SPATIO-TEMPORAL SEARCH LOG ANALYSIS

Nelson Veríssimo
Instituto Superior Técnico
Universidade Técnica de Lisboa
Lisboa, Portugal

2009

ABSTRACT

This work relates to the field of search log analysis. Applications for performing this type of analysis are rare, their development details, source code and documentation are not made available to the research community and they are restricted to work only with specific search logs. Having acknowledged this limitation and identified the existence of this gap in this field of research, this work has had the objective of planning, designing, developing and testing an open-source web application that is adaptable and capable of performing useful and relevant spatio-temporal analysis for a multitude of different search logs. It includes a number of innovative features as well as a combination of characteristics found in existing tools such as Google Trends and Google Zeitgeist but, unlike these, it is not restricted to a specific log. The application provides mechanisms to deliver information from different perspectives. It provides several overall log statistical measures as well as a set of custom perspectives created by a combination of three independent filters: the query level, the time period level and the location level.

Keywords

Search log analysis, web application, georeferentiation, log statistics.

1. INTRODUCTION

A search log is a specific type of log in which the recorded transactions refer to the user's actions and the search engine responses. It provides a scalable approach for collecting large amounts of data, in an unobtrusive way, leading to unbiased data collection that reflects real user behaviour based on real information needs. Search log analysis is a field of research that is concerned with extracting, organizing, filtering, analyzing and summarizing information from search engine logs, ultimately giving meaning to otherwise complex, uninformative and unstructured raw datasets. The information that is obtained is valuable for answering a wide range of questions about what people search about and how people search the web. There are several different reasons for conducting log analysis. These include:

- **Improving web search systems** - by analyzing how people search for information, search engine developers can design more efficient and accurate mechanisms for providing better results. Several studies have been made to improve search engine results, including the organization of search results [1], the recommendation of better queries (e.g. by using query expansion) [2,3,4] or the recommendation of related documents [4]. Search logs have also been analyzed for geographical purposes (e.g. for providing geographically-related results based on the user's location) [5,6,7,8,9].
- **Discovering informational trends and user interests** - user information needs change over time, and as these changes occur, log analysis enables the understanding of what web search trends and user interests are emerging and declining, within the various global regions [4,10,11]. The benefits obtainable from this source of information are vast. For instance, advertising companies can have a better insight on how users access the web and can target adverts to specific groups of users [12].

1.1 Problem statement, objectives and contribution

Applications for providing useful information from search logs exist but are rare, typically proprietary, and are not applicable to general search engines. Web applications such as Google Trends and Google Zeitgeist are examples of search log analysis systems that are available for general public use. However these applications are restricted for use with the logs from the Google search engine. For that reason, the aim and the contribution of this work was to design, develop and test an open-source web application for performing log analysis over general search engine logs. The Web application was built on top of a relational database and combines features from Google Trends and Google Zeitgeist and also presents some extra new features like general log statistics information and support additional spatio-temporal filtering mechanisms.

Validation was performed by means of a characterization study over a search log. This study served as the basis for evaluating the practicality and usefulness of the information that can be drawn from the output interfaces of the Search Log Analysis application. The search log used for this purpose was a portion of the *Biblioteca Nacional de Portugal* (*the National Library of Portugal*) search engine logs, with approximately 1GB in size and comprising about two months of search activity.

The tests performed with the search log analysis web application show that the application is able to provide answers to "what?", "where?", "when?" and "who?" questions, hence proving that it has useful applicability.

1.2 Overview of this document

This document is composed of six sections plus references. It is organized as follows: Section 1 is the current section and introduces the topic of search log analysis, presents the problem statement and proposes a solution. Section 2 introduces a brief definition of the concepts that are commonly used in the related literature, and presents a brief survey in the field of search log analysis. It describes the typical stages involved in the process, and presents some log visualization and exploration tools. Section 3 describes the solution approach, focusing on the data model and the general decisions regarding data storage and indexing. Section 4 presents an overview of the Web application. Chapter 5 presents the method used for evaluating the web application. It describes tests with a 1GB log from the *Biblioteca Nacional de Portugal* (*National Library of Portugal*) search engine. Finally, Chapter 6 summarizes the conclusions.

2. CONCEPTS AND RELATED WORK

This chapter presents a survey on the subject of log analysis, mainly focusing on search log analysis. The first section introduces the terminology and the main concepts used in the context of this research area. The second section, the typical processing stages involved in search log analysis are described, with basis on a study of the related literature. Afterwards, some interactive applications for performing log analysis are mentioned.

2.1 Concepts and terminology

In the IR literature, a **term** is any series of characters separated by white space (e.g. a word). A **query** represents a user's need for information and consists on the verbalization of that specific need, using a set of one or more terms. Queries are submitted to search engines in order to satisfy the specific user requirement and, in response, the search engine provides **result pages** containing ordered links to web documents. The queries are recorded in the **search engine logs**, as well as the result clicks that are associated. A **result click** consists in the document selections that individual users make from the result pages. The study of the result is also known as **clickstream analysis** and is useful for extracting user activity information. A **query instance** is a single query submitted to a search engine in a defined point of time, followed by zero, one or more result clicks. A **query session** (or simply session) consists of a sequence of related query instances performed by a single user within a small range of time, together with all the result clicks made during this range of time. Sessions are used to identify a single information need, as well as to identify the refinement process of the queries performed by the user.

2.2 Typical stages of log analysis

This section outlines the three typical stages of search log analysis, namely i) data collection, ii) data preparation and iii) data analysis [13].

2.2.1 Data Collection

Data Collection consists of gathering resources for the analysis. Transaction logs are usually generated automatically by the Web servers, and collected in server access logs (i.e. the log files). Search logs are a specific type of transaction logs. The set of fields that compose a search log may vary from search engine to search engine although the following are often present:

- User identification – the IP address of the client's computer that interacts with the system, or some unique identifier based on a server cookie..
- Date/Time – a timestamp indicating when the query was submitted.
- Query terms – the words that define the search, exactly as submitted (i.e. the query).
- Result click – a document from the results that the user has selected.

