Figure 3.1: Query Frequency along the day [PCT06].

Beitzel et al. [BJC+07] analyzed two query logs from America Online's (AOL) search engine. While the first contained all queries from an entire week (several hundred million) the second one contained the entire query stream from AOL's Web search service over a continuous 6-month period from September 2004 through February 2005 (several billion queries). In this analysis, the authors found certain trends that are stable over time despite a continuous fluctuation in query volume. Certain topical categories can exhibit both short-term (over hours in a day) and long-term (over several weeks or months) query

trends, and these trends and their behavior may vary wildly depending on the category and the length of time being studied. For instance, queries about a popular actor can have a stable query frequency along the time and have a peek at a certain date, if there is some action that triggers his popularity.

In an interesting study [HS00], Hölscher and Strube, showed that there are significant differences between expert and novice users when using a search engine. Performing interviews was possible to distinguish participants between novice users and expert users. Novices tend to reformulate more often their queries and they often make small and ineffective changes to their queries. Domain knowledge is also very important as searchers with background knowledge about the domain spend less time reading information about it and are more aware of what to search next. Experts tend to use longer queries, use more often boolean operators, and commit less formatting errors on their queries .

Teevan et al. [TRM11] compared Twitter searches with Web search engines. People search Twitter to find temporally relevant information (e.g. breaking news, real-time content and popular trends) and information related to people. Microblogging content has very different properties than content on the Web. Tweets are short, frequent, and do not change after being posted. Twitter search is used to monitor content, while Web search is used to develop and learn about a topic. Twitter search includes more social content and events information, while Web results contain more basic facts and navigational content. Search engines could use social information finding behavior to improve search experience. An hashtag method as the one used in Twitter is suggested to be adapted to Web search results. Using content analyses of the tweets that match a query might help to disambiguate the most common query intents; pages that match popular query specific Twitter topics, could be ranked higher.

Chapter 4

Methodology

4.1 Overview

At the beginning of this project the main metrics and analysis dimensions were already defined. Nevertheless, some metrics used in other works could not be used because of the lack of information in the dataset (e.g. no information about query results' clicks). Frequently in exploratory works like this researchers discover new details about the stored information and consequently new forms of data analysis. This fact was important to decide how to do the analysis. The two possibilities considered were: use a database/data Warehouse and conduct the analysis from there or use ad hoc data analysis from the UNIX command line and a scripting language. The second option was used. The preparation work for conducting the analysis from a database/data warehouse would be much bigger. Using the UNIX command line and a scripting language there wouldn't be this initial overhead and data inconsistencies would be much more easier to solve.

The dataset contained was stored in 165 compressed text files. These files have really good compression ratios which allow them to have relatively small sizes; extracting these files would consume large disk space. To avoid doing this, using the *zcat* command the files' content is written on standard output and available for processing. Another UNIX facility very used was pipes. Pipes allow to redirect the output of a program to another one or to a file, which is perfect to chain a set of processes. As in other data mining projects, most of the work is based on regular expressions i.e., matching strings of text, allowing to retrieve certain parts of the dataset relevant for the analysis; from the earlier stages to the end of the project regular expressions were used on a daily basis. Perl was the scripting language chosen mainly for its efficiency and flexibility when dealing with regular expressions. The main UNIX commands used were:

- **awk** - Used for simple data manipulation (e.g. switch the order of two columns) and calculations;
- **grep** - Used to match a given regular expression;

- **sort** - Used to sort certain fields (e.g. sort by IP address);
- **uniq** - Used to retrieve only unique results;
- **cut** - Used to retrieve specific parts of a line bounded by a delimiter.

Along the analysis process dozens of commands were executed more than one time. Since inserting all these commands manually was impracticable, these were distributed across some bash scripts that will run them and store their output whenever needed. Another programming language used was *R*[1], that offers many features for statistical computing and graphics. To draw graphics it was used an R package, *ggplot2*[2].

The analysis focused on three main levels: sessions, queries and terms. These were also the main levels of analysis used on other studies, so a direct comparison of the results is possible. The most common metrics used in similar studies (explained at section 2.4.3) were used at each one of these levels. Unfortunately the dataset doesn't have information about query results' clicks therefore an analysis at this level was not possible.

4.2 SAPO's Dataset

The dataset provided has information recorded about queries made on SAPO's search engine, covering a period of about six months in the year of 2010 (January, 29 to July, 12). This information was stored in 165 compressed files with a total size of 7.6 Gigabytes. In each one of these files the information about queries is stored in a kind of markup structure (like XML). Each record (query) is bounded by an empty line, so that it is possible to distinguish between different records. Not all these fields are important to the analysis process so such fields can be discarded. The fields used in the analysis process and their explanation are the following:

- *ip* stores the IP address of the user who made the query; the fields *country* and *city* are, respectively, the country and city registered for this IP address;
- *date* is a field with 14 numerical digits representing the date when the query was done in the format "YearMonthDayHourMinutesSeconds";
- *browser* is the user-agent string that identifies the application used by the searcher to interact with the search engine;
- *keywords* shows the query terms entered by the searcher.

Although the same structure is used for all queries, most tags don't always have content, and they are represented by empty tags (e.g. <browser \>). It is imperative to know

[1] Available at http://www.r-project.org/
[2] Available at http://had.co.nz/ggplot2/

which fields always have content (especially those needed in the analysis process) and to avoid problems in future stages if all are always present, with or without content. From the verifications done it is possible to conclude that all these fields are always stored. Only two fields always have content, *date* and *ip*; the remaining fields when there is no content the empty-tag is always present.

The original dataset has about 140 million queries. As the next chapters will show the final number of queries will be much more diminished because the original dataset has a large number of non-human queries (bots).

4.3 Dataset Preparation

When doing analysis of textual content, especially when using regular expressions, we must be certain about how the information is organized. When dealing with regular expressions one must be sure about the text pattern so that the regular expression can match it for all situations. For this reason an early task was to check if the information stored obeyed to the specified structure and find possible cases that could bring problems in the future. This was an ungrateful task as much of the problems were not found at this initial stage but over the course of the project, leading to redo and rethink a lot of work.

One of the problems found at this early stage was related to the last stored query in each one of the 165 files. For unknown reasons (related with the dataset collection process), in all of them the information about the last query was always incomplete. As all the files were read sequentially, and because these last queries stored incorrectly don't have a line break to signal the end of the line, their information would be connected to the first query of the next file. As this could introduce incorrect results these last queries were discarded; one piece of text bounded by an empty line, was only accepted as a query if it was found the opening and closing of the notification tag. Sometimes the field that stores the query terms, *keywords*, had the line break character leading to the occurrence of empty lines inside a query record. This event was solved by defining that when an opening of the *keywords* tag is found, that specific line must end with the closure of that tag; if this was not true the line-break of that line was removed. All leading and trailing white-spaces in every line were also removed.

These procedures were applied on the bots' elimination phase, explained next, where a new file with a structure identical to the original dataset is created eliminating all these problems.

4.3.1 Bots Elimination

The original dataset was not object of any kind of cleaning, meaning that it was possible to encounter many queries made by software applications (bots). As this study is only

concerned in analyzing queries made by human users, queries made by bots should be part of this analysis. Remove all the queries made by bots with 100% certainty, is a difficult task that surpasses the main goals of this work. Using the field *browser* and matching its content with some known user-agent strings of bots, would be possible to remove many unwanted queries. When this field has no content (empty tag) it would be impossible to tell if it is a bot, so these queries remain in the dataset.

For this method it was obvious that some kind of user-agents' list had to be done or use an existing one. Two different approaches were considered to produce this user-agents' list: having a list of all the existing browsers and remove all the queries whose user-agent string is not there, or start with a list containing all the bots (crawlers, robots, spiders,etc) and remove queries whose user-agent string is in that list. To start the word "all" could not be used in any of these two approaches. There are so many browsers/bots, with different versions, names, OS (Operative System) specific, and other specificities that would not be possible to find and store all of them. For these reasons the list could also not contain the exact user-agent string for every browser/bot version and try to do an exact match with the *browser* field content; this matching process needed to be regular expressions' based. Taking all these factors into account the second approach was chosen, because it was much more probable to wrongly discard queries using the first approach; also it is more desirable that some queries made by bots remain in the dataset than to wrongly remove non-bot queries.

To construct this list were used two online sources [Use, And] that include a list of the most known user-agents' strings. Browsing trough these two sources, a list of words that are present on user-agents' strings was created. This list (check Table A.1 in Appendix A) does not contain the exact string of the bot, but only a word to be used as a regular expression and match the *browser's* field content. For example the bot's string "Blaiz-Bee/1.0 (+http://www.blaiz.net)" could be stored in the list as only "Blaiz-Bee" because there is no need to do a more complicated regular expression (to match the version number). Another strategy was to store common words on the list so that its length was lower. Words like "bot", "crawler", "feed" or "spider" could match a large number of bot's strings; also this way it was not necessary to store strings like "Arikus Spider", "Googlebot/2.1", "grub crawler" because those common words would already match these strings. Extra care was taken to not include any word that could match a browser user-agent string. The final result is a new file with the same structure of the original dataset without the queries made by bots removed.

Some verifications were done to check if the right content was removed or if it was possible to discover new bot's strings to remove. Two files were created with the list of the user-agents' strings removed and the remaining. These files were sorted by decreasing order of occurrences, and in each one of them the strings with larger occurrences were checked. With this verification was possible to add new bots' strings to be removed and

check if any of the already removed were, indeed, bots. The original dataset had about 140 million queries (140,112,498). After remove the bots the number of queries was about 102 million queries (101,795,370). This number is quite impressive because around 30% of the queries (38,317,128) stored on the dataset were considered as being made by bots and removed.

4.3.2 Sessions Creation

After cleaning the original dataset from bots, the obvious next step would be the creation of sessions. The original structure of the information stored was very "noisy" especially because the information was spread along many lines and not all of it was really important to analyze. Furthermore, with this original structure it was not trivial to extract specific parts of the dataset. For this reason, all the fields needed in the analysis phase were stored in a file with a different structure. In this file each field is separated by a tab character, separating the fields by columns making easy the access of each field using the UNIX *cut* command. The fields stored in this new file were (by this order): *date*, *ip*, *country*, *browser*, *keywords* and *city*. Figure 4.1 shows some lines of the created file. When one of these fields had no content it was stored in the file the word “EMPTY_FIELD” (e.g. “EMPTY_KEYWORDS” or “EMPTY_COUNTRY”). There were three distinct types of empty queries relative to the *keywords* field: when there was an empty tag for the field *keywords* and it was stored in the new dataset as “EMPTY_KEYWORDS”; when the field *keywords* had no content (e.g one or more white-spaces); the third one was the *empty taxonomy*. Sometimes along with the query terms the tag *taxonomy* was found. Frequently only the taxonomy information was stored and no query terms (e.g. “taxonomy:TOP/DESPORTO”); these queries were considered as empty as they don’t contain any content. To simplify the sessions creation this file was sorted first by the *ip* field and then by the *date* field. Having the file sorted by these two fields would be much less complicated to create the sessions’ file.

```
20100604121053  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121053  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121112  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   o faz tudo      Vila Real
20100604121112  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   o faz tudo      Vila Real
20100604121113  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121113  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121124  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   o faz tudo      Vila Real
20100604121124  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   o faz tudo      Vila Real
20100604121125  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121125  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121611  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   o faz tudo      Vila Real
20100604121611  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   o faz tudo      Vila Real
20100604121612  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
20100604121612  <ip>  PT    Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0)   EMPTY_KEYWORDS  Vila Real
```

Figure 4.1: The new dataset structure.

Having this file the next step was to create the sessions. The sessions were delimited using the IP address and date of the query. An interval was defined to separate different

information needs of the same user. Two consecutive queries are included on different sessions if they have an inactivity of at least 30 minutes between them. This gap value was used in other studies [CS10, CVn01] and using one avoids having sessions with very long durations which would not represent the reality. Very often the exactly same query (same date, IP address, query terms, etc) was stored in the dataset more than one time; these queries were not included in the sessions' file. As in other studies [SMHM99, JS05, JSP05, CS10] a cutoff value was used to delimit the maximum number of queries per session. Any sessions with more than 100 queries were excluded, since sessions with so many queries were likely to come from bots. All the three types of empty queries were also excluded.

