Data Min Knowl Disc (2012) 24:663–696
DOI 10.1007/s10618-011-0228-8

Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction

Maristella Agosti · Franco Crivellari · Giorgio Maria Di Nunzio

Received: 8 January 2010 / Accepted: 21 June 2011 / Published online: 14 July 2011

Abstract In the last decade, the importance of analyzing information management systems logs has grown, because log data constitute a relevant aspect in evaluating the quality of such systems. A review of 10 years of research on log analysis is presented in this paper. About 50 papers and posters from five major conferences and about 30 related journal papers have been selected to trace the history of the state-of-the-art in this field. The paper presents an overview of two main themes: Web search engine log analysis and Digital Library System log analysis. The problem of the analysis of different sources of log data and the distribution of data are investigated.

Keywords Web log · Query log · Search log · User study

1 Introduction

The interaction between the user and an information access system can be analyzed and studied to gather user preferences and to "learn" what the user likes the most. This information can then be used to personalize the presentation of results. User preferences can be learned either explicitly, for example by asking the user to fill-in questionnaires, or implicitly, by studying the actions of the user which are recorded

Responsible editor: Myra Spiliopoulou, Bamshad Mobasher, Olfa Nasraoui, Osmar Zaiane.

M. Agosti · F. Crivellari · G. M. Di Nunzio (✉)
Department of Information Engineering, University of Padua, Via Gradenigo, 6/a, 35131 Padova, Italy
e-mail: dinunzio@dei.unipd.it

M. Agosti
e-mail: agosti@dei.unipd.it

F. Crivellari
e-mail: crive@dei.unipd.it

in the search log of a system. The second choice is certainly less intrusive, but it does require more effort to reconstruct each search session a user has made in order to learn his preferences.

Log is a concept commonly used in computer science; in fact, log data are collected by programs to make a permanent record of events during their usage. Log data can be used to study the usage of a specific application, and to better adapt it to the objectives the users were expecting to reach. The analysis of transaction logs for studying automatic information access systems has a long history, much longer than the World Wide Web as we know it today.

An important example of log analysis can be traced back to the beginning of the 1980s (Tolle 1983). From 1981 to 1983, the office of the Online Computer Library Center (OCLC) conducted a research study into online public access catalogs (OPACs) to study the use at that time of online catalogs by means of transaction log analysis and focus group interviews. Those analyses were carried out to determine to what extent current system features were used. The assumptions they made are still valid today: the design of the system can be compared with actual, recorded system usage. This comparison is essential if we are to improve, redesign and create better systems.

At that time logs were recorded on tapes, and transaction log tapes were different for each OPAC system. Therefore, breaking down, reformatting, parsing of transaction logs into user sessions and developing state codes was a major part of the analysis. Considerable effort was dedicated to the development of the methodology for the transaction log analysis (analysis still to be done at that time). The methodology proposed consisted of studying the probability of transitions from state to state (basically actions) starting from a zero order analysis, which consisted of only the frequency of occurrences of individual states considered independent, and then moving on to a higher order analysis to study the probability of going from an initial state to the destination state.

One particularly important message the authors expressed in the paper was the following:

> One method to ease the level of effort in the future would be to clearly define the requirements in the production of data log tapes before the system is put into operation.

This advice reminds us that log analysis requires clean and well-defined log records, in the sense that the structure has to be as simple as possible to make the work of the analyst easier and at the same time complex enough to capture all the possible aspects of the system under study. This means that the effective work on log analysis can start only when the transaction log data have been collected and prepared for the analysis. The collection and preparation activities that need to be conducted have been analysed and reported in Agosti and Di Nunzio (2007), where a general methodology for gathering and mining information from Web log files was proposed, together with a series of tools to retrieve, store and analyze the data extracted from log files. A review of the general work to be done for the collection, preparation and analysis to conduct Web search transaction log analysis was previously presented by Jansen (2006), where the reader was informed about the different aspects that need to be faced to be in the condition to start an effective analysis of the data.

In this paper, we build on the experience acquired in the collection and preparation of log data and concentrate on the work that can be conducted if the collections of log data have been prepared. We present an overview of the research on log analysis carried out over the past 10 years and identify two main different lines: Web search engines log analysis and Digital Library Systems log analysis. The work of this overview started by collecting papers and posters about log analysis from five major conferences: the International Conference on Research and Development in Information Retrieval (SIGIR), the European Conference on Information Retrieval (ECIR), the International World Wide Web Conference (WWW), the ACM/IEEE Joint Conference on Digital Libraries (JCDL), the European Conference on Digital Libraries (ECDL). We found about 50 papers and posters which could be easily classified as log studies (they all contain at least the word "log" in the title). In addition, for each author of these papers we looked for related published journal papers. At the end, we added about 30 more journal papers.