2.2.2 Data Preparation

Before data analysis is performed, several operations must be made in order to transform the data into a form which can be examined for patterns and relationships. In the case of the analysis procedures that rely on a relational database, the preparation stage typically consists of the following steps:

1. Parsing and importing the data into the database - involves decomposing the log file into individual fields and also indexing the fields [14].
2. Cleaning the data - consists of searching, identifying and removing the corrupt and empty record entries that exist in the data. Record corruption is usually characterized as the existence of fields with no content or with an incorrect type of content [15].
3. Identifying sessions - The session identification step is concerned with defining the boundaries of the sessions in order to separate different information needs, specified by multiple queries within the session by a single user. Existing methods rely on the IP address and timestamp associated with each request, and are usually based on heuristics or statistics [4,16,17,18].

To structure the log database for his research, Jansen used a relational database corresponding to the Entity-Relation (ER) diagram shown in Figure 1 [13].

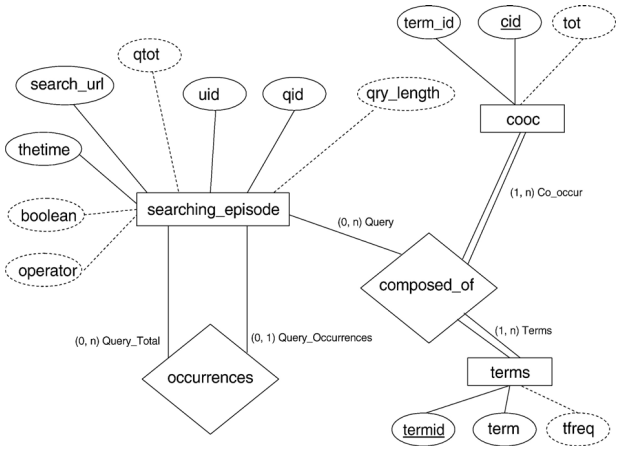

Figure 1. Entity Relation (ER) diagram for modeling web searching log. (Jansen [13]).

2.2.3 Data Analysis

Logs are typically analyzed using two types of approaches: the statistical approach [7,10,11,19,20,21] and the detailed log analysis approach [7,21,22]. Each one has its own strong and weak points. Detailed log analysis can uncover important aspects, but it may also provide biased and inaccurate results due to the small samples that are usually studied. The statistical approach is not proficient at uncovering very specific aspects but it provides scalable means for analyzing large populations.

There are three levels of analysis when dealing with search engine logs: i) session level analysis, ii) query level analysis and iii) term level analysis. This work addresses the three levels, although giving particular emphasis to the query level.

2.2.3.1 Session level analysis

Session level analysis is concerned with aspects such as finding the number of sessions during a certain period of time, finding the session durations, finding session patterns (e.g. similarity among sessions) and finding the document resources that have been consulted in the context of sessions (result clicks) [13,23]. Silverstein et al, [16] claimed that queries regarding a single information need come clustered in time, and that there is a gap before the user returns to the search engine.

2.2.3.2 Query level analysis

Query level analysis focuses on the queries that the users submit, monitoring their frequency and their changes over time and space, as well as examining and categorizing them

according to different classification schemes. Based on the user's intentions, Broder classified queries into 3 categories: Navigational queries, Informational queries and Transactional queries [22]. A navigational query is defined as a user's expression of intent to reach a particular site that he or she assumes that exists. An informational query is a query that expresses the intent to gather information regarding one or more topics that may be present in several result pages. A transactional query is one that expresses the intent to reach a site where sophisticated interactions will take place, usually related to the acquisition of goods such as images, videos or music.

This classification has been adopted and extended by several other studies [12,20,24,25].

In the past, search queries have also been analyzed for geographical purposes. Gan et al, defined a geographic search query as a "*query that employs geographical terms in an attempt to restrict results to a particular region or location*" [7]. Several studies have been made regarding these queries [5,6,7,9]. Sanderson and Kohler presented a characterization study over geographic queries, measuring their frequency, topics, length and spatial relationships [6]. Zhang et al. studied how users reformulate their geographic terms after non-satisfactory results are presented by the search engine [9].

2.2.3.3 Term level analysis

Term level analysis focuses on the study of individual query terms, addressing features like term occurrence in queries and occurrence distribution, query length in number of terms, high and low usage terms and co-occurring terms.

Co-occurrence measures the association between terms, that is, the frequency with which any two terms occur together within the same queries in the search log [13,26]. If two terms occur in the same query then there is a bond between them.

2.3 Exploratory interfaces

The goal of visualizations is to provide a new way of exploring data, helping to provide better insights into vast amounts of data given the human visual capabilities to spot trends, patterns and anomalies. Session Viewer [21] is a log visualization tool that enables both statistical and detailed log analysis and supports hypothesis generation. It enables users to navigate in the visualization hierarchy (session populations, sessions and queries) and see the impact of certain aspects on the displays of each level. This application is proprietary and not available to the research community..

Google Trends[1] is a web application for term level query log analysis. It charts how often a particular search term is presented in queries in relation to the total search volume across various regions of the world. It also allows the user to compare the volume of searches between two or more terms. The main interface is based on a time series chart plotting term frequencies. On another graph, popularity is shown by region, city and language. It is possible to highlight the main graph by region and by time period.

Google Zeitgeist[2] is a website that provides annual reports of the most common queries, organized by several categories that were submitted to the Google search engine.

[1] www.google.com/trends/

[2] www.google.com/zeitgeist/

The main limitation of applications like Google Trends and Google Zeitgeist is that their use is restricted to proprietary logs.

The Search Log Analysis Web application that is proposed here is based on the Google Trends and Google Zeitgeist Web applications. It inherits many of the concepts from these two applications and tries to fill the gap left by these applications by widening the array of possible logs to be analyzed and also introducing new features. In addition, the Search Log Analysis Web application features new functionalities that include session, query and term analysis, as well as spatio-temporal analysis and filtering capabilities.

3. Solution Approach

Figure 2 illustrates the ERD for a database that can support search log analysis applications.