Using the dataset sorted by the IP address and date, the creation of sessions was simple. As the queries with the same IP were all together and sorted by date at each step is only necessary to subtract the date of the current query with the last one and check if the difference is less than 30 minutes and the IP address is the same. If these two conditions are true then the session is the same otherwise the query belongs to a new session. When there is a new session the last one is only accepted if the number of queries is not greater than 100.

875	2010-06-07 15:02:27	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	FACEBOOK	Mirandela
875	2010-06-07 15:32:15	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	FACEBOOK	Mirandela
876	2010-06-07 16:29:16	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	bershka	Mirandela
876	2010-06-07 16:35:25	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	zara	Mirandela
877	2010-06-07 17:41:51	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	Facebook	Mirandela
877	2010-06-07 17:43:01	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	zara	Mirandela
877	2010-06-07 17:52:56	<ip>	PT	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)	bershka	Mirandela

Figure 4.2: Some lines of the sessions' file.

To the last dataset structure (see Figure 4.1) is added a new column that represents the session identifier. The date is also stored in a human-readable format (Year-Month-Day Hours:Minutes:Seconds). Figure 4.2 shows some sessions of the file. At the end of the session's creation the total amount of queries dropped considerably. The total number of queries was now 45,413,607, a drop of almost 60% from the total number of queries (101,795,370), with a total number of sessions of 15,767,954. This was mainly because of the removal of sessions with more than 100 queries. There were 14,272 sessions in these situation totalizing 30,474,741 queries. The number of empty queries removed was 14,725,859 (where 1,880,616 were "empty taxonomy") and repeated queries were 11,181,163. A total of 56,381,763 queries were removed at this stage. Table 4.1 summarizes these results.

Table 4.1: Queries Removed in Sessions' Creation.

	Queries	**% (from the last dataset)**
Empty Queries	14,725,859	14.47%
Empty Tag	12,834,738	12.61%
Empty Taxonomy	1,880,618	1.85%
Empty Content	10,503	0.01%
Repeated Records	11,181,163	10.98%
+100 Sessions	30,474,741	29.94%
Total	**56,381,763**	**55.39%**

Methodology

Chapter 5

Data Analysis

5.1 Overview

Table 5.1: General Statistics.

Metric	Value
Queries	45,413,607
Sessions	15,767,954
Terms	89,609,923
Mean Queries per Session	2.88
Mean Terms per Query	2.03
Mean Characters per Term	6.86
Unique Queries	19.59%
Unique Terms	3.03%
Queries Never Repeated	11.78%
Terms Never Repeated	1.5%

The main results of this analysis are shown on Table 5.1. The queries distribution has a defined pattern along the week and hours, as shown in Figure 5.1, Figure 5.2 and Figure 5.3. The mean number of queries per day is 275,234 (median of 290,505). The maximum number of queries in a day was registered on 2010-02-08, with 375,558 queries; the minimum was on 2010-06-02 with only 11,346. This last value could be due to an error in the collection process. Except from the first and last months the number of queries per month is similar. These two months should not be compared with the others because the queries are not logged for a full-month period. The number of queries on the other months is between 8 and 9 million queries. The only exception is the month of June where the total amount of queries drops to about 7 million.

The weekly pattern is clearly seen in Figure 5.2. Along the week there is a clear downward tendency. From Monday to Wednesday the number of queries is almost always above the average. From Thursday the the tendency of decrease is noticeable. The weekend period is when less queries are made. On all Saturdays the number of queries is

below average and the behavior on Sundays is almost the same. In a study from Vivisimo [KSJ06], a meta-search engine, the percentage of queries per weekday also clearly drops on weekends from values between 14% and 17% to 10% and 11% on Saturday and Sunday, respectively. In AOL's study [BJC+07] the results were a little different. In this study the highest percentage of queries in a weekday is registered on Sundays and from here there is a downward tendency until Friday where is seen the lowest number of queries in a weekday. Unlike this and Vivisimo studies, in AOL's study from Friday until Sunday the percentage of queries per weekday raises.

In this period there are two drops on midday and 7:00 p.m.; the first one could be related with the lunch break. The highest traffic is verified between 2:00 p.m. and 5:00 p.m., where the number of hourly queries is more than 50% above the hourly mean. This amount of traffic is only again seen on 9:00 p.m. . The hour with the lowest number of queries is 1:00 a.m. and the highest 3:00 p.m.. These results were a bit different from an AOL study [BJC+07] with the same kind of analysis. In this study the day with less queries was Friday and the maximum peek was on Sundays. The downward tendency from Monday until Friday was also there, but the number of queries raises steadily on weekends. In this study the hour with less queries was 6:00 a.m. and from this hour until almost the end of the day the number of queries raises almost continuously. The hour with most queries was between 9:00 p.m. and 10 p.m.

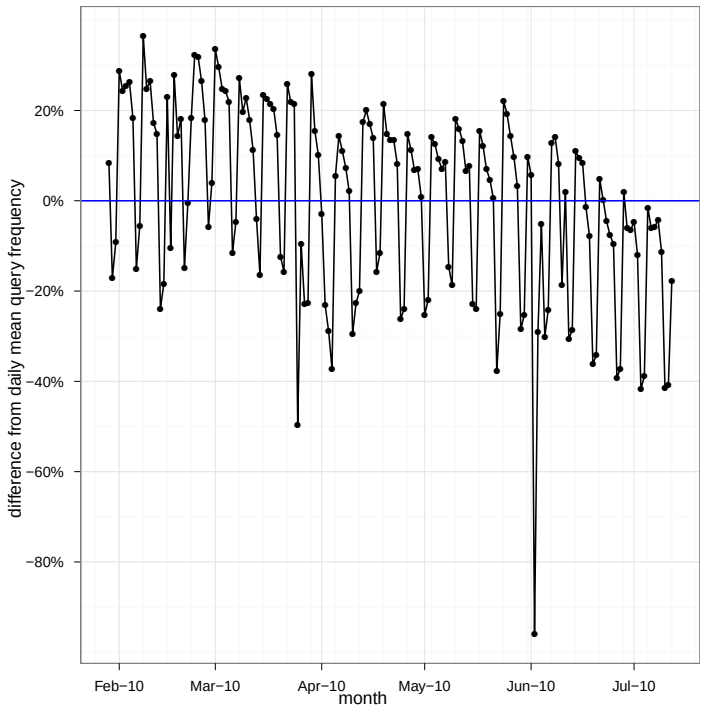

Figure 5.1: Difference from Daily Mean Query Frequency.

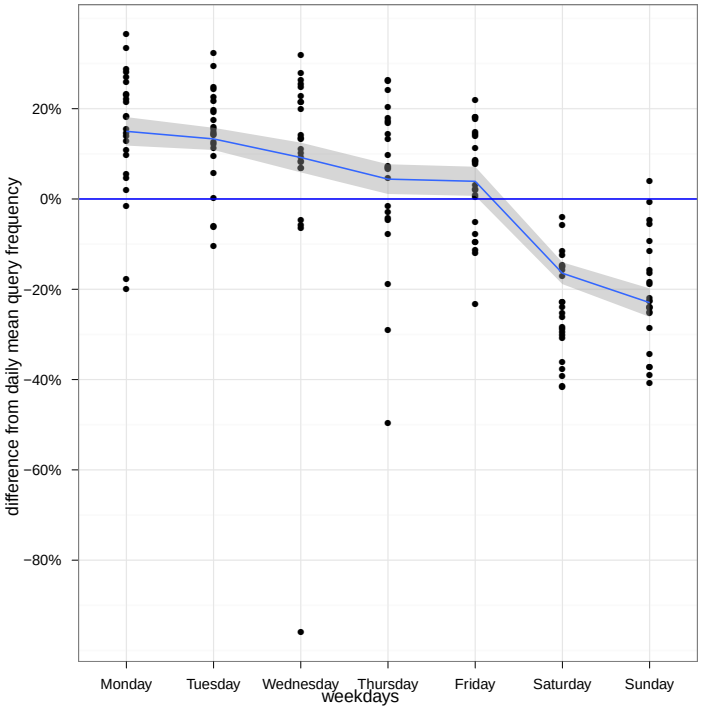

Figure 5.2: Difference from Weekday Mean Query Frequency.

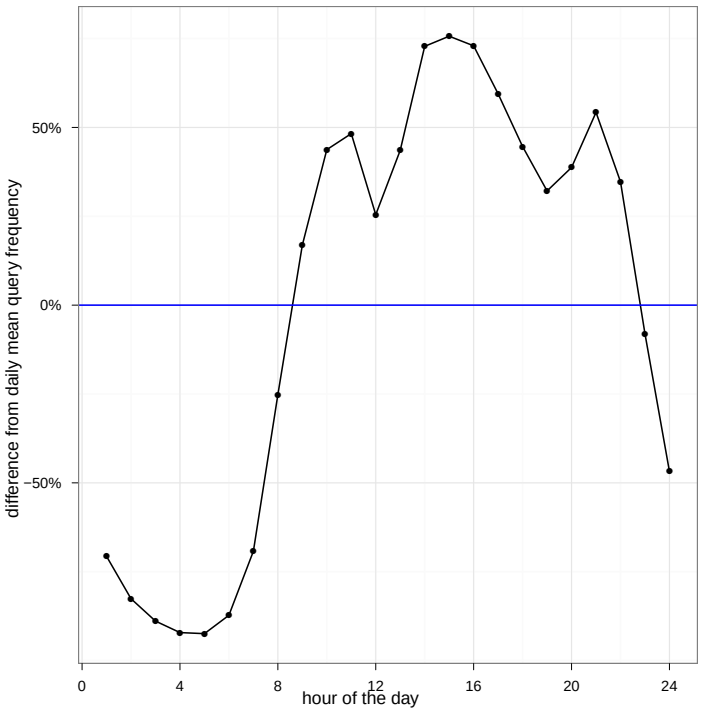

Figure 5.3: Difference from Hourly Mean Query Frequency.

Using the *browser* field was possible to make some analysis about the browsers that users choose to do their searches on. This analysis was only possible when the field had this information; this happened for around 99% of the queries (44,992,247). The analysis was done for the 5 main browsers in the year of 2010 [W3S]: Internet Explorer, Firefox, Chrome, Safari and Opera. Table 5.2 shows the distribution of these 5 browsers and Figure 5.4 the ten most used browser versions. SAPO's users do not follow the general statistics that give Firefox the first place in this ranking [W3S]. The browser usage along the months was similar for every browser except Internet Explorer. Internet Explorer registered a downward tendency which was also verified in browsers' usage statistics along the year of 2010 [W3S].

Table 5.2: Most Used Browsers.

Browser	**Queries**	**% Queries**
Internet Explorer	37,848,289	84.12%
Firefox	4,749,292	10.56%
Chrome	1,262,578	2.81%
Safari	559,829	1.24%
Other Browsers	444,905	0.99%
Opera	127,354	0.28%
Total	44,992,247	100%

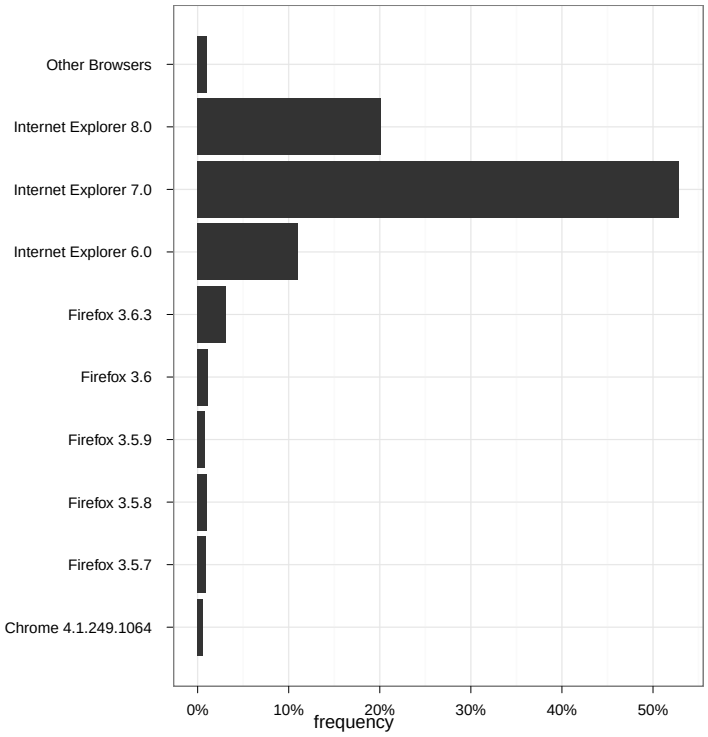

Figure 5.4: Most used Browsers and Versions.