We decided to organise the presentation of our findings along the two themes of Web search engines (WSE) and Digital Library Systems (DLS), mainly for the following reasons:

– Both WSE and DLS can be used as information access systems accessible through the Web; however, the collections of documents the user can access are drastically different. In fact WSE retrieve documents that are in general Web pages, while DLS retrieve documents that have been chosen after a quality control performed by professionals.
– The quality of documents influences the quality of logs. This aspect has not been adequately taken into account in previous studies, especially in relation to data mining and knowledge discovery aspects. We believe that log analysis will benefit by making this difference explicit.

From the justified choice of addressing the log analysis study along the two main themes of Web search engine and Digital Library System log analysis it derives that the review of the papers of the last decade is divided into Sect. 2 and Sect. 3. Section 4 presents the problem of verifiability and repeatability of log analysis experiments and the problem of log collection distribution. In Sect. 5, we analyze the possible trends in future log analysis research. Lastly in Sect. 6 we give our final remarks.

2 Web search engine log analysis

Web Search Engines (WSE) deal with the representation, storage, organization of and access to information items which are essentially Web pages. Users seeking some information should be provided with an easy way to represent and access the information they need (Baeza-Yates and Ribeiro-Neto 1999). However, characterizing the user information need is no simple task, and this problem can roughly be divided into three aspects: how users make their requests to the search engine, how users interact with the search engine and how the search engine organizes the results.

In the following we will focus on each of the three aspects. In particular in Sect. 2.1 we present the problem of how to help users with the formulation of their information needs; in Sect. 2.2, we discuss the issue of how to interpret user interaction with the

search engine; Section 2.3 reviews the approaches of how to organize the results of a search engine.

2.1 Query suggestion, expansion, and classification

The transformation of the user's information need into a query is a challenging activity, particularly when the users need is vague, when there is a lack of knowledge about the collection and when the retrieval environment is unfamiliar. The vagueness of the information need is usually reflected in the very limited number of words that users write to express their needs. Web users typically submit very short queries to search engines, with the average length of Web queries being less than three words[1] (Levene 2010). Short queries usually lack sufficient words to cover useful search terms and thus negatively affect the performance of Web search: users queries are too short to contain a sufficient number of useful terms for discriminating amongst ambiguous documents. For this reason, query suggestion and query expansion techniques are implemented by Web search engines to extend the original query with new search terms to narrow the scope of the search. Query suggestion aims to explicitly suggest full queries that have been previously formulated by users so that query integrity and coherence are preserved in the suggested queries, while query expansion aims to implicitly expand the query without suggesting anything to the user. As query classification concerns, automatic approaches to develop Web taxonomies have been studied and proposed.

2.1.1 Query suggestion

Typical methods for query suggestion exploit query logs and document collections, by assuming that in the same period of time many users share the same or similar interests which can be expressed in different ways. In Chuang et al. (2000), the automatic construction of a thesaurus to provide efficient interactive search and/or term suggestion techniques is proposed. On the basis of the live thesaurus, a front-end search engine with the capability of automatic term suggestion and meta search is built. The thesaurus was constructed from query logs of two Taiwanese search engines: the Dreamer[2] and GAIS.[3] Different orders of analysis were carried out. The first-order analysis involved structuring subject taxonomies for building a classification scheme, and categorizing search terms from the Dreamer's log into certain subject categories in terms of users' possible intentions. The second-order analysis included deriving various statistics to describe users' information needs and seeking patterns from both the Dreamer and GAIS logs. The third-order analysis was to realize the limitations of the search term log analysis and apply the findings to information retrieval research.

In Pu et al. (2002) and Chuang and Chien (2003a), the authors further investigated experimental results and performance evaluation of their classification scheme. The experimental results demonstrated that the approach is scalable and adaptable

[1] http://www.keyworddiscovery.com/keyword-stats.html.

[2] http://www.dreamer.com.tw/.

[3] http://gais.cs.ccu.edu.tw/.

when used to categorize query terms into predefined subject taxonomy within the dynamic Web environment. The approach was designed to combine with the search process of real-world search engines to extract features from the retrieved highly ranked search result snippets for each query. A clustering algorithm for generating a multi-way-tree cluster hierarchy was proposed. The algorithm was a hierarchical agglomerative clustering algorithm, which generated a binary tree hierarchy at first, followed by a hierarchical cluster partitioning (Jain et al. 1999).