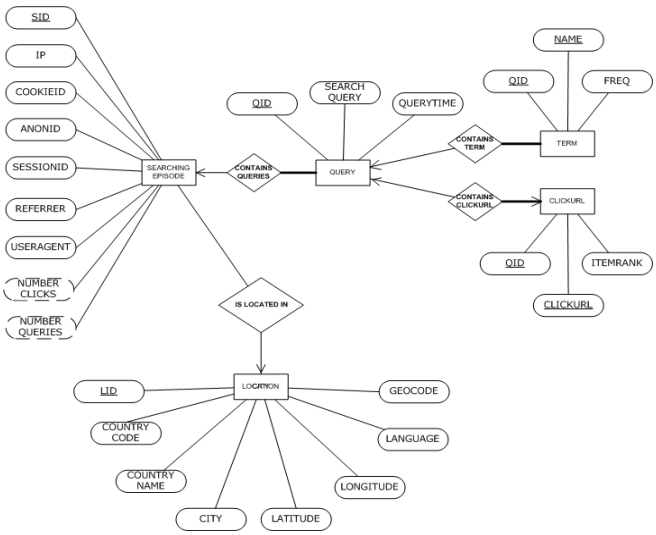

Figure 2. The Search Log Analysis Entity Relational Diagram.

The data model comprises a set of five entities: the Searching Episode, the Query, the Term, the Clickurl and the Location. The Searching Episode entity, which stores client and session related data, comprises the sequences of search interactions performed by a single user. Each Searching Episode includes the following attributes that may be or not be defined according to their availability on the specific log dataset source:

- User identification which may be the IP address of the client's computer or another user identification code.
- The cookie identifier for helping track the history of the user activities.
- A session identifier for helping separating searching contexts and analysing the intentions and the outcomes of the user searches.
- The referrer which is the address of the previous webpage from which a link led to the current page.
- The user-agent for identifying the client's platform details. It usually identifies the browser and operating system. It may also identify the usage by indexing robots or spiders which are automated programs searching the internet.

A Searching Episode contains one or more queries. The Query entity stores the complete search queries, as introduced by the user, and includes a time stamp registering the date and time of the requests. The Query entity has relations to both the Term entity and to the Clickurl entity. The first is a one Query contains many Terms relation, as the Term entity stores the set

of individual terms existing in a specific query and their frequency. The second is a one Query contains one Clickurl relation, as for each query there is at most a result that is selected. The Clickurl's itemrank attribute corresponds to the ranking of the result as provided by the search engine results page. Finally, the remaining entity is the Location entity which provides the system with a set of geographical related features. These features are essentially names of regions and their corresponding geospatial coordinates. The relation with the Searching Episode entity means that each Searching Episode may have its geographical location determined, allowing the reports to the include location related information as well as allowing location related searches and filters. In more detail, the location entity stores the following set of location attributes: 1) country code, 2) country name, 3) city name, 4) latitude, 5) longitude, 6) language, 7) geocode (which will be described ahead. It is important for pinpointing locations on the map and for performing searches and filtering results based on map areas).

In order to increase system performance, some adaptations were made to the data model. Some denormalization techniques have been followed. For instance, since each query belongs to one and only one Searching Episode, and each Searching Episode as its origin in one location, it was decided to provide each query with data regarding its origin, including, for that matter, a column location identifier in table Query.

For optimizing data retrieval, three types of indexes are applied to the different types of data: B-tree index, Fulltext index and Spatial index. The B-tree index is the most used for indexing the data in this system's database and unless stated otherwise it is the default index. A Fulltext index is used for providing complex text searching capabilities against data stored in character-based fields, while a Spatial index optimizes the performance of spatial queries and the retrieval of all geometry based data. Indexes add a certain amount of overhead for inserting rows into the database, and a trade-off had to be made. The main idea has been to optimize the retrieval of data in sacrifice of insertion speed.

The data model described in this section integrates most of the concepts proposed by Jansen [13] and presented in the entity-relational diagram of Figure 1. Just like in that data model, the Searching Episode stores the user and session details, with the exception that, in this work, queries are stored in a separate table. The rationale is that each Searching Episode may contain several queries and it is not necessary to store repeated data regarding the user and the session. Another important difference between this and the model proposed by Jansen is that in his work he decided to store term co-occurrences, while, in this work, what is stored are the terms for each query. The reason for this is because the terms that need to be presented in the analysis result pages correspond to the top terms forming the search queries that match the analysis criteria.

This data model also contemplates geographical data which is responsible for the georeferentiation features of this system.

3.1 Working with GeoLocation data

The data that is used in this work consists in an extensive list of IP address ranges and their assignment to geographic locations, down to the city level of granularity, including several details such as the country name and code, city name, latitude and longitude.

For providing the georeferentiation functionalites, this system takes advantage of the spatial functionality offered by MySQL's geographic information system (GIS) extension. Two GIS classes are particularly relevant in this work: the point class and the polygon class. According to the MySQL documentation, "a *Point is a geometry that represents a single location in coordinate space" and "a Polygon is a planar Surface representing a multisided geometry. It is defined by a single exterior boundary and zero or more interior boundaries*". Points are used to define geographic coordinates in terms of latitude and longitude parameters. Polygons, on the other hand, are used for defining and representing areas.

One GIS function is particularly important in this work: the MBRContains function. It receives two geometry elements and indicates if the Minimum Bounding Rectangle of one element is contained in the Minimum Bounding Rectangle of the other element. In this work, this function is used to indicate which points are contained inside a given area or polygon.

Finally, the GIS extension provides spatial indexing support for optimizing operations of georeferentiation.

3.1.1 Determining the locations

Knowing the IP address range assigned to each specific city enables the possibility to, given any client's IP address, determine that client's approximate location.

Figure 3 shows a sample of the geolocation data used in this work. It consists of two tables where the left table maps IP ranges to locations and the right table provides the location details. Location is determined by means of a table key (i.e. the location ID).

IP RANGES to LOCATION

IP_FROM	IP_TO	LOCID
75714560	75714815	25445
75714816	75715071	55874
75715072	75715199	10487

LOCATION

LOCID	COUNTRY_CODE	COUNTRY_NAME	CITY	LAT	LON
10487	US	United States	Grand Prairie	32.6606	-97.0249
25445	US	United States	Richardson	32.9716	-96.7058
55874	US	United States	Irving	32.8569	-96.9629

Figure 3. A sample of the geolocation data used in this work, showing the connections between the two tables.