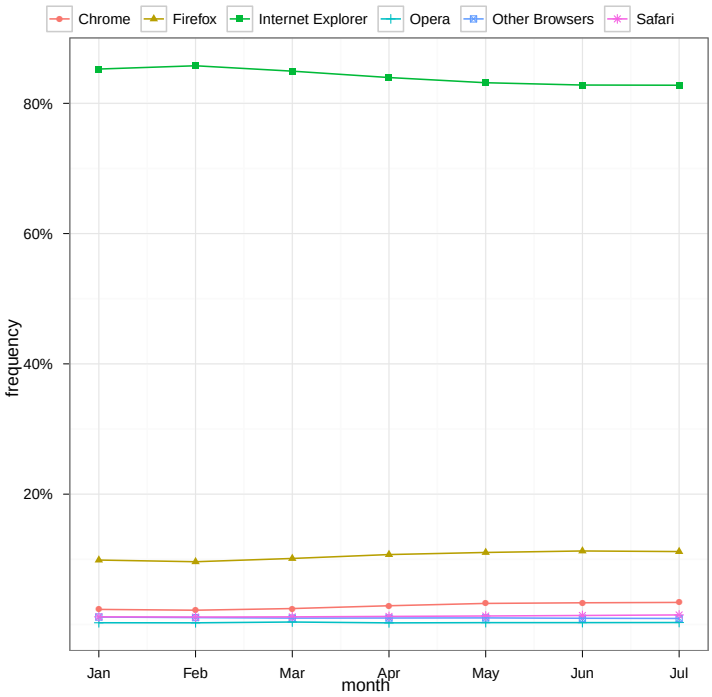

Figure 5.5: Cumulative Browser Usage Evolution.

Table 5.3: Top 10 Countries.

Rank	Country	Queries	%
1	Portugal	36,806,631	90.08%
2	Brazil	722,661	1.77%
3	France	696,019	1.70%
4	Switzerland	377,214	0.92%
5	United Kingdom	326,568	0.80%
6	United States	313,450	0.77%
7	Spain	236,945	0.58%
8	Germany	228,282	0.56%
9	Canada	137,558	0.34%
10	Netherlands	116,179	0.28%
	Total	39,961,507	97.8%

The main goal of this study is to characterize Portuguese web searches. Using the information from the *country* field was possible to check the searcher's country. Although this information could be misleading and not precise (e.g. the use of a proxy) this was the easiest way to check whether a query was made in Portugal or not. When the *country* field had no information stored (empty tag) this verification could not be made; this happened for around 10% of the queries (4,549,826). Table 5.3 shows the ten countries with most queries, for all the queries that had this information (40,863,781 queries, around 90% of

the total). Only less than 10% of these queries were not made by users with a Portuguese associated IP address. Also around 95% of these queries were made from European countries. These values strongly indicate that almost all the queries were made by European users, mostly Portuguese.

A similar analysis was made for the *city* field. This particular analysis should be deal with extra care. These cities represent the Internet Service Provider city and could not represent accurately the user's city. In 12% of the queries this field had no content. Not surprisingly the 10 cities with most queries were all Portuguese. To see a city from another country we would have to go to place 60th (São Paulo from Brazil). These cities account only around 34% of all queries and only represent 0.05% of all the cities registered on the dataset, as there are a total number of 17,868 unique cities. Furthermore, these ranking is not paired with the number of the 10 Portuguese cities with most population (from 2006 census [Ins07]). Only 4 of them (Lisboa, Porto, Amadora and Guimarães) are present in both rankings and appear by the same order in both.

Table 5.4: Top 10 Cities.

Rank	**City**	**Queries**	**%**
1	Lisboa	8,308,065	18.29%
2	Porto	2,203,595	4.85%
3	Carnaxide	951,029	2.09%
4	Sintra	710,530	1.56%
5	Coimbra	694,234	1.53%
6	Amadora	626,401	1.38%
7	Almada	541,167	1.19%
8	Maia	538,107	1.18%
9	Loures	511,406	1.13%
10	Guimarães	481,391	1.06%
	Total	15,565,925	34.36%

5.2 Bots' Analysis

In many works on this area few references are made to the analysis of bots mainly because researches have access to datasets already without bots. Such analysis could give us an insight about bots' behavior on search engines. The bots' removal process was done first by checking the user-agent string and comparing it with some known bots' strings. At the stage of sessions' creation, longer sessions (more than 100 queries) were removed as they were considered as queries made by bots. During these processes many queries were removed: at the first step 38,317,128 and over sessions' creation 30,474,741 which adds a total of 68,791,869 removed queries. This value represents about half of the queries of the original dataset. A query traffic so high as this one could give a better knowledge about bots if some analysis is done.

Observing Figure 5.6 and Figure 5.7 it is possible to draw some conclusions. The number of queries made by bots seems to be consistent over nearly all months. In June was registered the maximum number of bots' queries, on the 4th day of the month. Surprisingly the minimum was also registered on the same month on the 2nd day. This month accounts for almost 25% of the total queries made by bots. The graphic clearly shows an increasing in the number of bots from the month of April until June. In March and April this percentage was around 15% and raised to 20% in May. In the month of June the number of human queries decreased about 15% (from 8,498,272 to 7,227,360). The daily mean of non-human queries is about 416,920 (median of 379,953) and from human queries this value is 275,234 (median of 290,505). These values are proven by Figure 5.6 where we can see that in almost every day the percentage of non-human queries is superior. Figure 5.7 shows the difference between the frequency of bots removed in the preparation phase and bots removed at the sessions' creation. The two types are balanced with a slight advantage of the bots removed on the first stage, in almost every month. Again the month of June was the exception. In this month the number of sessions bots was clearly lower than the other bots' type.

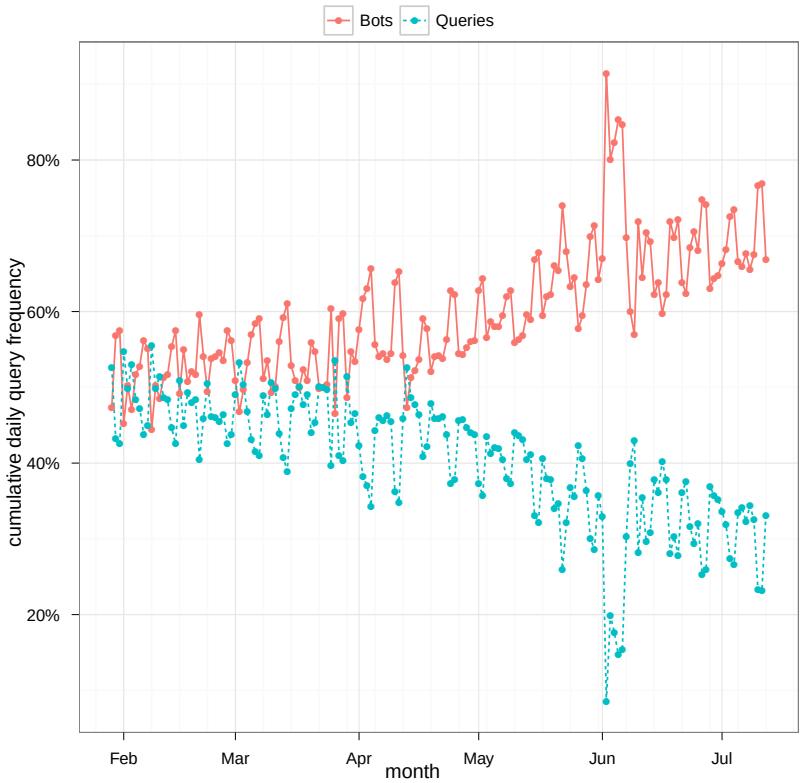

Figure 5.6: Cumulative Daily Bots/Queries Frequency.

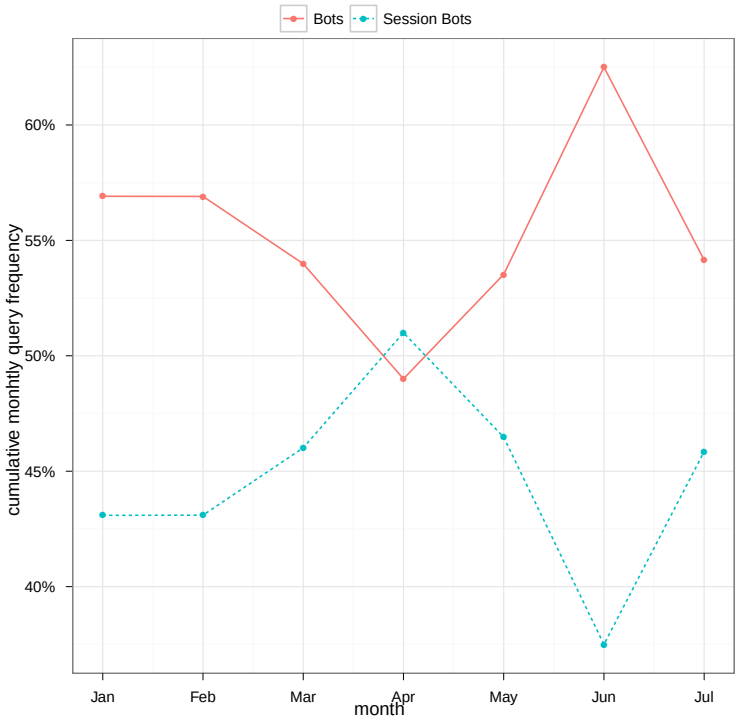

Figure 5.7: Cumulative Monthly Bots Frequency.

From these values it is possible to conclude that the abnormal high value of bots in this month was mainly related with bots removed at the first stage. It is important to say that this situation wasn't because of an decrease of sessions' bots but because of a strange increase of the other kind of bots. From May to June there was an increase of these kind of bots of almost 42%. Regarding the sessions' bots, another important matter was to find how they were distributed; the distribution of sessions with more than 100 queries is shown in Figure 5.8. The graphic shows a clear downward tendency with some upsides along the way. The most seen number of queries per session was indeed the first one to be removed, 101, which is found a total of 303 times; although this is only 2% of the total sessions removed (14,272). The maximum value of queries per session is incredibly big number, 1,138,241; the mean is about 2,135 queries, with a median of 167. This mean value seems to be affected by some very high queries per session values that occur few time, so the median value seems to be much more accurate and close to the reality. The interval of queries per session from 101 until 1500 accounts for 95% of all the sessions removed, thus confirms the last statement. Table 5.5 shows the 10 most frequent automatic processes and Table 5.6 their most frequent queries. Although SAPO search engine uses the Robots Exclusion Protocol this seems insufficient as many known automatic processes are still seen. Surprisingly the most frequent query seems made by an human user. As expected the empty query appears on the 4th place. The other queries'

content have no logical meaning for a human user. These queries account around 15% of the total amount of queries made by bots.

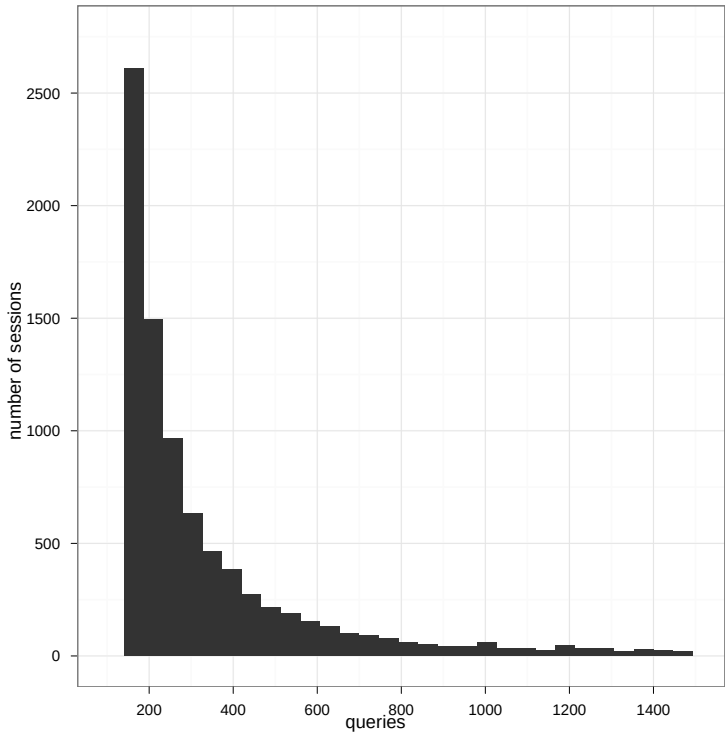

Figure 5.8: Distribution of Sessions with More than 100 Queries.