The use of terminological feedback mechanisms to offer search term suggestions is another possible solution. Here term suggestions are generated based on user relevance judgements of previously retrieved documents by interactive relevance feedback. Several Web search engines have begun offering short lists of search refinement suggestions to encourage the interactive narrowing of query result sets. Anick (2003) reports on the results of several log-based studies of user interaction with the AltaVista[4] Prisma assisted search tool. In particular, they study the degree and nature of user uptake of the feature, as well as the conditions and effectiveness of its use within information seeking sessions based on an analysis of anonymous user activity logs from the AltaVista search site. Using document clicks as a heuristic measure of reformulation effectiveness, it was found that feedback-based refinements, when applied, were as effective at locating relevant documents as the average manual reformulation. Nevertheless, the vast majority of reformulations were still done manually. This may have been for many reasons, such as users' habits, impatience with scanning term lists, lack of knowledge or understanding of the tool, or simply that it is difficult to capture all the possible refinement needs of a diverse set of users in 12 terms.

In Zhang and Nasraoui (2006, 2008), the authors propose a method for clustering similar queries by adding up the similarity values for many same query pairs of many query sessions, and by keeping a query's most similar queries in the final clusters. They used Chinese search engine logs[5] to evaluate the coverage and the recommendations that matched the queries of a session. In Zhang and Nasraoui (2008), the authors further combine the association or correlation-type information with the textual content between queries in order to improve coverage and compensate for the sparsity of the query terms. A soft relation matrix is built to store the relation between consecutive queries that occur within the same session. Queries that are submitted immediately one after the other receive a maximal relation value, while this relation value is dampened for queries that are farther apart from each other during the same search session.

2.1.2 Query expansion

Query expansion involves adding new words and phrases to the existing search terms to generate an expanded query. The expansion can be computed by finding relationships between query terms and document terms in terms of probabilistic correlations or association rules. It can also be approached by analyzing the implicit actions that a user performs during the search.

[4] http://www.altavista.com/.

[5] http://corp.sina.com.cn/.

In Cui et al. (2002, 2003), accumulated information on user interactions was exploited for adapting a search engine to the users. In particular, the authors presented an approach to find out what queries were used to retrieve what documents of the Encarta Web site,[6] and from that, to extract strong relationships between query terms and document terms and use them in query expansion. By exploiting correlations between terms in documents and user queries mined from user logs, the query expansion method can achieve significant improvements in retrieval effectiveness compared to other query expansion techniques. The central idea of the method is that if a set of documents is often selected for the same queries, then the terms in these documents are strongly related to the terms of the queries. Thus some probabilistic correlations between query terms and document terms can be established based on the query logs, and these probabilistic correlations can be used for selecting high-quality expansion terms from documents for new queries. The proposed log-based query expansion method has three other important properties:

– since the term correlations can be pre-computed offline, the initial retrieval phase is no longer needed;
– since query logs contain query sessions from different users, the term correlations can reflect the preference of most users;
– the term correlations may evolve along with the accumulation of user logs. The query expansion process can reflect updated users interests at a specific time.

Discovering related queries, or related query terms, can also be further used for query expansion or search optimization. Shi and Yang (2006, 2007) used an improved association rule mining model to mine related queries from query transactions in query logs. The model presented an algorithm that: first, segments the user sessions identified in query logs into query transactions; then it mines association rules of related queries using an improved association rule mining model. This mining model utilizes not only the co-occurrences between distinct queries but also the Levenshtein (1966) distance similarity between them.

Query expansion can also be approached by means of pseudo-relevance feedback (PRF) algorithms Baeza-Yates and Ribeiro-Neto (1999). PRF assumes top-ranked retrieved information sources are relevant, and takes terms from those sources and offers them to searchers as query refinement alternatives. PRF algorithms have more knowledge of term distribution statistics than searchers, can provide recommendations representative of the current content of highly-ranked sources, and can generate refinements for all queries for which there are search results. White et al. (2007) reports the results of a comparison of PRF and query log-based refinement (called Query Extension QE). Knowledge of when the techniques differ and when they are similar is vital for designing query suggestion algorithms that use multiple sources of evidence; this is a long-term direction for our continuing research in this area. The study showed that the source, the amount of feedback and the query type affect the similarity between QE and PRF. Conceivably, both techniques could be deployed in parallel and refinements offered based on query classification. For example, the techniques appear interchangeable for navigational queries, but complementary for informational queries, and PRF

[6] http://encarta.msn.com/.

is better able than QE to serve rare queries. In addition, when QE and PRF were least similar, queries seemed ambiguous.