Another common operation involves determining the locations existing inside a specific area. Each location has Latitude and Longitude values that can be used to define a point class element. This is the case of the Geocode attribute that was mentioned in the data model section. A point may or may not be inside in a specific area (i.e. a polygon). This is determined by using a MySQL GIS function (the MBRContains function), which receives the polygon and the point and indicates if the point is "contained" inside the polygon. In addition, using a similar method, the same function can be used for determining which points exist inside any specific area.

3.2 Data Preparation

The data preparation stage involves preparing the log and the geolocation data. It is achieved by running a MySQL script containing the preparation details. This script uses the UTF-8 character encoding to allow the representation of any character in the Unicode standard.

3.2.1 Preparing the Log Data

Data preparation starts by parsing the log data. Logs from different sources are stored in different databases and usually require a different parser as it must be custom made to reflect the log's structure and data. The parser scans the data, identifies

tokens, breaks the distinct data fields and stores them in the adequate tables and columns. The data fields depend on the availability of the specific log dataset and thus may include the following set or a subset of the fields: 1) IP address, or another user identifier, 2) cookie identifier, 3) session identifier, 4) referrer, 5) user-agent, 6) search query, 7) querytime, 8) clickurl 9) itemrank.

Fields 1 to 5 are stored in table Searching Episode, fields 6 and 7 in table Query, and fields 8 and 9 in table Clickurl.

IP addresses are usually represented by a series of four one-to three-digit numbers separated by periods. In order to simplify operations and optimize system performance they are stored in the database as an integer.

In logs such as *Biblioteca Nacional de Portugal*, a subset of data fields are contained in a URL address, so it is necessary to detect, extract and decode those fields from the URL and store them in the database. Queries are stored and indexed with a fulltext index and its terms are separated, have the non alphanumeric characters removed, and then are stored in table Term. Afterwards, a copy of the table Term is created, and the stopwords of the English, Portuguese, Spanish, German, French and Italian languages are removed.

3.2.2 Preparing the geolocation data

It is during the geolocation data preparation stage that the location of every request is determined, removing that computation load from execution time.

The entire location data must be stored in the database. During this step, each table row gets an additional field, the geocode, which is a point defined in terms of the specific Latitude and Longitude values of that location. This field is indexed using MySQL GIS spatial index.

As Figure 3 illustrates, the approximate location of a given IP address can be determined by finding the range in which that value belongs to, that is where:

```
IP FROM <= IP ADDRESS <= IP TO
```

Every request having an IP address is associated with its corresponding location. This is done by including the location identifier in each of the rows of the Searching Episode table and in each row of the table Query.

An additional step in the geolocation preparation consists in determining the most common language for each location and storing it the Location table. It is important to note that the language is based on the official language of the countries where the queries were performed. In cases where more than one official language applies, the most frequent language is selected.

4. Web Application for search log analysis

The Search Log Analysis web application is a system designed for analyzing logs, specifically the logs of search engines. The application provides mechanisms to deliver information from different perspectives, ranging from a global or general perspective to a set of custom perspectives created by a combination of filters. The global perspective provides a general picture of the searching activity, including several overall statistical measures. The filtered perspectives may adapt to meet the specific information needs of the users. In terms of the filtering capabilities, this system allows the analysis to be conducted over three distinct levels: the query level, the time period level and the location level. The three levels are independent, meaning that it is possible to conduct the analysis using any combination of filters from the three levels.

The query level covers three distinct types of filters. In this context the user may:

- Filter for query terms. The application returns a set of results reflecting the search log queries that comply with the selected set of terms. These results include the most frequent locations where the searches were made, the most common queries, the most common terms existing in the queries containing those terms, the most common websites associated with the search of those terms, and the users that most often performed searches using those terms.
- Filter for users. Given that the log source provides data for identifying users, such as an IP address or other identifier, it is possible to analyze information regarding user activity, including the location of the user or users, the most common searches, the frequency of searches over time and the most visited websites.
- Filter for URL addresses. Given that the log source provides data regarding the selection of URL addresses in the search results, it is possible to determine the most common queries and query terms associated with the webpage, the frequency of user clicks over time, the list of the most common visiting users and the geographical locations with most visitors.

Up to five queries, delimited by commas, may be analyzed simultaneously. This enables the comparison of the analysis results. The reason to this limit is due to the compromise between usefulness and performance.

The time level is taken into consideration when a user selects a specific time period for the analysis. The user may restrict the analysis results to match a specific day, month or year.

The location level allows the restriction of the results to be based on a selected region. The selection of a region enables the retrieval of information regarding that specific area. This is possible through the analysis of the IP addresses contained in the logs. Note that this level of analysis is disabled for logs that don't contain the IP addresses of the machines performing the queries. Filtering results based on a location is done by selecting a country, selecting a city or by defining a custom area on the world map.

The information is provided in both tabular and visual formats. The tabular information covers a set of different categories, namely location names, queries, terms, users and websites. The displayed categories depend on the context of the analysis and consist in the list of its most frequent elements.

The visual information is displayed through the use of charts displaying frequencies and through a thematic world map pinpointing contextual locations.

4.1 Analysis through information categories

The results of the analysis are presented as a set of distinct categories. The set of categories depends on the available data provided by the log source and the type of search query. Table 1 describes the categories that exist for each type of search query, given that the required data is provided in the log dataset. Each category is composed by the set of its top ranked elements. For every element there is a value corresponding to its absolute frequency in the analysis results, and a percentage value that informs the proportion of the absolute frequency in relation to the whole set o results. Up to ten results are displayed, and beyond that value there is information of the number of the remaining elements, their summed absolute

frequency, and once more, the percentage that they represent to the whole set of results.

Table 1. Description of the existing categories for each type of search.

Category	Type of Search			
	Term-related	User-related	Website-related	Area-related
Top Countries	YES	NO	YES	YES
Top Cities	YES	NO	YES	YES
Top Languages	YES	NO	YES	YES
Top Search Queries	YES	YES	YES	YES
Top Terms	YES	YES	YES	YES
Users also searched for (Top)	YES	NO	YES	YES
Top IP addresses	YES	NO	YES	YES
Top Website addresses	YES	YES	YES	YES

4.2 Visual data analysis through charts

Charts are often used to make it easier to understand large quantities of data and the relationships between different parts of the data. The web application uses bar charts and line charts and to present query frequencies. Bar charts are used to display absolute frequencies of queries. Line charts are used to present time series data, plotting the query frequencies over time within the selected time period (year, month, day, entire log span).