Table 5.5: The 10 Most Frequent Automatic Processes.

Rank	Description	Occurences	% Total Bots
1	Radian6 RSS Feed Crawler	12,831,927	18.65%
2	Google Robot	7,168,775	10.51%
3	GigaMedia/NTT DoCoMo Robot	6,768,236	9.38%
4	SimplePie RSS Parser	3,668,146	5.33%
5	Magpie RSS Parser	1,664,385	2.42%
6	Twiceler Web Crawler	719,626	1.05%
7	SAPO::HTTP	712,935	1.04%
8	SAPO RSSWorks	629,188	0.91%
9	Baidu Spidering Engine	246,946	0.36%
10	libwww Perl Module	213,942	0.31%

Table 5.6: The 10 Most Frequent Bots' Queries.

Rank	Query	Occurences	% Total Bots
1	portugal	3,861,500	5.61%
2	sevices	2,836,660	4.12%
3	keyword	1,317,448	1.92%
4	EMPTY_KEYWORDS	611,645	0.89%
5	style.css	269,328	0.39%
6	search	157,038	0.23%
7	adigms	153,112	0.22%
8	nossasfotos00851com	123,055	0.18%
9	wp contentthemeslifestyle_10imagesbg_small_top.gif	62,268	0.09%
10	"component,myblog"	50,839	0.07%

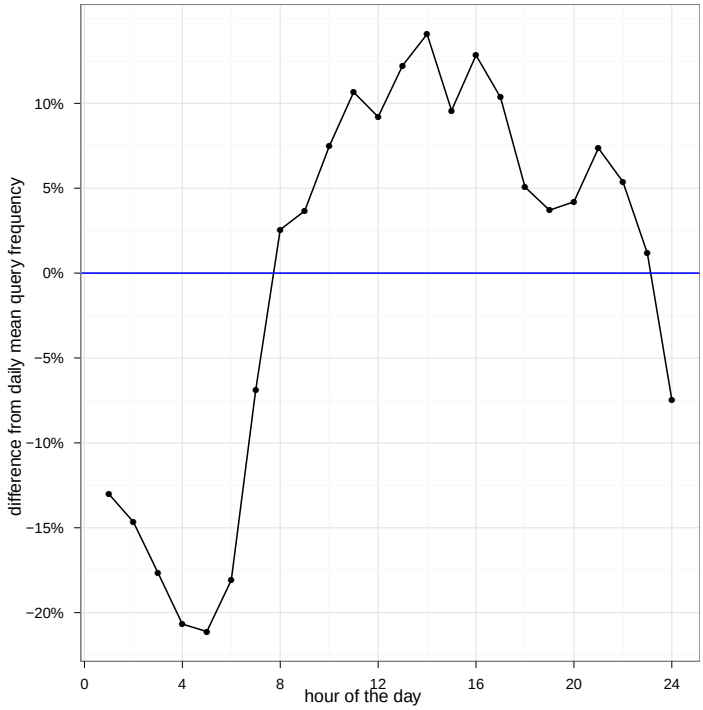

Figure 5.9: Difference from Hourly Mean Bots Query Frequency.

The distribution of queries made by bots along the 24 hours of the day is shown on Figure 5.9. The first seven hours of the day are below the daily mean. From 5:00 a.m. the number of queries made by bots increases steadily until 11:00 a.m.. From 8:00 a.m. until 11:00 p.m. the amount of queries is above the daily mean. The maximum value is registered at 14:00 p.m. and the minimum at 1:00 p.m. (accounting almost more 15% above and below the daily mean, respectively). From 11:00 a.m till 17:00 p.m the number of bots' queries is around 10% above the daily mean. This interval could be considered as the bots' "rush hour". Oddly this hourly distribution is very similar for queries made by

human users. Figure 5.10 shows the distribution of bots' queries by the day of the week. The variation in the number of bots per weekday is reduced, and much more constant than for queries made by human users. Viewing this graphic it is possible to see that from Monday to Wednesday there is a tendency to the number of the bots' queries to decrease and then rising until Friday when it starts to descend again until Sunday. From this graphic it is possible to see the weekday of the maximum and minimum values of bots' queries.

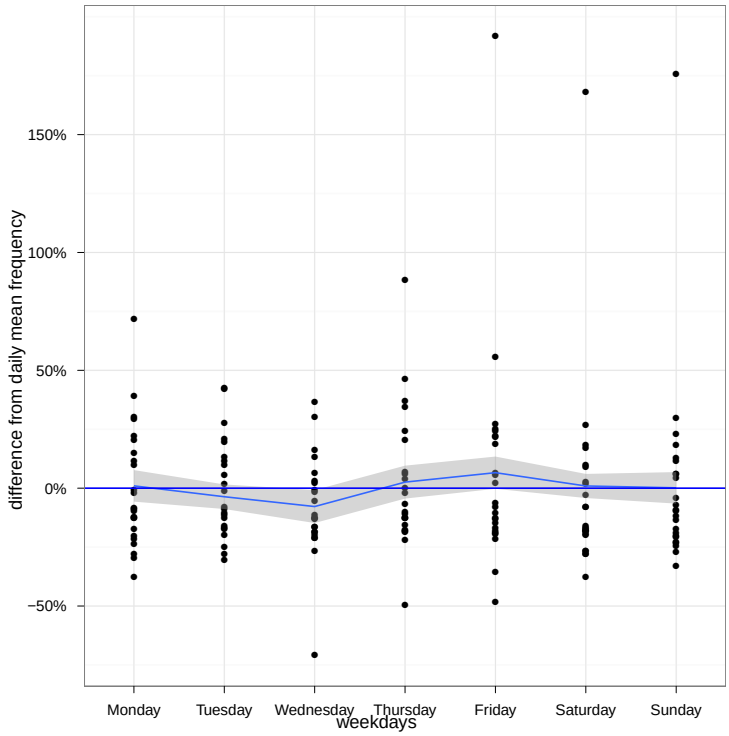

Figure 5.10: Difference from Weekday Mean Bots Query Frequency.

5.3 Session Level Analysis

Analyzing the sessions could be quite useful to infer the searchers' behavior during a search episode. The session duration and its length are the main metrics used at this level. The total number of sessions is 15,767,954. Figure 5.11 shows that there is no visible big discrepancy between number of sessions and queries. Clearly when one of them raises the other shows the same behavior and the same happens on descents.

The session duration is measured from the first query in that session until the last one. This value ignores the time users spend viewing web pages or doing other tasks. Table 5.7 summarizes these results. Sessions with a duration greater or equal than 0 and less than 1 are those with the greatest percentage. Sessions with a single query are considered as having a duration of 0 minutes; the large number of sessions with only one query (check

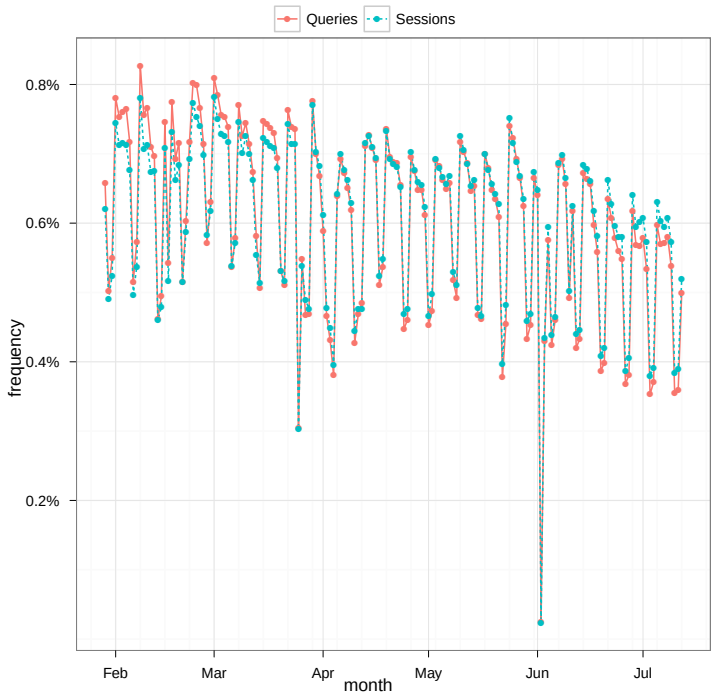

Figure 5.11: Sessions and Queries Distribution along months.

Table 5.8) falls in this interval and, naturally, increases this value. Around 65% of the sessions have a duration lower than 1 minute, and almost 88% of sessions have a duration lower than 15 minutes. Sessions with duration of 1 hour or more are very rare (only 1%). From these values it is possible to conclude that search's sessions end quickly. The mean session duration was 5 minutes and 23 seconds. There is a clear downward tendency as the duration increases, only interrupted by sessions with duration between 15 minutes and less than 30. This event also happened in Tumba!'s study [CS10] where the percentage of sessions within this interval raised (from 5.88% to 8.89% and 4.80% to 7.29% in the years 2003 and 2004 respectively). The exactly same behavior happened on Vivisimo search engine [KSJ06] registering raises on this interval of around 3% from the last interval (10 minutes and less than 15 minutes). AlltheWeb's study [JS05] shows an irregular behavior along the session duration time very different from the other studies. In this study the percentage has ups and downs along these intervals (from 26.2% for sessions with duration less than 5 minutes lowering 6.2% in the interval from 5 to 10 minutes and increasing again in the interval from 10 to 15 minutes, and this pattern repeats). The mean session duration seems consistent with Tumba!'s study [CS10] with 6 minutes and 31 seconds and 5 minutes for the years of 2003 and 2004. AlltheWeb and Altavista [JS05, JSP05] sessions' duration are much longer (2 hours and 58 minutes respectively). This could be due to the nonexistence of an inactivity period (session delimitation) between sessions which

could incorrectly evaluate the session duration. This study registered one of the lowest session durations amongst the observed studies. The maximum duration value registered was 2595 minutes (about 43 hours). In this session the query had always the same term ("sex") on the 97 queries, and this was done without an inactivity of 30 minutes between two followed queries. This could probably be a bot but with the information provided and without another bot removal process (e.g. defining a maximum for a session duration) this session was not removed. Table 5.8 shows the sessions' length distribution. More than half of the sessions had only 1 query and almost 90% of the sessions (85.92%) had up to 5 queries. Only about 5% of the sessions had 10 or more queries. The mean session length was around 3 queries per session (2.88) with a median value of 1. This value of session length is close to values of European search engine studies like AlltheWeb [JS05] that registered a mean value of 2.2. Although the Portuguese Tumba! search engine study had much lower values (1.45 and 1.42 for 2003 and 2004). Tumba!'s values are much closer to US based search engines like Altavista [SMHM99] (1.39) or Excite [WSJS01](1.6) where users seem to make less queries per session as the mean value is lower and the percentage of sessions with only one query is higher than on European ones. Figure 5.12 shows the distribution of session duration and length. This graphic clearly illustrates that the size and duration of a session are not clear proportional, but bigger sessions have a tendency to have more queries. For earlier duration intervals this tendency is reduced. Even when the session length increases most sessions tend to have a relatively small duration and this value doesn't increase proportionally. The most common pair (duration, length) is 0 and 1 showing the fact that there are many sessions with a single query as seen before. Beyond 500 minutes there are few sessions (only 522) and after 1000 minutes only 50, so these sessions with longer durations are very rare.

Table 5.7: Distribution of Sessions' Duration (minutes).

Duration	**Sessions**	**% Sessions**
[0,1[	10,271,728	65.14%
[1,5[	1,929,860	12.23%
[5,10[	992,000	6.29%
[10,15[	639,500	4.06%
[15,30[	1,209,199	7.67%
[30,60[	570,018	3.62%
[60,120[	133,628	0.85%
[120,180[	14,690	0.09%
[180,240[	3,952	0.03%
[240,∞[	3,379	0.02%
Total	15,767,954	100%

Table 5.8: Number of Queries per Session.