A different concept of query expansion can be found in Parikh and Kapur (2006), where the authors present the problem of the vagueness given by short queries from a human being point of view: human beings do not naturally think in terms of queries, but rather in terms of natural concepts. For this reason, the authors propose an algorithm to generate "units" or concepts. Units are generated using an iterative statistical approach using canonicalized query logs. The unit generation algorithm takes as its input a consolidated query file, which has distinct queries and their frequencies over a specific time interval. In the first iteration, all single words in all queries are considered as units or concepts and their frequencies are stored. One of the important applications of units is the generation of query refinements. Query refinements are determined from what are called extensions and associations: an extension of a unit is a larger unit that contains all the words in the first unit, whereas an association of a unit is another unit with which the first unit appears often in queries.

Automatic query rewriting can help user web search, by augmenting a users query, or replacing the query with one likely to retrieve better results (Jones et al. 2006). One example of query-rewriting is spell-correction which means changing words to synonyms or other related terms. In this work, the authors show the opportunities for improving results in the case of the Japanese language which are greater than for languages with a single character set, since documents may be written in multiple character sets, and a user may express the same meaning using different character sets.

2.1.3 Cross-lingual query expansion

Cross-Lingual Information Retrieval deals with the problem of retrieving documents written in a language different from the one used in the query. There are many applications on the World-Wide Web that make use of this multilingual approach (Gao et al. 2007; Wang et al. 2006; Hu et al. 2008). In this context, query suggestion and expansion is performed between different languages too. The Cross-lingual query suggestion (CLQS) task is to determine one or several similar queries in the target language from the query log.

A method of calculating the similarity between source language query and the target language query was proposed by Gao et al. (2007, 2010). In addition to the translation information, the method exploited a wide spectrum of bilingual and monolingual information, such as term co-occurrences, and query logs with click-through data. A discriminative model is used to learn the cross-lingual query similarity based on a set of manually translated queries, and the model is trained by optimizing the cross-lingual similarity to best fit the monolingual similarity between one query and the translation of the other query. A one-month English query log MSN search engine[7] is used as the target language log, and a monolingual query suggestion system was built based upon it. A sample of French queries were selected randomly from a French

[7] http://www.microsoft.com/presspass/newsroom/msn/factsheet/searchanswers.mspx.

query log, and were manually translated into English by professional French-English translators. By leveraging additional resources such as parallel corpora, Web mining and log based monolingual query expansion, the final system was able to cover 42% of the relevant queries suggested by a CLQS system with precision as high as 79.6%.

Another approach to cross-lingual query similarity is by exploiting different types of monolingual and bilingual information as a training set for a discriminative learning algorithm. In Gao et al. (2010), the support vector machine (SVM) regression algorithm (Smola and Schölkopf 2004) is used to learn the cross-lingual term similarity function. The key to this approach is to learn a cross-lingual query similarity measure between the original query and the suggestion candidates. A discriminative model to determine such similarity by exploiting different types of monolingual and bilingual information.

In Wang et al. (2006) a Web-based approach for dealing with the translation of unknown query terms for cross-language information retrieval is presented. With the proposed term extraction and term translation methods, it is feasible to translate unknown terms and construct a bilingual lexicon for key terms extracted from documents. The authors present an extraction method that is a hybrid of the local maxima method and the PAT-tree-based method (Wu et al. 2000). First, they construct a PAT tree data structure for the corpus, in this case, a set of search-result pages are retrieved using the source term as the query. By utilizing the PAT tree, the association measurement of every character or word n-gram in the corpus can be efficiently calculated, and the local maxima algorithm can be applied to determine the most possible lexical boundaries to extract the terms.

Another view of the problem of CLQS was given by Hu et al. (2008). Millions of users across the world issue queries to a search engine in various languages daily, and they formulate queries and click on returned Web pages based on their language knowledge, which generates a large-scale and cross-lingual click-through data source. A method to generate query translations based on the analysis of click-through data was proposed. This method extracts query translation pairs from click-through data based on two assumptions. The first assumption is that there may exist some naming convention in URLs which specifies the language information of the corresponding pages. The second assumption is that the clicked URLs are relevant to the query. This latter assumption provides the connection between URLs and queries. Based on these two assumptions, the proposed method consisted of two stages: identifying bilingual URL pair patterns and mining query translation pairs. They experimented the method using click-through data collected from a commercial Web search engine for eight months of only those sessions containing English and Chinese queries.