The chart display provides users with a simple and powerful mechanism for analyzing query usage trends over time and for drilling down the time periods from a high-level view (covering several years, for example) down to more detailed views (such as displaying data by hour or by day, for example).

Given that users can type up to five different queries (separated by commas), the time series chart can show up to five different lines, each representing a different query and having a different color. This provides a visual and direct comparison of the usage patterns across different queries for the selected time period.

4.3 Visual data analysis and exploration through maps

The developed application supports two ways to visualize information on a map. The first visualization type consists of a map having a set of markers displaying the contextual locations also presented in tabular format. Figure 5 illustrates this type of visualization. The information depends on the type of query. For user related queries the map displays specific user locations. For term related queries and website related queries, the map displays the locations where those terms or websites where most frequently used. Up to ten markers may exist on the map corresponding to the highest related cities presented in the Top Cities category. In order to relate the mapping information with the tabular information, both the city names in the Top Cities category and the markers on the map are identified with letters.

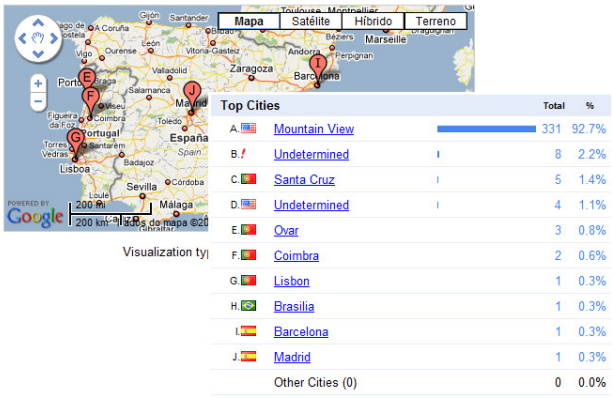

Figure 5. An example of a map illustrating the regions where the specific search was conducted.

The second visualization type is a dot map which is a geographic representation that uses colored dots to display the distribution and density of the search activity across a geographic area. The dot map has a color code that uses three colors: green, yellow and red to represent increasing levels of density. Areas with no color indicate the absence of search activity for a given analysis context. Figure 6 portrays an example of a dot map displayed by this application.

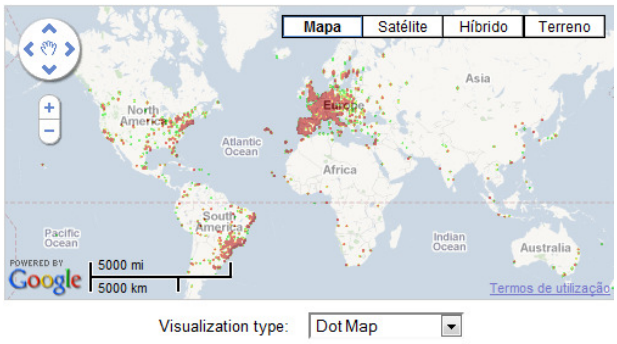

Figure 6. Example of a dot map illustrating the distribution and density of the search activity across the globe.

5. EXPERIMENTAL EVALUATION

Through testing and usage of the Search Log Analysis application, this section illustrates the type, practicality and usefulness of the analysis and check whether the hypothesis presented in Section 1 is validated and confirmed. The testing is considered successful, proving the hypothesis, if it is shown that the application can be used to perform spatio-temporal search log analysis, providing useful and relevant information.

In order to formalize the testing and the validation of the hypothesis, the DIKUW hierarchy model is used as a framework [27]. Ackoff classifies the content of the human mind into five categories:

1. **Data**: Symbols, in a raw format, without a significance or meaning of itself
2. **Information**: Data that are processed to be useful; provides answers to "what", "where", "when" and "who" questions
3. **Knowledge**: Application of data and information; answers "how" questions
4. **Understanding**: Appreciation of "why"

5. **Wisdom**: Evaluated understanding

Considering the above, a characterization study of a search log is presented in the next subsection in order to help testing if the information layer can be reached from the analysis of the log data.The log used for this study originates from the search engine of *Biblioteca Nacional de Portugal* (*National Library of Portugal*), is approximately 1GB in size and contains data for approximately two months of search activity

5.1 Characterization study of the *Biblioteca Nacional*'s Search Log

Testing focuses on the interpretation of the results obtained from the General Statistics interface, presented in Figure 7, and in the Overall Search Results interface, illustrated in Figure 8.

General Statistics regarding the Dataset
Biblioteca Nacional (PT)

Time Span	
Start Date/Time of Log:	1st December 2008 23:59:48
End Date/Time of Log:	29th January 2009 23:59:44
Days in Log:	59
Queries	
Queries:	2387075
Average queries per day:	40458
Average queries per week:	283206
Average queries per month (30 days):	1213740
Distinct queries:	364882
% of distinct queries:	15.3%
Terms	
Terms in queries (including stop words):	9784895
Terms in queries (excluding stop words):	7007165
Distinct terms (including stop words):	165496
Distinct terms (excluding stop words):	143335
% of distinct terms (including stop words):	1.7%
% of distinct terms (excluding stop words):	2.0%
Stop words:	2777730
% of stop words:	28.4%
% of non stop words:	71.6%
Average terms per query (including stop words):	4.1
Average terms per query (excluding stop words):	2.9
IP addresses and Sessions	
Distinct IP addresses:	50473
Distinct sessions registered in the log:	511851
Queries covered by the registered sessions:	2282519
Queries not covered by the registered sessions:	104556
% of queries covered by the registered sessions:	95.6%
% of queries not covered by the registered sessions:	4.4%
Average queries per IP address:	47.3
Average queries per session:	4.5
Locations and Languages	
Countries where queries were performed:	111
Cities where queries were performed:	2818
Languages (of countries where queries were performed):	66

Figure 7. General Statistics for the *Biblioteca Nacional*'s log.