Queries	Sessions	% Sessions
1	8,262,027	52.40%
2	3,019,483	19.15%
3	1,426,250	9.05%
4	838,916	5.32%
5	524,011	3.32%
6	363,898	2.31%
7	255,982	1.62%
8	190,489	1.21%
9	145,437	0.92%
≥ 10	741,461	4.70%
Total	15,767,954	100%

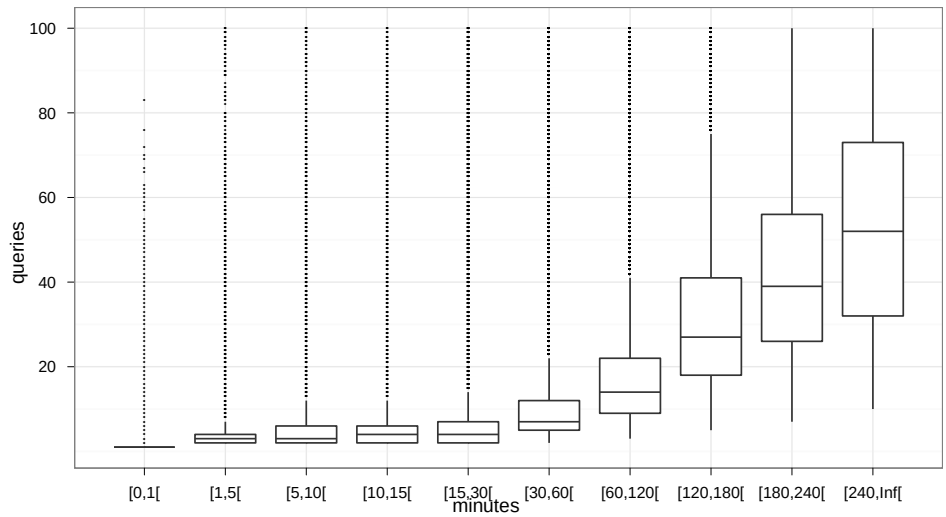

Figure 5.12: Distribution of Session Length vs Duration.

5.4 Query Level Analysis

When some analysis related with terms was made there was always the decision to include or not too common words (like “de” or “o”) that are not relevant for an analysis. This kind of words are called *Function Words* and might include prepositions, pronouns, articles, particles or other words that have little lexical meaning As these words are related with the dataset a list of function words (check Table A.2 in Appendix A) was built from observing the most common English and Portuguese words (the two main languages found on queries). Observing the list of the 1,000 most frequent unique terms (corresponding to half of the total terms) a list of function words was made. For comparison purposes

these words were kept in some cases to analyze the effects of removing these words. Beyond that the accents in all queries' words were removed and all words were converted to lower case since the SAPO's search engine pre-processes the queries by lower casing its terms and by removing accents [Por] (e.g. "Educação", "EdUcacao" or "educacao" are equal terms). As seen before, the total number of queries is 45,413,607. Only 19.59% (8,898,205) of these queries are unique. This means that in every 5 queries of the dataset 4 are repeated, indicating that in many different searches the same queries are made. The number of queries that are never repeated (logged exactly one time) is 5,350,549 (11.78%). All these values are calculated with the removal of function words. If function words were not removed these values would be a little different. The number of unique queries would raise (20.31%) and the same would happened to the number of queries that are never repeated (12.34%). Figure 5.13 shows these results. In this graphic the unique queries were ranked by their decreasing frequency along with the cumulative percentage of all queries; its distribution fits the power law as in other studies [CS10, BYGJ+08]. By caching around 0.4% of the most frequent queries the search engine could respond to 50% query requests, and caching around 20% the search engine could deal with 80%. Removing or not the function words in this case does not seems to have great effect on these results as they are practically the same, with the only difference that without function words the search engine could deal with a little more query requests. The distribution of terms per query (with function words removed) is showed on Table 5.9. Around 90% of the queries have at most 3 terms and only 1% of the queries have 7 or more terms. The mean terms per query is 2.03 without function words (2.31 if function words were not removed). From this values we can conclude that users tend to make short queries. The maximum number of terms in a query was of 1,160. The query text in this case seems to have been copied from elsewhere and used in a search. These very long queries are not frequent, as the number of queries with 10 or more terms is very low (0.43% of the total).

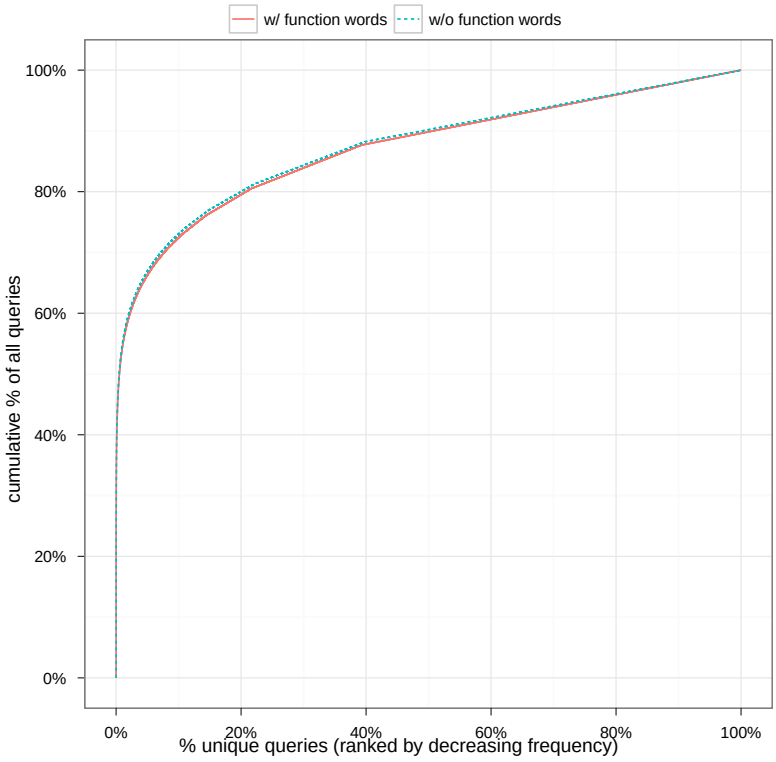

Figure 5.13: Cumulative Distribution of Queries.

Table 5.9: Distribution of Terms per Query (without function words).

Terms	**Queries**	**% Queries**
1	19,970,275	43.97%
2	14,058,607	30.96%
3	6,786,832	14.94%
4	2,661,717	5.86%
5	1,033,327	2.28%
6	409,897	0.90%
7	179,739	0.40%
8	83,189	0.18%
9	36,983	0.08%
≥ 10	193,041	0.43%
Total	45,413,607	100%

Another important topic of analysis was about how users refine or reformulate queries. For each session the first query is defined as the initial one, and all the next ones in the same session are subsequent queries. Following the ideas in other studies [CS10, JSS00], a modified query is defined as a subsequent query with the same information need than the last query; two queries share the same information need if they have at least one equal term. With this in mind and counting the number of terms from a query in a session and

the next one it was possible to see if users tend to add terms (specialization), remove terms (generalization) or both at the same time. In this analysis function words were not removed because some words (e.g "de") could be seen as a specialization of a query.

Table 5.10: Types of Queries Distribution.

Type of Query	**Queries**	**% Queries**
Initial	15,767,954	34.72%
Identical	15,133,268	33.32%
New	8,372,205	18.44%
Terms Swapped	37,084	0.08%
Modified	6,103,096	13.44%
Total	45,413,607	100%

The number of initial queries was 15,767,954 (34.72%), and it is obviously equal to the total number of sessions. Around 33% (15,133,268) of the queries were repeated, meaning that in a particular session two queries in a row were exactly the same. This can happen for many reasons such as the refresh of the search engine's results page or a back-button click leading to the submission of the same query other than a user intentional submitted the same query (this number is very high for this case only). Another kind of queries was new queries, queries with only new terms, indicating a new information need. The number of this type of queries was 8,372,205 accounting around 18.44%. In few cases two queries in a row had the exact same terms but in a different order (e.g. "cidades portugal" and "portugal cidades"). This situation only happened in 37,084 queries which is less than 1% (0.08%). The remaining queries were modified queries. Table 5.10 summarizes these results. Table 5.11 shows the number of terms changed for modified queries. When a user refines or reformulates a query it does that more often by adding terms than by removing. In 43.71% of the modified queries one or more terms were added opposing to 24.14% of the queries where one or more terms were removed. Users tend to reformulate queries by specialization instead of generalization. Also users don't make radical changes to their queries, as almost 60% of the modifications result of removing/adding at most two terms. Zero length changes had the highest occurrence. In this situations a user changes one or more terms but the total number remains the same.

SAPO search engine has some advanced operators [Por] to aid users in the process of search. The MORE operator ("+") which is included in search by default (e.g. search for "Gil + Vicente" is equal to "Gil Vicente"); the NOT operator to exclude results (e.g. Figo -"Luís Figo"); the AND and OR operators to search for pages that include all the terms or at least one of them, respectively (e.g. futebol AND sporting and futebol OR sporting). The SITE operator to match specific domains or websites (e.g. site:.pt); the FILE operator to search for documents in a specific format (e.g. filetype:pdf); the PHRASE operator to indicate that documents should contain that terms by that exact order (e.g. "Gil Vicente");

Table 5.11: Number of Terms Changed per Modified Query.

No. Terms	Queries	% Queries
≤ -5	90,950	1.49%
-4	67,753	1.11%
-3	135,376	2.22%
-2	367,483	6.02%
-1	811,967	13.30%
0	1,962,114	32.15%
+1	1,626,658	26.65%
+2	655,675	10.74%
+3	212,929	3.50%
+4	86,292	1.41%
$\geq +5$	85,899	1.41%
Total	6,103,096	100%

the FILL operator which uses the wildcard "*", telling the search engine to find the best matches (e.g. univ* which matches "universidade", "universo",etc). These operators were not very used by SAPO's users; in only 676,040 queries (1.5%) at least one of these operators was used. Table 5.12 shows the percentages of co-occurrences between each two operators (the intersection between the same operator gives the percentage of queries where that operator is the unique used on a query). The operator most used is the PHRASE operator which is present in almost 80% of the advanced queries. The other advanced operators are very rarely used, namely the FILE operator which is used on less than 1% of the advanced queries. The most used combination of two operators is NOT and PHRASE which makes sense, since this operator could be useful to exclude results by giving a specific phrase (e.g. Figo -"Luis Figo"). Advanced operators seem unknown for SAPO's users or the advantages they provide are insufficient.

Table 5.12: Correlation Between Advanced Operators.

Operator	AND	FILE	FILL	MORE	NOT	OR	PHRASE	SITE	Total[1]
AND	6.387%	0.001%	0.004%	0.008%	0.006%	0.71%	1.42%	0.009%	8.5%
FILE	0.001%	0.029%	0%	0.002%	0.002%	0.0003%	0.258%	0.005%	0.3%
FILL	0.004%	0%	1.427%	0.003%	0.002%	0.0001%	0.01%	0.0001%	1.4%
MORE	0.008%	0.002%	0.003%	6.911%	0.002%	0.0001%	0.05%	0.004%	6.7%
NOT	0.006%	0.002%	0.002%	0.002%	5.456%	0.001%	1.887%	0.009%	7.4%
OR	0.71%	0.0003%	0.0001%	0.0001%	0.001%	0.282%	0.854%	0.002%	1.8%
PHRASE	1.42%	0.258%	0.01%	0.05%	1.887%	0.854%	72.154%	0.107%	76.7%
SITE	0.009%	0.005%	0.0001%	0.004%	0.009%	0.002%	0.107%	1.577%	1.7%

[1]Sum may not be 100% due to rounding

The overall results at this level seem similar with other studies [SMHM99, WSJS01, JS05, JSP05, CS10, JSS00]. There is a lack of studies with the same magnitude of this for comparison; there are some very big collections [SMHM99, BJC+07] with hundreds of millions of queries and many smaller studies [CS10, JS05, JSP05, WSJS01, JSS00] with

hundreds of thousands queries or around one million queries. The number of queries on this study falls in the middle and some comparisons are difficult to be made. The percentage of unique queries is similar with studies having bigger collections like AltaVista [SMHM99] (around 27%) but lower than smaller studies like Tumba! [CS10] (44% and 48% in 2003 and 2004) or an Excite's study [JSS00] (35%). The percentage of never repeated queries is also lower than the Portuguese Tumba!'s study with values of 30% and 34%, which is almost 20% more. As the number of unique queries is low, it is only necessary to cache a few percentage of the most frequent unique queries to deal with most of query requests. Tumba!'s study requires a caching of a little more than 10% to deal with 50% opposing to the percentage of 0.4% of this study. In other studies [SMHM99, WSJS01, JS05, JSP05, CS10, JSS00] the value of mean terms per query is around 2-3; SAPO's users seem to make even shorter queries than in the studies observed (mean value of 2.01); this happens mostly because the percentage of queries with only one or two terms is higher than in other studies (between 20% in AltaVista's study [JSP05] and 40% on Tumba! [CS10]). The number of terms changed in modified queries is very similar with other studies [CS10, JSS00] mainly for zero length changes where this percentage is around 30%. The tendency to keep the same number of terms, or remove/add one term on modified queries evidenced on these studies is also present on SAPO's search engine. SAPO's search engine follow the tendency stating that users from European search engines use less advanced operators than users from US-based search engines [JS06]. The number of advanced queries is much closer to European-based search engines studies like Tumba! [CS10] (11%-13%) and AlltheWeb [JS05] (1%) than to US-based search engines like AltaVista [SMHM99, JSP05] studies with higher percentages (around 20%).