2.1.4 Query classification

Manually collecting and organizing term vocabularies and thesaurus for Web applications is neither practical nor cost effective due to problems related to information updating, time consumption, and scalability, for this reason different automatic approaches to create Web classifications have been proposed.

In Chuang and Chien (2003b), Huang et al. (2004), and Hung and Chien (2007) three different automatic approaches of creation of Web taxonomies are presented.

In Chuang and Chien (2003b), a real-world query term filter for filtering out pornography-related terms for Web image search has been successfully developed. The logs from three different search engines in different time periods were collected and a two-level subject taxonomy with 14 major categories and 100 subcategories was constructed to express the popular search subject areas. About 20 thousand of high-frequency query terms were categorized properly by human experts and used as ground truth to evaluate the classification algorithm based on the term-frequency/document-frequency distribution of terms.

A Web-based approach to generate thematic metadata is presented in Huang et al. (2004). The main point of the proposed approach is to use Web search result snippets as the feature source, and use the structural information inherent in a thematic hierarchy to train a classifier. The authors developed a technique called Hier-Concept Query Formation Method which generates effective queries to send to search engines to acquire a suitable training corpus. A k-Nearest Neighbor algorithm (Jain et al. 1999) was used to cluster results and generate thematic metadata.

In Hung and Chien (2007), the authors proposed an approach that automatically finds training documents for text classification from the Web, requiring only user-defined categories and some relevant keywords. The greedy Expectation Maximization algorithm, was applied to determine the number of concepts in a category through clustering. The empirical results showed that the generated keywords were comparable to the keywords extracted from Yahoo!s subcategories in terms of classification accuracy.

The approach, presented in Chuang et al. (2000) and that has been analysed in Sect. 2.1.1, can be also considered as a query classification approach, since it supports query suggestion through the automatic construction of a thesaurus using query logs.

2.2 User behavior and interaction

Log studies can be used not only for helping users to reformulate their information need, but also for understanding the interaction and the usage of the information access system. For example, studying the disorientation of users when browsing a Web portal can be used to improve the structure and the presentation of the portal itself. Many Web sites have a hierarchical organization of content and the organization chosen by the Web designer may be quite different from the organization expected by visitors to the Web site.

2.2.1 Finding and re-finding results

In Srikant and Yang (2001) an algorithm to automatically find pages in a Web site whose location is different from where visitors expect to find them is proposed. The idea is that visitors will backtrack if they do not find the information where they expect it: the point from where they backtrack is the expected location for the page. This problem presents an important issue: for some Web sites there is a clear separation between content pages and navigation pages (for example Amazon); product pages on these Web sites are content pages, and category pages are navigation pages. Other

Web sites may not have a clear separation between content and navigation pages. For example, Yahoo! lists Web sites on the internal nodes of its hierarchy, not just on the leaf nodes. In this case, a time threshold for distinguishing whether or not a page is a target page should be used. The Web server log of the Wharton Business School, University of Pennsylvania[8] was used to evaluate the proposed algorithm.

If finding information is a challenge, another important problem concerns re-finding information a user already accessed at least once. A study of re-finding behavior through log analysis trying to glean from the data which queries were intended to re-find information rather than find new information requires. The work presented in Teevan et al. (2006) attempted to capture re-finding intent by looking for repeated clicks on the same search results in response to queries issued by the same user at different times. Using a log of the queries and result clicks issued by the anonymous users of 114 Web browsers over a period of 365 days, the authors explored issues of re-finding by considering all instances of the same query string that occurred within 30 min to be a single query. The result of the study showed that:

- 40% of all observed queries led to a click on a result that was also clicked during another query session by the same user. The query a person issued was a good indicator of whether the searcher was going to click on a previously viewed result or not.
- Approximately 71% of the queries that resulted in repeat clicks involved the same query string.
- Not all identical queries led to a repeat click, but 87% did. It was significantly less common for searches with the same query string to result in clicks on different results.