5.1.1 Interpretation of the General Statistics interface

The statistics on this interface are a general summary or overview of the entire log, with no filters or constraints applied.

Providing answers to "when?" type of questions:

The Search Log Analysis application presents the exact dates and times for the start and end of the log. This information is important in order to allow the user to identify the exact timing, relevancy of duration and context of when the searches were carried out and the data was captured.

Providing answers to "how many?" type of questions:

The General Statistics interface is mostly made of statistics representing frequencies and averages answering several "how many?" type of questions. Among several other metrics, the interface informs the user about the total number of days contained in the log, the total number of queries performed during that period and the average number of queries performed per day, per week and per month.

The "Average Queries per Day", "Average Queries per Week" and Average Queries per Month" statistics supply simple but useful information as they provide the user with an understanding of how much the Search Engine or the database is being used on a daily, weekly and monthly basis. This information is useful to the management team of *Biblioteca Nacional*. For instance, it would add value to its Director as it provides insight regarding the number of searches carried out and the levels of usage of the library. It would be of interest to the Marketing Director to enable comparisons of actual data versus usage targets and assess deviations and required actions to achieve the objectives.

The application makes a distinction between "queries" and "distinct queries" in order to give users an understanding of how repetitive or unique the queries being performed are.

The Search Log Analysis application calculates and displays the total number of terms searched, including and excluding stop words. It is useful to know the difference between both in order to understand the percentage of stop words and the percentage of non stop words. In this example of *Biblioteca Nacional*, stop words represent nearly 30% of all words used. After excluding such relatively unimportant words one can get a better view of the number of key words entered in the searches. The Average Terms per Query statistic, particularly the one excluding stop words, provides users with an understanding of how long or how complex queries tend to be on average.

Similarly, the Search Log Analysis application calculates and displays the total number of distinct terms, including and excluding stop words, as well as the percentage of distinct terms. In the example of *Biblioteca Nacional*'s log only about 2% of terms used were distinct, which quickly shows that the large majority of the terms searched are repeated and non-distinct (around 98%). In generic terms, this statistic gives the user the information regarding how varied or repeated the terms searched tend to be.

As seen in figure 7, the Search Log Analysis application calculates and displays the total number of distinct IP addresses and the total number of distinct sessions registered in the log. This information is useful in order to understand how many different computers and how many different sessions generated the queries. The application also calculates the average number of queries per session and per IP address and informs the user about the total number of different computers (IP addresses), countries, cities and languages from where queries were performed. Some of these statistics nearly relates to "where?" type of questions. This information can provide the users of the Search Log Analysis application, namely the management of *Biblioteca Nacional*, with a view of how popular or how spread the service is across the globe.

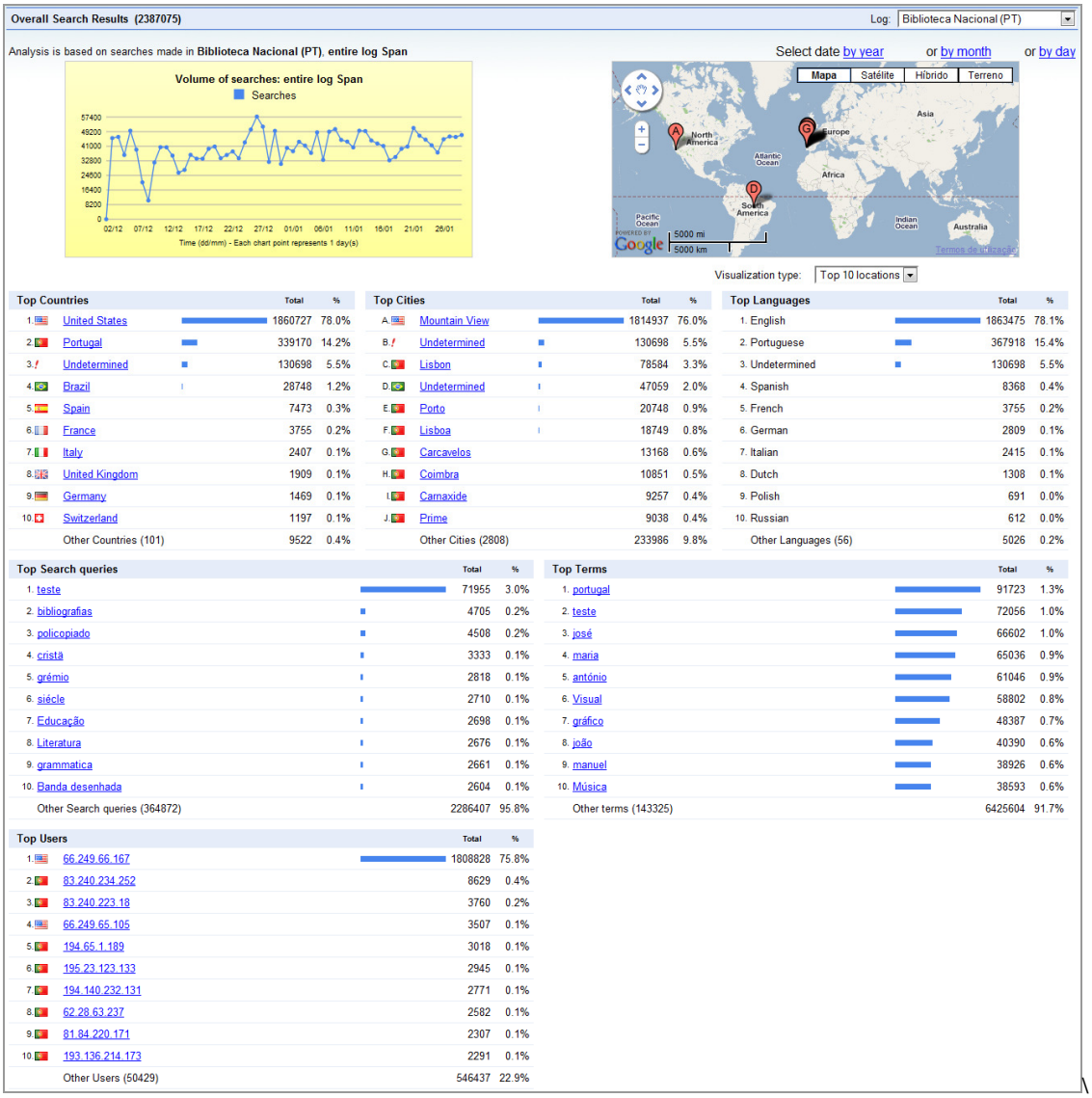

Figure 8. Overall Search Results output interface for the *Biblioteca Nacional*'s log (with no filters applied).