5.5 Term Level Analysis

A term is viewed as a sequence of characters bounded by white spaces. The only exception to this definition were advanced queries operators as they are not counted as terms. As stated before all terms were unaccented and lower cased. In this a distinction was also made between the removal or not of function words. The number of total terms was 89,609,923 (104,608,338 if function words were not removed). Despite this very high number of total terms, only around 3% are unique. The number of terms that are never repeated is the same with or without function words, 1,344,217 corresponding to 1.5% of the total terms (1.3% with function words). The mean number of characters per term is 6.86. This value is lower if no function words were removed, 6.18, which makes sense since most function words are short words with length between 1 and 4 characters. The maximum number of characters in a term was 6,227 and was only resisted for one term. Figure 5.14 shows the distribution of the number of characters per term. The number of terms with more than 40 characters is only 66,626 which is 0.07% of the total. Observing

the graphic it is possible to see that most terms had between 4 and 8 characters. This tendency was kept even after the removal of function words. Around 85% of the terms have less than 10 characters. Figure 5.15 shows the 250 most frequent terms represented in a cloud. The total number of terms is surprisingly high when compared with other studies like Tumba! [CS10] or AltaVista [JSP05] where the number of total terms is around 1 million, which is almost 90 times less than the value observed for this study. Unfortunately in studies from bigger search engines like AOL [BJC+07] or AltaVista [JSP05] the total number of terms is not revealed. The percentage of unique terms is low like in other studies [SMHM99, WSJS01, JS05, JSP05, CS10] registering the lowest percentage of the observed studies. The number of characters per term is very connected to the words used in a specific language. Tumba!'s study [CS10] was the only with the majority of Portuguese. The value obtained in this study is almost the same as Tumba!'s study (value of 6.99).

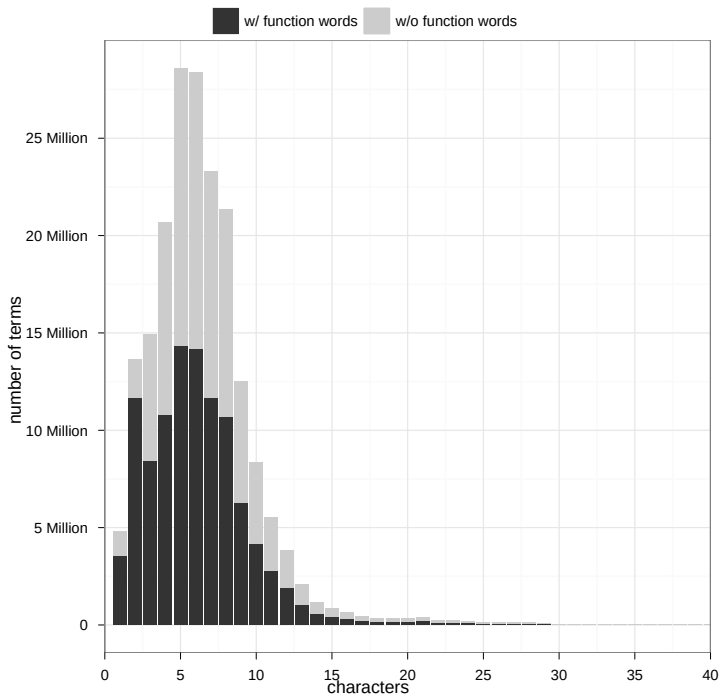

Figure 5.14: Distribution of Characters per Term.

Like the cumulative distribution of queries, this process was repeated for terms (Figure 5.16); the unique terms were ranked by their decreasing frequency along with the cumulative percentage of all terms. This distribution fits a power law. Caching only 2% of the most frequent terms the search engine could deal with 90% of queries containing

[1]Generated using http://www.wordle.net

Figure 5.15: A cloud of the 250 most frequent terms[1].

these terms (the values are again very similar with or without function words). These results are consistent with other studies [CS10, BYGJ+08].

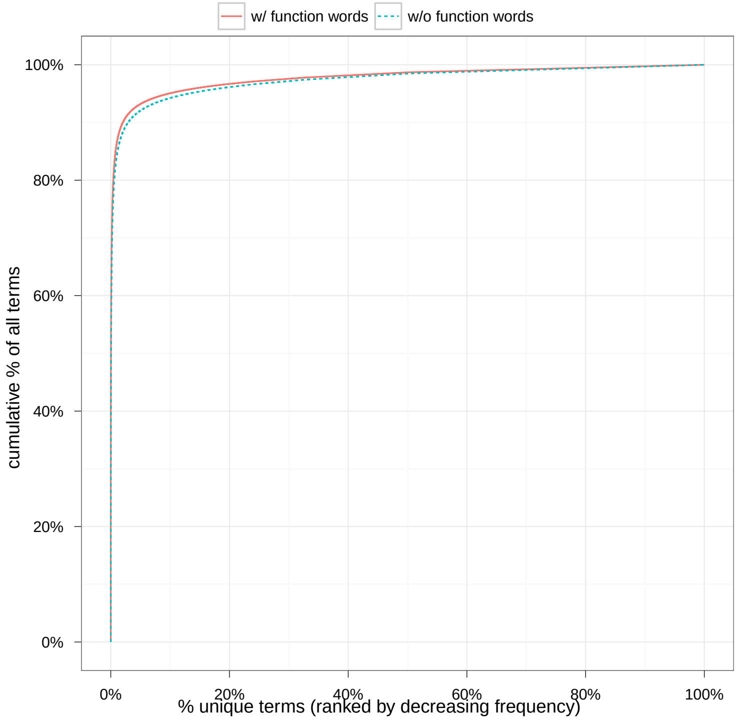

Figure 5.16: Cumulative Distribution of Terms.

5.5.1 Topical Analysis

To find out what kind of information users search for, a random sample of 2,500 queries was manually classified by two evaluators into 11 general topic categories defined by Spink et al. [SJWS02] (presented on Table 5.13). This process was used in many studies [JS05, WSJS01, JSP05, CS10]. By calculating the *finite population correction factor (fpc)* it is possible to see that the sample size is suitable. *Fpc* measures how much extra precision is achieved when the sample size becomes close to the population size [AWW+10]. The equation is shown next (Equation 5.1), where N is the population size and n the sample size. With a sample size of 2,500 and a population size of 45,413,607, *fpc* has a value of almost 1 (0.99997) so with a bigger sample size there is almost no effect on the precision [AWW+10]. In order to have a greater coverage of the overall data the queries were randomly retrieved from the last query of different sessions, which should indicate the last information need that a user tried to satisfy in a particular search session.

$$fpc = \sqrt{\frac{N-n}{N-1}} \tag{5.1}$$

To measure the agreement between the two evaluators Cohen's kappa coefficient was used. Cohen's kappa (Equation 5.2) measures the agreement between two evaluators who each classify N items into C categories (mutually exclusive), and gives a value, k, which is simply the proportion of agreement after chance agreement is removed from consideration [Coh60]:

$$k = \frac{\rho_0 - \rho_e}{1 - \rho_e} \tag{5.2}$$

ρ_0 is the observed proportion of ratings where the evaluators are in agreement and ρ_e is the hypothetical proportion of chance agreement, which is the proportion of agreements that would be expected between the raters if they were scoring randomly. For a good level of reliability kappa should be at least 0.6 or 0.7 [Woo07].

The first classification made by the two evaluators registered a very low level of agreement with a kappa value of 0.43. In only 1,312 of 2,500 queries there was an agreement between the evaluators. The categories were not very specific and many times more than one category was correct for the same query. Having this in mind the discrepancies were solved by making some "rules" about the category to use according to the query's content. Queries about weather were classified as *Commerce, Travel, Employment or Economy* because traveling seems the most logical reason for a person to search for this content. The same explanation can be given for queries searching for hotels, maps, etc. Queries where a user searches for a specific website are classified as *Computers or Internet* because they imply that a user is trying to reach that specific online content. Queries searching for emails or social networks also were classified with this category. When a user searches

for something related to tradition/culture (e.g. search for a recipe) the query is classified as *Society, Culture, Ethnicity or Religion*. The category of *People, Places or Things* is very general especially for *Things*. As such, this particular category was used only if the query could not be allocated into other category. For example the query "Hospital S.João" was classified as *Health or Sciences* as it implies that a user is trying to get some information about health care, but could also be classified as a *Place* (that particular hospital). For queries that no category could be assigned (another topic or unknown content) they were classified as *Unknown or Other*. Table 5.13 shows the results of this analysis ordered by category decreasing frequency.

Table 5.13: Topic Categories Ranking.

Rank	**Categories**	**Queries**	**% Queries**
1	Computers or Internet	672	26.88%
2	People, Places or Things	566	22.64%
3	Commerce, Travel, Employment or Economy	414	16.56%
4	Entertainment or Recreation	290	11.60%
5	Society, Culture, Ethnicity or Religion	143	5.72%
6	Unknown or Other	121	4.84%
7	Sex or Pornography	116	4.64%
8	Education or Humanities	58	2.32%
9	Health or Sciences	53	2.12%
10	Government	46	1.84%
11	Performing or Fine Arts	21	0.84%
	Total	2,500	100%

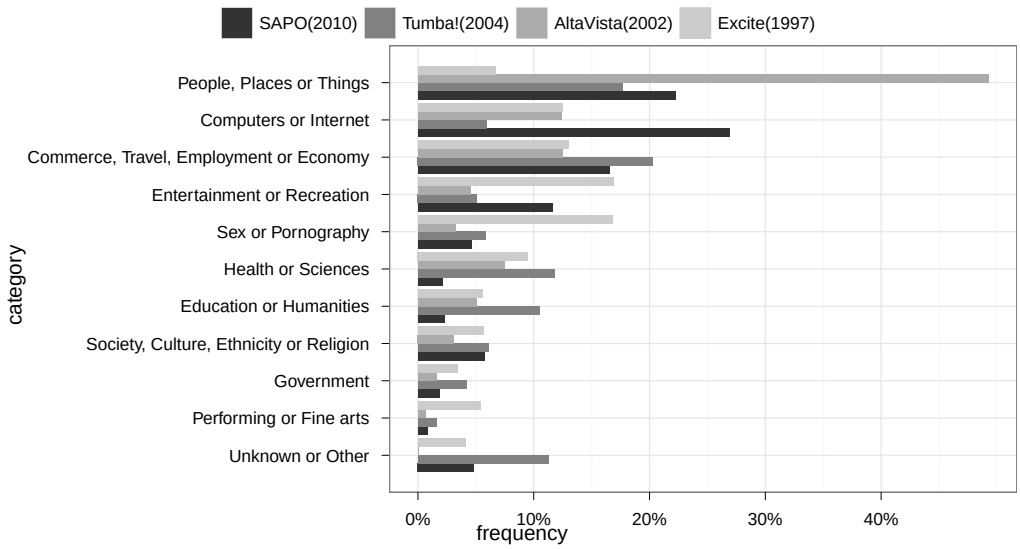

Figure 5.17: Topical Analysis Comparison Between Studies.