Following the work on re-finding similar information presented in Teevan et al. (2006), researchers tried to understand which search engine features help and which negatively impact the re-finding objective of the users (Teevan et al. 2007; Teevan 2008). Search engines are constantly attempting to improve results through the discovery of new resources and the creation of new ranking strategies. Queries where the user makes the same query and always clicks on one and only one result are assumed to have a navigational intent. Navigational queries, as defined above, tended to be for specific corporate Web sites, and were likely part of a daily routine or at least daily life. By far the largest category of professional queries contained searches for stores or businesses. While this benefits users who are looking for the best new information, the rank change of previously viewed search results can adversely impact those users attempting to re-find. When a previously clicked result changed position, users were less likely to re-click results. This suggests that changes to result ordering caused people to re-find less information and view more new information. Changing the rank of a previously clicked result appears to hinder re-finding, so click predictions should be used carefully by search engines to customize search results in a manner consistent with the search habits of the individual users.

In Smyth et al. (2004), Freyne et al. (2004), and Smyth and Balfe (2006), approaches which rely on query repetition are presented. These approaches are based on the

[8] http://www.wharton.upenn.edu/.

hypotheses that such repetition is accompanied by selection regularity; in other words, not only do searchers often user similar queries, they tend to select similar results in response to these queries. These works describe a collaborative approach to Web search that captures the search histories of communities of related users and reuses this information to re-rank search results to better reflect community preferences. The collaborative filtering method exploits a graded mapping between users and items. This relationship is captured as a hit matrix in which each element $H_{i,j}$ contains a value $v_{i,j}$ to indicate that a number of users have found page p_j relevant for query q_i.

2.2.2 *Topic shift*

Instead of re-finding the same information, the study of the temporal changes in popularity and uniqueness of topical categories is another important aspect. Understanding how queries change over time is critical for developing effective, efficient search services. Beitzel et al. (2004) put emphasis on changing query stream characteristics over this longitudinal time aspect of query logs. The work focused on Circadian changes in popularity and uniqueness of topical categories. A search log consisting of hundreds of millions of queries from a major commercial search service, the American Online search engine,[9] over a seven day period from December 2003 through January 2004 was examined. The result was that the average number of query repetitions in an hour does not change significantly on an hourly basis throughout the day. Most queries appear no more than several times per hour, and these queries consistently account for a large portion of the total query volume throughout the course of the day. The queries received during peak hours are more similar to each other than their non-peak hour counterparts.

In Beitzel et al. (2007a,b), the same authors investigated also the problem of classifying the long tail of the query stream, that is rare queries which are low in individual frequency but of collectively high volume that are not often matched by categorized topical lists. Tracking changes in the query stream tail provides insight into whether rare queries are changing similarly to popular queries. A system for automatic Web query classification that combines manually classified queries with the perceptron with margins algorithm (Krauth and Mezard 1987) was proposed.

Repetition and temporal changes can be studied also for cause-effect relation between queries and sequential patterns of repeated queries. Causal relation is also referred to as cause-and-effect relation, which is one of the basic relation types in ontology. If one query change causes another query change within a time period, we say that there is a causal relation between the two queries. The authors of Sun et al. (2007) study a new problem of mining causal relations between different queries based on the time series data extracted from query logs. If a query is submitted much more often than normally, it usually means that a specific event related to this query is happening. The dataset used was collected from Microsoft search engine as a daily aggregation of users submitting frequency for each query, from December 2003 to August 2005.

[9] http://www.aol.com/.

Learning algorithms can successfully exploit information encoded in temporal profiles to improve tasks such precision prediction and query triage (Jones and Diaz 2007). Since the quality of retrieval results is correlated with the distribution in time of the documents retrieved, one can identify temporally ambiguous queries as good candidates for disambiguation with a small time series interface. A relevance modeling (Lavrenko and Croft 2001) solution to this estimation problem was adopted in order to look at the temporal information each of the top N documents provide and weight this information according to the documents probability of relevance, $P(Q|D)$, that is the probability of a query Q given a document D.

2.2.3 Robots instead of humans

Another important aim of Web log researchers is to recognize humans and isolate automatic crawlers. In fact, a search engine log contains not only individual user transactions and not all logged users are humans; a client which is an agent rather than a human should be discarded too. A reliable method for excluding agents is to only draw on clients who have accepted cookies. However, agents frequently "assume the mask of a human" and accept cookies. To detect agents the Web analysis uses the following rule: if a number of requests submitted by a client is greater than a certain threshold the client is detected as an agent and is excluded. The criterion of agent detection is only probabilistically reliable.