5.1.2 Interpretation of the Overall Search Results interface

Providing answers to "where?" type of questions

The Search Log Analysis application provides a ranking of the top 10 countries and the top 10 cities where most queries have been performed, as a whole for the entire log or in relation to the specific search entered by the user.

This is useful information to anyone analyzing patterns of user behavior and geographical spread or to anyone responsible for the log or for managing the company that produces the searches as it provides insight regarding the main locations on the planet from where the searches have been carried out. Any Sales Director or any Marketing Director will want to know where most of their customers are, which markets are more important and which countries have got the largest potential for customer growth.

Looking at both the Top Countries and the Top Cities tables in parallel, it is interesting to notice that out of all queries carried out at *Biblioteca Nacional* during the period analyzed, over 3/4 of all searches, the large majority, were performed in the United States, and particularly in Mountain View (97.5% of all of queries known to be performed in the United States were actually carried out from this single location in California). Moreover, examining also the table showing the Top IP Addresses used, it can be calculated that nearly all searches (97.2%) carried out in the United States originated from one single IP address in Mountain View. The second most frequent country is Portugal, where 14.2% of all queries were carried out. The application shows that the United States and Portugal combined together represent over 92% of all queries. This insight is valuable as it helps decision makers to focus on their main target markets. Looking at the table displaying the top cities, we can observe that 7 of the top 10 cities are Portuguese (although representing just 6.9% of all queries, a far lower proportion than the 76% achieved by Mountain View).

Searches have been carried out from a large number of cities (2818) but 2808 cities only account for 9.8% of all queries. The third most common country querying the database is actually an undetermined country, given that there wasn't a valid match for that IP address. Only 1.2% of all queries came from Brazil (4th in the ranking). Although queries were performed from several different countries (110 countries to be exact), the reality is that approximately 99% of all queries came from just these four countries.

By default, the map within the application pinpoints the location of the top 10 countries. However, through selecting "Dot Map" from the drop down list below the map, the Search Log Analysis application allows the user to also visualize a Dot Map highlighting the relevant locations reflecting activity. Figure 6 presents a dot map displayed by this application. The application has shown accurate results when plotting points on the map. As planned, it is observed that the Search Log Analysis application plots 3 different colors on the map, depending on the frequency queries performed. The locations of high activity are displayed in red, while the locations of lower activity are displayed in yellow and green.

In relation to the language of the queries used, over 3/4 of all queries were in English and 15.5% were in Portuguese. These two languages combined account for 93.5% of all queries. If we include the third most common language used (which is actually undetermined and different from the ones listed in the next sentence), then that equates to 99% of all queries. Spanish, French, German, Italian, Dutch, Polish and Russian combined together represent less than 1% of all queries.

Providing answers to "what?" type of questions

Looking at the searches carried out during the period analysed, it is clear that the searches tend to be largely varied and not very repetitive. The top 10 most common searches correspond to only 4.2% of all queries. This information may be useful to detect behavioural trends, growth rates, seasonality and identify the hot topics for the period being analysed. For example, that could guide the National Library in terms of enabling a better understanding of what most readers want to search and influence the library's marketing and investment decisions.

Providing answers to "when?" type of questions

Figure 8 shows the line chart created by the Search Log Analysis application when the log for *Biblioteca Nacional* is selected and no constraints in terms of time, queries, terms or space are placed. That chart plots the number of searches carried out for each of the 59 days available within the *Biblioteca Nacional*'s log, where each point represents one day. At a glance one can observe that the month of December is much more volatile in terms of the volume of searches than the month of January. The overall trend for the 2-month period is of a slight growth in the number of searches conducted.

The line chart presented by the Search Log Analysis application adds value because it provides quantitative answers to the question "how many queries?" and adds information regarding "when". It can be observed that the line chart does provide useful and relevant information to the user in a quick and visual manner.

Providing answers to "who?" type of questions

Over 3/4 of all searches were carried out from a single IP address in the United States. This is by far the most used IP address given than the second in the ranking of usage (located in Portugal) only managed to represent a mere 0.4% of all queries searched during the studied period. The importance of that single IP address in the United States is such that it actually corresponds to 98.3% of all 10 top IP addresses combined together.

The top 10 most used IP addresses correspond to 77.1% of all queries and cover just the United States and Portugal. The remaining IP addresses used (and they are over 50,000) represent 22.9% but they are spread by a large variety of IP addresses, each one representing less than 0.1% of all queries carried out.

6. Conclusion

Every day large volumes of search data are stored in the logs of search engines. On its own, this raw data is just a set of transaction records between users searching for information and the systems that handle and serve those requests. However, if properly analyzed this resource can prove to be very valuable, as it provides comprehensive details of how a search engine was used, namely answering questions such as what and when was searched, where the searches where conducted and by whom.

Whilst some applications have already been developed with the purpose of answering these and other questions, those applications are still rare, their development details and source code are not made available and they are restricted to work only with specific search logs. Acknowledging this limitation and realising the existence of a gap in this field of research, I hypothesized that an open-source Web application, based on a relational database, could be developed and used to perform spatio-temporal search log analysis. In addition, this application should provide useful and relevant information and should not be restricted to work with a specific log but, instead, be adaptable to a myriad of search logs.

The main aim of this work was to prove the hypothesis by designing, developing and testing such a Web application. Validation was performed by means of a characterization study over a search log from *Biblioteca Nacional de Portugal(National Library of Portugal)* with the objective to prove that the Search Log Analysis web application provides useful and relevant information. The framework used to prove the usefulness and relevancy of the information was the DIKUW (Data, Information, Knowledge, Understanding and Wisdom) hierarchy model, according to Ackoff.