Figure 5.17 shows a comparison of the results from this study with three other search engines: Excite [WSJS01], AltaVista [JSP05] and Tumba! [CS10]. These studies have a time range between 1997 and 2010. The most searched category was *Computers or Internet*. Many users search for this type of content, namely for social networks, specific websites, mail services, and even for another search engine. *People, Places or Things* and *Commerce, Travel, Employment or Economy* are close to the first categories. The three first categories account around 66% of the analyzed queries. As in other studies [CS10, WSJS01, JS05] the categories *Commerce, Travel, Employment or Economy* and *People, Places or Things* are in the top 3 positions, supporting Jansen and Spink's [JS06] statement saying that *"The overall trend is towards using the Web as a tool for information or commerce, rather than entertainment."*. Queries related to *Sex or Pornography* this category only accounted for around 5% of the queries analyzed. This behavior was also seen on the Tumba!'s study [CS10] (percentage of 4.90% and 5.80% in 2003 and 2004 respectively). Furthermore the decrease of queries about this topic is clear in AlltheWeb's study [JS05] with a decreasing from 10.8% to 4.5% between the two years analyzed (2001 and 2002); looking at an older study (1997) from Excite search engine [WSJS01] this topic accounted for 16.8% being the second most searched category, which demonstrate this downward tendency. Comparing with the Portuguese Tumba!'s study [CS10] there are some differences. SAPO's users seem to search a lot more about *Computers or Internet* than Tumba!'s users. Furthermore, *Education or Humanities* and *Health or Sciences* are topics rarely searched by SAPO's users unlike in Tumba! where these categories are ranked in the first 5 places. The tendency to observe few searches from categories *Performing or Fine arts* and *Government* is also observed in other studies [CS10, WSJS01, JS05].

To have more backup information about this analysis, Table 5.14 shows the 20 most frequent queries and terms. With these queries and terms it is also possible to check the most frequent topics. Seeing the most frequent 20 terms/queries the first places show many examples of the most seen category like *google, facebook, hi5 or gmail*. Again referring to the 20 most frequent terms and queries most of them fall into these categories. Observing Table 5.14 it is possible to identify many terms and queries related to weather forecast like *"tempo"* (weather) or *"meteorologia"* (meteorology) indicating that SAPO's users are concerned with this topic. This seems to be specific to this search engine or Portuguese users because this behavior was not identified in other studies.

Regarding these 11 topic categories an analysis was made about their hourly and weekly distribution, showed in Figure 5.18 and Figure 5.19. Again as we are not analyzing the full dataset this information could not be very accurately as could exist hours or weekdays from some category that are not in the sample dataset, and still those queries could be present in the full dataset. From the collection of observed studies, only one from AOL search engine [BJC+07] made a similar analysis. In nearly all categories until

Table 5.14: The 20 most frequent terms and queries.

Rank	Query	Occurrences	% Queries	Term	Occurrences	% Terms
1	google	1,191,384	2.62%	google	1,325,675	1.48%
2	facebook	665,548	1.47%	facebook	701,249	0.78%
3	hi5	330,931	0.73%	portugal	516,148	0.58%
4	gmail	297,783	0.66%	sexo	465,670	0.52%
5	youtube	232,589	0.51%	jogos	431,412	0.48%
6	tempo	175,230	0.39%	hi5	366,529	0.41%
7	sexo	157,006	0.35%	www	365,064	0.41%
8	hotmail	148,198	0.33%	tempo	306,587	0.34%
9	financas	113,902	0.25%	lisboa	305,432	0.34%
10	correio manha	96,687	0.21%	sapo	304,528	0.34%
11	banca jornais	84,268	0.19%	gmail	304,074	0.34%
12	meteorologia	83,344	0.18%	2010	296,472	0.33%
13	euromilhoes	77,542	0.17%	youtube	269,130	0.30%
14	cgd	73,457	0.16%	-	269110	0.30%
15	jogos santa casa	72,186	0.16%	casa	265,701	0.30%
16	www	72,025	0.16%	videos	260,734	0.29%
17	rfm	68,640	0.15%	porto	250,218	0.28%
18	portal financas	66,536	0.15%	financas	245,924	0.27%
19	record	66,464	0.15%	gratis	222,387	0.25%
20	porno	65,135	0.14%	porno	201,047	0.22%

8:00 a.m. the amount of query traffic is very low and from 8:00 p.m. a downward tendency is observed. The 3 top categories (*Computers or Internet*, *People, Places or Things* and *Commerce, Travel, Employment or Economy*) show a similar hourly distribution. The highest traffic on these categories is established between 10:00 a.m. and 4:00 p.m. and a clear decreasing after 8:00 p.m.. As a matter of fact in most categories the maximum query traffic is reached on the afternoon period. *Entertainment or Recreation* and *Performing or Fine Arts* are two exceptions to this, reaching the maximum query traffic very late (from 8:00 p.m.). Searches about the topic *Government* also present a different behavior with a maximum query traffic before midday and registering low traffic after 2:00 p.m.. The distribution along the days of the week shows a clear downward tendency as the weekend approaches, in many categories. *Entertainment or Recreation* on the other hand has an higher query traffic on Saturdays and Sundays. The categories *Sex or Pornography* and *Government* show interesting patterns. The first shows a big increase in query traffic on Fridays unlike *Government* that presents high query traffic from Monday until Thursday, decreasing a lot from there.

Although the categories used in AOL study [BJC+07] were not the same it is possible to make some comparison. The hourly distribution does not shows such fluctuation as in this study. In this study the *Sex or Pornography* category shows a very different behavior with very high traffic before 6:00 a.m. and the *Entertainment* category shows a similar behavior. Other categories used in this study like *Health* or *Shopping* showed almost no

hourly variation. The weekly distribution made in this study was made globally (global percentages for the 11 categories) and not independently for each category. This analysis shows a constant weekly popularity for almost all categories. As the categories were not individually identified a direct comparison could not be done.

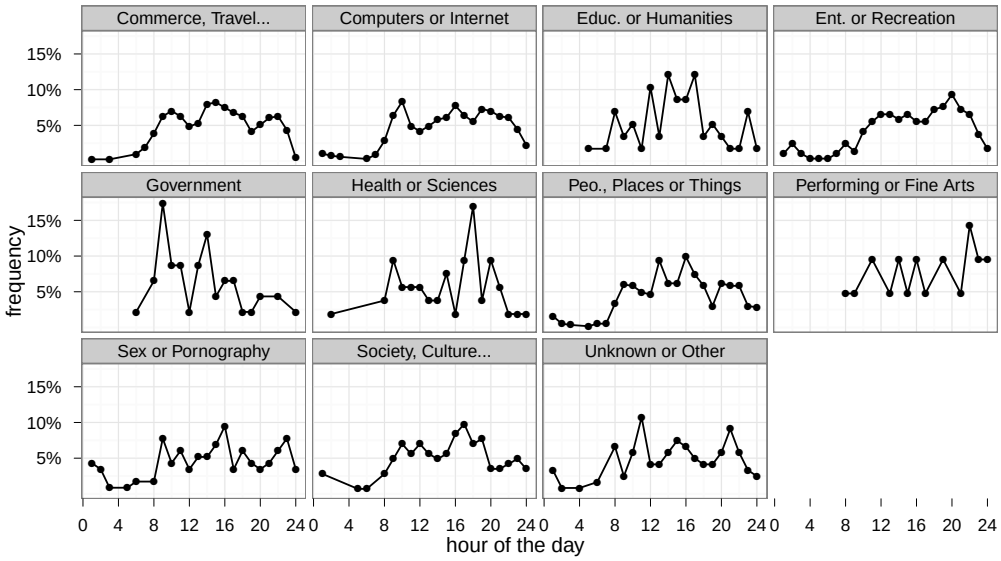

Figure 5.18: Hourly Query Traffic by Category.

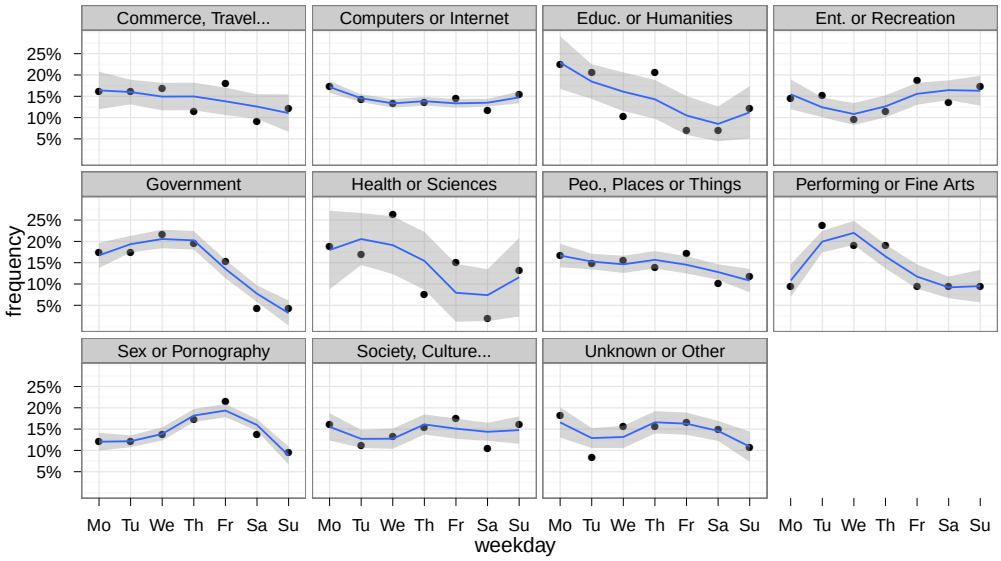

Figure 5.19: Weekday Query Traffic by Category.

Chapter 6

Conclusions

The Web is used daily by people who want to fulfill their information needs. To search and retrieve relevant information for users, information retrieval systems play a very important role. On the Web, search engines are a kind of information retrieval systems and are commonly used as an access point to information. Query Log Analysis appears as research field to provide better knowledge about what and how users search in a search engine. Studying the transactional log files that store the queries made by users, it is possible to characterize the users' behavior and the most searched topics. This analysis can have a direct impact on website-optimization strategies and search engine design. Query Log Analysis has three distinct phases: collection, preparation and analysis. Collection methods are in many cases server-side, and it is common to collect information about user identification (IP address), date and time of the query and the exact query (terms entered by the user). In the preparation phase researchers make different decisions trying to make the stored information more accurate; empty queries and non-human queries are removed and in some cases limitations on the session's duration and length are made, defining a maximum duration and a maximum number of queries per session, respectively. In the analysis phase the information is analyzed according to different parameters, most of them common across studies making the comparison between them possible.

In the preparation phase the removal of bots was essential. Using the *browser* field was possibly to correctly identify a large number of bots. After this step, the removal of too long sessions (more than 100) was also executed. This second step removed many queries and maybe some of them incorrectly. Different approaches could be used to support this method. Defining a number of maximum queries per minute or a minimal interval between queries [DF09] are different approaches for this process that could be safer and produce good results. Oddly the hourly distribution of bots is similar to the hourly query distribution.

The mean session duration was 5 minutes and 23 seconds, with around 65% of sessions lasting less than 1 minute. This value is consistent with Tumba!'s study [CS10] with 6 minutes and 31 seconds and 5 minutes for the years of 2003 and 2004 but much lower than studies were there is no session delimitation such as AlltheWeb [JS05] or AltaVista [JSP05] with values of 2 hours and 22 minutes and 58 minutes and 10 seconds respectively. The mean number of queries per session was 2.88 which is a value close to European based search engines like AlltheWeb [JS05] (2.2) but higher than Tumba!'s study [CS10] (1.45 and 1.42 for 2003 and 2004) or some US based search engines like AltaVista [SMHM99] or Excite [WSJS01] with values of 1.39 and 1.6 respectively. SAPO's users submit few information on each query with a mean of 2.03 terms per query. In other studies this value was between 2.35 in an 1999 AltaVista's study [SMHM99] and 2.92 again in an AltaVista's study [JSP05] from 2002. SAPO's study registered one of the higher percentages of queries with only one term (43.97%). The percentage for Tumba!'s study was about 40% and the lower percentage observed was for AltaVista study [JS05] with only 20.4%. When modifying a query, SAPO's users take 3 main decisions, by this order: change one or more terms but maintain their number (32.15%), add one term (26.65%) or remove one term (13.30%). This exactly behavior is seen on an Excite study [JSS00], with percentages of 34.76%, 19.03% and 16.33% for the same type of modifications. On the other hand, on Tumba!'s study [CS10] the most frequent modification is adding one term (around 35%). SAPO's users follow the tendency to rarely use boolean operators observed in most studies (only 1.5% of the queries). Furthermore, as an European based search engine, this value is close to similar ones like AlltheWeb [JS05] (1%) and lower than US based search engines (values around 20%). Tumba!'s study is between these studies with values of 11% and 13%. Although the number of terms is very high (89,609,923) only around 3% of them are unique. These terms are also short with a mean number of character per term of 6.86.