In Buzikashvili (2006) a uniform method of agent detection, which creates equal conditions regardless of the observation period, was proposed. The uniform agents detection is based on the sliding temporal window technique by selecting a sliding window size, assigning a certain threshold to this window, sliding the window over a time series of the client transactions, and comparing number of client requests covered by the window.

Users who issue more than one query per session is an interesting point for studying human multitasking. During multitasking a user executes tasks, interrupting one and returning to the interrupted task later. A true multitasking search is rare and cannot be investigated manually. Buzikashvili (2006) elaborated an automatic procedure of task session detection which distinguish between sequential execution of several task sessions when the tasks are executed one-by-one and a parallel execution when a searcher switches between unfinished tasks. Log samples of two search engines, the Russian Yandex[10] (2005, one week's sample, 175,000 clients) and Excite[11] (2001, one day's sample, 305,000 clients) were used to test the following hypothesis:

1. How frequent are parallel task sessions?
2. How are temporal sessions distributed over a session width?

The analysis carried out showed that searchers follow the principle of least effort and select the cheapest tactics in any situation; more than 98% of temporal sessions are sessions of sequential execution; during multitasking a searcher uses only two tasks, and finally he frequently executes them in an "enveloped manner": he interrupts one

[10] http://www.yandex.ru/.

[11] http://www.excite.com/.

task session, starts and completes the second task session and returns to the unfinished task.

2.3 Organization and presentation of results

Helping users to formulate their queries and study the interaction with the system is as important as studying how relevant documents are presented and organized by the search engine. The utility of a search engine is affected by multiple factors. One of the primary factors is the soundness of the underlying retrieval model and ranking function. These ranking functions may incorporate user feedback stored in the log data to improve the quality of the results. In this section, we grouped the works in three categories: re-ranking the list of retrieved documents by means of user feedback; using contextual aspects to present the results; studying the type of a query in order to organize results differently.

2.3.1 Document re-ranking

In Miller et al. (2001), log data are used to incorporate user feedback in the adjacency matrix of the HITS algorithm (Kleinberg 1999); the authors propose an algorithm named Usage Weighted Input. This is a modification of the adjacency matrix which replaces the original matrix with a link matrix. The link matrix weights connections between nodes (i.e. pages) based on the usage data from Web server logs of traffic on the Web site, in particular by incrementing the strength of the link from node i to node j every time is recorded in the log the action "the user travels from i to j". Studying the transition from one page to another does not give any information about the relevance of the page itself. Assessing the relevance of a document is a complicated matter which involves the subjectivity of each individual who poses the query.

The relevance of one or more documents is usually exploited to re-rank or re-organize results according to user preferences. Relevance feedback can be very effective, but it relies on users assessing the relevance of documents and indicating to the system which documents contain relevant information. In real-life Web searches, users may be unwilling to browse Web pages to gauge their relevance. The alternative is the so-called implicit feedback, which means the observation of implicit user actions (e.g. the time spent on a Web page, the actions performed on a page, and so on). In White et al. (2002, 2005), researchers verify the hypothesis that implicit and explicit feedbacks are interchangeable as sources of relevance information for relevance feedback. Through developing a system that utilized each type, they were able to compare the two approaches from the user perspective and in terms of search effectiveness. An interface was designed to connect to the Google search engine and offer users a summary of the documents retrieved for investigating Web search behavior. The implicit system re-ranks the list when the user moves the mouse over a document title. Through these means the user no longer has to be concerned with marking a document as relevant, the system has mined their interaction and made an educated assumption based on this, under the assumption that viewing a document summary is an indication of user interest in the contents of the document.

How to organize and present search results is also a very important factor that can affect the utility of a search engine significantly. However, when the search results are diverse, usually due to ambiguity or multiple aspects of a topic, the ranked list presentation would not be effective; in such a case, it would be better to group the search results into clusters so that a user can easily navigate into a particularly interesting group. In Wang and Zhai (2007), a proposal for a strategy to partition search results is presented which imposes a user-oriented partitioning of the search results. The strategy consists of two steps. The first step involves learning interesting aspects of similar topics from search logs and then organizing search results based on these interesting aspects. For example, if the current query has occurred many times in the search logs, we can look at what kinds of pages are viewed by the users in the results and what kind of words are used together with such a query. The second step involves generating more meaningful cluster labels using past query words entered by users. Assuming that the past search logs can help us learn what specific aspects are interesting to users given the current query topic, it may be expected that those query words entered by users in the past that are associated with the current query can provide meaningful descriptions of the distinct aspects. Experiments on MSN search log data set released by the Microsoft Live Labs in 2006 showed that the proposed log-based method can consistently outperform cluster-based methods and improve over the ranking baseline, especially when the queries are difficult or the search results are diverse. Furthermore, the method can generate more meaningful aspect labels than the cluster labels generated based on search results when we cluster search results.