The validation of the Search Log Analysis application has proved successful and the hypothesis has been confirmed as the tool has shown that large amounts of unstructured raw data can be converted into relevant, summarized and useful information that enables its users to answer "when", "how many", "where" and "what" and "who" questions. Table 2 summarizes the results.

Table 2. Summary of hypothesis validation results

Type of Question	Validation Status	Result
When?	The application is able to inform when the queries and the search terms where carried out (year, month, day and hour).	**PASS**
How Many?	The application is able to calculate and inform about several frequency metrics and averages, including the number of queries, distinct queries, terms, distinct terms, stop words, distinct IP addresses, sessions, countries, cities, languages, average queries per day, average queries per week, average queries per month, average terms per query, average queries per IP address, average queries per session, etc.	**PASS**
Where?	The application is able to inform where searches are carried out (top 10 countries and top 10 cities).	**PASS**
What?	The application is able to provide what specific searches and what specific terms have been most in use.	**PASS**
Who?	The application is able to identify who are the most frequent users, by IP address, querying the database.	**PASS**

Moreover, and referring to the DIKUW hierarchy model, the interpretation and application of the information provided by

the Search Log Analysis application helps users to answer some "how" questions, going beyond the information level, converting that insight into knowledge. The outputs provided by the Search Log Analysis application provide enough relevant and useful information to allow the users to make conclusions and decide what actions they may need to perform and how.

7. REFERENCES

[1] Wang, X., Zhai, C.: Learn from Web Search Logs to Organize Search Results. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (2007)

[2] Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query Recommendation using Query Logs in Search Engines. In International Workshop on Clustering Information over the Web (2004)

[3] Cui, H., Wen, J., Nie, W.: Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering. Volume 15, Issue 4. USA (2003)

[4] Baeza-Yates, R., Hurtado, C., Mendoza, M., Dupret, G.: Modeling User Search Behavior. In Proceedings of the Third American Web Congress. Washington DC, USA (2005)

[5] Xiao, X., Wang, L., Xie, X., Luo, Q.: Discovering Co-located Queries in Geographic Search Logs. In Proceedings of the first international workshop on Location and the web. ACM International Conference Proceeding Series, Volume 300 (2008)

[6] Sanderson, M., Kohler, J.: Analyzing geographic queries. In Proceedings of the Workshop on Geographic Information Retrieval. 27th Annual International ACM SIGIR Conference. Sheffield, UK (2004)

[7] Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of Geographic Queries in a Search Engine Log. In Proceedings of the first international workshop on Location and the web. In CM International Conference Proceeding Series. Volume 300 (2008)

[8] Zhuang, Z., Brunk, C., Giles, C.: Modelling and Visualizing Geo-Sensitive Queries Based on User Clicks. In Proceedings of the first international workshop on Location and the web. ACM International Conference Proceeding Series. Volume 300 (2008)

[9] Zhang, V., Rey, B., Stipp, E., Jones, J.: Geomodification in Query Rewriting. In Proceedings of the 3th ACM workshop on Geographical information retrieval (2006)

[10] Jansen, B., Spink, A.: How are we searching the world wide web? A comparison of nine search engine transaction logs. In Information Processing & Management: an International Journal, Volume 42, Issue 1. Tarrytown, NY, USA (2005)

[11] Pass, G., Chowdhury, A., Torgeson, C.: A Picture of Search. In Proceedings of the 1st international conference on Scalable information systems. ACM International Conference Proceeding Series, Volume 152. Hong Kong (2006)

[12] Rose, D., Levinson, D.: Understanding User Goals in Web Search. in Proceedings of the 13th international conference on World Wide Web. International World Wide Web Conference. New York, USA (2004)

[13] Jansen, B.: Search log analysis: what it is, what's been done, how to do it. In Library & Information Science Research, Volume 28, Issue 3 (2006)

[14] Zobel, J., Moffat, A., Sacks-Davis, R.: An Efficient Indexing Technique for Full-Text Database Systems. In Proceedings of 18th International Conference on Very Large Databases. Canada (1992)

[15] Beitzel, S., Jensen, E., Chowdhury, A. Grossman, D. Frieder, O.: Hourly Analysis of a Very Large Topically Categorized Web Query Log. Illinois Institute of Technology, Chicago, USA (2004)

[16] Silverstein, C., Henzinger, M., Marais, H., & Moricz, M.: Analysis of a very large Web search engine query log. In SIGIR Forum, Volume 33, Issue 1. New York, USA (1999)

[17] He, D., Goker, A., & Harper, D. J.: Combining evidence for automatic web session identification. Information Processing & Management (2002)

[18] Montgomery, A., & Faloutsos, C.: Identifying web browsing trends and patterns. IEEE Computer (2001)

[19] Chau. M., Fang, X., Sheng, O.: Analysis of the Query Logs of a Web Site Search Engine. In Journal of the American Society for Information Science and Technology Volume 53, Issue 13 (2005)

[20] Baeza-Yates, R., Calderón-Benavides, L., González-Caro, L.: The Intention Behind Web Queries. String Processing and Information Retrieval (2006)

[21] Lam, H., Russel, D., Tang, D., Munzner, T.: Session Viewer: Visual Exploratory Analysis of the Web Session Logs. In Visual Analytics Science and Technology IEEE Symposium (2007)

[22] Broder, A.: A taxonomy of web search. In ACM SIGIR Forum, Volume 36, Issue 2. New York, USA (2002)

[23] Nettleton, D., Calderón-Benavides, L., Baeza-Yates, R.: Analysis of Web Search Engine Query Sessions. ACM SIGKDD Workshop on Web Mining and Web Usage Analysis, WEBKDD 2006. Philadelphia, USA. (2006)

[24] Fujii, A.: Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval. In Proceeding of the 17th international conference on World Wide Web. China (2008)

[25] Lee, U., Liu, Z., Cho, J.: Automatic Identification of User Goals in Web Search. In Proceedings of the World Wide Web Conference. Chiba, Japan. (2005)

[26] Wolfram, D.: Search Engine Queries: An Analysis of Excite Data Set. School of Library and Information Science. University of Wisconsin – Milwaukee (1999)

[27] Ackoff, R. L.: From Data to Wisdom, Journal of Applies Systems Analysis, Volume 16, (1989)