Topical analysis showed that the most seen category is *Computers or Internet*. As a matter of fact there are many queries related to this topic like queries about social networks, email services, searches for a website by a specific link and others. This study also followed the tendency of more queries about *People, Places or Things* and *Commerce, Employment, Travel or Economy* and less about *Sex or Pornography* and *Health or Sciences* [SJWS02].

Generally Portuguese users comply with users from European based search engines. The most surprising difference is about the most seen topic category, in which the first place is generally occupied by *People, Places or Things* and *Commerce, Employment, Travel or Economy*. Portuguese users prefer to make short queries and fast sessions. The session duration does not seem to be related with the users' origin or culture because this value was similar along studies. SAPO's users registered the highest percentage of sessions with less than 5 minutes with around 77%, where in the other observed studies the

highest was 74.98% from Tumba! [CS10]. SAPO's users, like in other European search engines' like AlltheWeb [JS05] or Tumba! [CS10] do more queries per session than in US based search engines like AltaVista [SMHM99] or Excite [WSJS01]. In US based search engines this value is close to 2 and on European ones closer to 3. Portuguese users do not see advantages on using Boolean operators or simply they are unknown to them or their perceived complexity does not encourage users to use them. Again this value is close to European search engines like AlltheWeb [JS05] (1%) but much lower than US based search engines like AltaVista [SMHM99, JSP05] where the percentage is around 20%. Portuguese users submit few information in their queries, like in other European based search engines like Tumba! [CS10] or AlltheWeb [JS05] where this value is around 2. In fact this value (2.03) is even lower than in the observed studies where the minimum was 2.17 on Tumba! [CS10]. This could be due to the removal of *function words* because if these words were not removed this value would be a bit higher (2.31). This value is higher on US based search engines like AltaVista [JSP05] where is close to 3.

Conclusions

References

[Ada07] Eytan Adar. User 4xxxxx9: Anonymizing query logs. In *Query Logs Workshop WWW*, volume 7. Citeseer, 2007.

[And] Andreas Staeding. List of User-Agents (Spiders, Robots, Crawler, Browser). Avaible at http://www.useragentstring.com/, accessed on June 17, 2011.

[AWW+10] S.C. Albright, W.L. Winston, W. Winston, C. Zappe, M.C.O.N. Broadie, and P.C.O.N. Kolesar. *Data Analysis and Decision Making*. Cengage South-Western, 2010.

[BBD+98] Deborah D. Blecic, Nirmala S. Bangalore, Josephine L. Dorsch, Cynthia L. Henderson, Melissa H. Koenig, and Ann C. Weller. Using transaction log analysis to improve OPAC retrieval results. *College and Research Libraries*, 59(1):39–50, 1998.

[BJC+07] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, Ophir Frieder, and David Grossman. Temporal analysis of a very large topically categorized web query log. *Journal of the American Society for Information Science and Technology*, 58(2):166–178, 2007.

[BJF+05] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David D. Lewis, Abdur Chowdhury, and Aleksander Kolcz. Improving automatic query classification via semi-supervised learning. In *Fifth IEEE International Conference on Data Mining*, page 8. Ieee, 2005.

[BYCBGC06] Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Caro González-Caro. The Intention Behind Web Queries. *String Processing and Information Retrieval*, 4209:98–109, 2006.

[BYGJ+08] Ricardo Baeza-Yates, Aristides Gionis, Flavio P. Junqueira, Vanessa Murdock, Vassilis Plachouras, and Fabrizio Silvestri. Design trade-offs for search engine caching. *ACM Transactions on the Web*, 2(4):1–28, 2008.

[CDMS08] Prabhakar Raghavan Christopher D. Manning and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Coh60] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[CS10] Miguel Costa and Mário J. Silva. A Search Log Analysis of a Portuguese Web Search Engine. *INForum 2010 - II Simpósio de Informática*, 2:525–536, 2010.

[CVn01] Fidel Cacheda and Ángel Viña. Understanding how people use search engines: a statistical analysis for e-Business. In *Proceedings of the eBusiness and eWork Conference and Exhibition 2001*, volume 1, pages 319–325, 2001.

[DF09] Omer Duskin and Dror G Feitelson. *Distinguishing humans from robots in web search logs*, pages 15–19. ACM Press, 2009.

[Fou] The Apache Software Foundation. Apache HTTP Server Version 2.0 Documentation - Log Files. Avaible at http://httpd.apache.org/docs/2.0/logs.html#other, accessed on June 17, 2011.

[GHL03] Luis Gravano, Vasileios Hatzivassiloglou, and Richard Lichtenstein. Categorizing web queries according to geographical locality. *Proceedings of the twelfth international conference on Information and knowledge management*, pages 325–333, 2003.

[Gro10] Miniwatts Marketing Group. Internet world stats, 2010. Avaible at http://www.internetworldstats.com/stats.htm, accessed on June 17, 2011.

[Hea09] Marti A. Hearst. *Search User Interfaces*. Number Ch 1. Cambridge University Press, 2009.

[HS00] Christoph Hölscher and Gerhard Strube. Web search behavior of Internet experts and newbies. *Computer Networks*, 33(1-6):337–346, 2000.

[Ins07] Instituto Nacional de Estatística. Estimativas de População Residente, Portugal, NUTS II, NUTS III e Municípios, 2007. Avaible at http://www.ine.pt/ngt_server/attachfileu.jsp?look_parentBoui=6446446&att_display=n&att_download=y, accessed on June 17, 2011.

[Jan06] Bernard J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library Information Science Research*, 28(3):407–432, 2006.

[JM05] Bernard J. Jansen and Michael D. McNeese. Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Journal of the American Society for Information Science and Technology*, 56(14):1480–1503, 2005.

[JP01] Bernard J. Jansen and Udo Pooch. A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3):235–246, 2001.

[JS05] Bernard J. Jansen and Amanda Spink. An analysis of Web searching by European AlltheWeb.com users. *Information Processing & Management*, 41(2):361–381, 2005.

[JS06] Bernard J. Jansen and Amanda Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.

[JSP05] Bernard J. Jansen, Amanda Spink, and Jan Pedersen. A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, 56(6):559–570, 2005.

[JSS00] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, 2000.

[JST08] Bernard J. Jansen, Amanda Spink, and Isak Taksa. Research and Methodological Foundations of Transaction Log Analysis. *Handbook of Research on Web Log Analysis*, pages 100–123, 2008.

[Kel04] Joyce D. Kelly. Understanding implicit feedback and document preference: A naturalistic user study. *SIGIR Forum*, 38(1):77, 2004.

[KSJ06] Sherry Koshman, Amanda Spink, and Bernard J. Jansen. Web searching on the vivisimo search engine. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 57(14), 2006.

[Kur93] Martin Kurth. The Limits and Limitations of Transaction Log Analysis. *Library Hi Tech*, 11(2):98, 1993.

[Miz97] Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[MT07] G. Craig Murray and Jaime Teevan. Query Log Analysis : Social and Technological Challenges. *SIGIR Forum*, 41(2):112–120, 2007.

[NHLW99] David Nicholas, Paul Huntington, Nat Lievesley, and Richard Withey. Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Online Information Review*, 23(5):263–269, 1999.

[PBY08] Barbara Poblete and Ricardo Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize web documents. *WWW Journal*, pages 41–50, 2008.

[PCT06] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. *Proceedings of the 1st international conference on Scalable information systems - InfoScale '06*, pages 1–7, 2006.

[Pet93] Thomas A. Peters. The history and development of transaction log analysis. *Library Hi Tech*, 11(2):41–66, 1993.

[Por] Portugal Telecom. Ajuda SAPO - Pesquisa. Avaible at http://ajuda.sapo.pt/servicos.html?servico_id=4813, accessed on June 17, 2011.

[RB83] Ronald E. Rice and Christine L. Borgman. The Use of Computer-Monitored Data in Information Science and Communication Research. *Journal of the American Society for Information Science and Technology*, 34(4):247–256, 1983.

[SDP06] Arun Sen, Peter A. Dacin, and Christos Pattichis. Current trends in web data analysis. *Communications of the ACM*, 49(11):85–91, 2006.

[SJWS02] Amanda Spink, Bernard J. Jansen, Dietmar Wolfram, and Tefko Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[SMHM99] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

[Tel06] Portugal Telecom. História do SAPO, 2006. Avaible at http://mundo.sapo.pt/artigos/2007/03/06/hist_ria_do_sapo/index.html, accessed on June 17, 2011.

[Tim06] New York Times. A Face Is Exposed for AOL Searcher No. 4417749, August 2006. Avaible at http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=1, accessed on June 17, 2011.

[TRM11] Jaime Teevan, Daniel Ramage, and Meredith Ringel Morris. TwitterSearch : A Comparison of Microblog Search and Web Search. *Search*, 2011.

[Use] User Agent String.Com. Avaible at http://www.user-agents.org/index.shtml, accessed on June 17, 2011.

[VY07] Ganesan Velayathan and Seiji Yamada. Behavior based web page evaluation. *Journal of Web Engineering*, 6(3):222–243, 2007.

[W3S] W3Schools. Browser Statistics- Web Statistics and Trends. Avaible at http://www.w3schools.com/browsers/browsers_stats.asp, accessed on June 17, 2011.

[Woo07] James M. Wood. Understanding and Computing Cohen's Kappa: A Tutorial. *Web Journal WebPsychEmpiricist*, 2007.

[WSJS01] Dietmar Wolfram, Amanda Spink, Bernard J. Jansen, and Tefko Saracevic. Vox populi: The public searching of the web. *Journal of the American Society for Information Science and Technology*, 52(12):1073–1074, 2001.

Appendix A

Appendix

A.1 Bots' Terms

Table A.1: Strings used as Regular Expressions for the Removal of Bots.

feed	bot	rss	^sapo
spider	crawler	yahoo	libwww-perl
check_http	lwp	autoproxy	Apple-Pubsub
Jakarta	Microsoft URL	AppleSyndication	^Microsoft Office
vb project	bloglines	zend	scoutjet
WordPress	Atomic_email	^Mozilla/4.0$	^Mozilla/5.0$
Microsoft-CryptoAPI	AppEngine-Google	Mail.Ru	PostRank
NSPlayer	Rome Client	FINLY	VERSION_TABLE
Akregator	Azureus	curl	utorrent
HTTP agent	OutlookConnector	System.Net.AutoWebProxyScriptEngine	Microsoft BITS
ESS Update	OSSProxy	aol/http	NetworkedBlogs
AdminSecure	PubSubAgent	SOAP	^w3af
charlotte	stackrambler	mediapartners-google	newsgator
bittorrent	Contacts	Anonym	BTWebClient
WLUploader	MPFv	JNPR	OpenCalaisSemanticProxy
checker	greatnews	Winhttp	Drupal - 37
vlc	DataCha0s	Apache-HttpClient	unchaos
DAP	validator	activesync	cache
kevin	webcopier	Publishing	Aberja
CE-Preload	Infoseek	Sitewinder	webcollage
^java	PF:INET	plagger	python
^PHP	liferea	Ruby	BlogzIce
^CFNetwork	ichiro	^Windows-Media-Player	Windows-Update-Agent
reaper	Twitturly	Bookdog	Referrer Karma
ia_archiver	findlinks	facebookexternalhit	CHttp
Kaspersky	Wget	Transmission	AltaVista
Reeder	reader	nutch	PuxaRapido
Sphider	!Susie	CheckLinks	vagabondo
agadine	research	Yandex	silk
larbin	l.webis	iTunes	

A.2 Function Words

Table A.2: Function Words Removed.

e	em	uma	la	your	if	me
de	no	por	of	about	either	by
da	na	in	era	you	sua	
do	nos	ao	nao	be	suas	
dos	nas	sobre	ou	our	onde	
das	para	the	onde	ser	quem	
a	com	se	entre	this	qual	
o	que	mais	sem	not	my	
as	como	sao	for	are	meu	
os	um	on	to	here	aos	