2.3.2 Studying contextual aspects

Capturing the context of a user's query from the previous queries and clicks in the same session may help our understanding of the user's information need. A context-aware approach to document re-ranking, query suggestion, and URL recommendation may improve users' search experience substantially, since users often raise multiple queries and conduct multiple rounds of interaction with a search engine for an information need. The whole search process can be modeled as a sequence of transitions between states.

Cao et al. (2009) proposed modeling query contexts by a variable length Hidden Markov Model (HMM) for a context-aware search: by letting q_t be the current query, the HMM can re-rank the search results by the posterior probability distribution $\mathrm{P}(s_t|q_t, O_{1\ldots t-1})$, where s_t is the current search intent hidden in the user's mind and $O_{1\ldots t-1}$ is the context of q_t, which is captured by the past queries $q_1 \ldots q_{t-1}$ as well as the clicks of those queries.

Clicks on a page can also be used as different source of information. For example, the ImpressionRank of a web page is the number of times users viewed the page while browsing search results. It captures the visibility of pages and sites in search engines and is thus an important measure, which is of interest to web site owners, competitors, market analysts, and end users. In Bar-Yossef and Gurevich (2008, 2009), an algorithm for estimating the ImpressionRank of a web page is proposed. This algorithm relies on access to three public data sources: the search engine, the query suggestion service of the search engine, and the web. In addition, the algorithm is local, uses modest

resources, and solves a novel variant of the keyword extraction problem: it finds the most popular search keywords that drive impressions of a page.

Temporal and geographical proximity between users and queries are investigated in Jones and Diaz (2007) and Jones et al. (2008). Temporal profiles provide a method for predicting relevant dates for a particular query. Learning algorithms can successfully exploit information encoded in these profiles to improve tasks such precision prediction and query triage. In Jones and Diaz (2007), the authors study the correlation of the quality of retrieval results with the distribution in time of the documents retrieved. The temporal information each of the top N documents provide and weight the temporal information according to the documents probability of relevance. The first documents retrieved are viewed as a proxy for the set of relevant documents, and weight each by the estimated relevance. By examining aggregated logs of queries containing place-names, and looking at the distances from IP-location and reformulated query-location, insights into user distance preferences in web search can be obtained. It is observed that queries are likely to be relatively close to the IP-location that issued the query. Therefore, the shape of the cumulative distribution for these distances could be used to inform priors on distances for retrieving documents when the query does not contain a location.

2.3.3 News, events, people

The type of query a user issues can be used to organize results differently. For example, news-related queries can be used for disambiguating user information needs (e.g. prompting the user with a link to an online news service), as well as for highly effective online news processing, including news clustering, summarization, and ranking.

The objective of Maslov et al. (2006) is a method for extracting queries related to recent, ongoing, or upcoming real-life events reflected in the news, or news-related queries. It aims at extracting queries related to real-life events from a general-purpose Web search engine log, using relative query frequencies and validating them against current news feeds. Using the Russian Yandex News service[12] they found that news-related queries have different length distribution compared to general Web queries. News-related queries are not simply longer: they are very condensed event descriptors, often tying together important aspects of an event (e.g. location, date, actors, or type of event). This feature is illustrated by three pairs of sample queries related to three events, that are: President Putins press conference, Oscar nominees announced, and computer virus warning.

Together with news-related queries, Named Entity (NE) recognition is another important research area that concerns search log data. Query logs are a very useful resource for this task, and present unique advantages over other knowledge resources, such as knowledge bases created by human labor (for example, Wikipedia) or unanalyzed Web data. Most importantly, query logs directly reflect people search interests. Sekine and Suzuki (2007) focuses on NE categories like people, location, and organization, because it is well-known that such entities account for a large portion of

[12] http://news.yandex.ru/.

queries. A method for building up ontological knowledge about semantic categories and their properties using search query logs is presented.

3 Digital library system log analysis

Digital Libraries (DLs) differ from the Web in many ways. Firstly, DL collections are explicitly organized, managed, described, and preserved by expert librarians, archivists and museum keepers, that is, by the professionals that are knowledgeable of the specific area of interest of the library collections. In the context of digital library systems, the term "library" identifies all types of memory institutions. Secondly, Web sites and Web search engines assume very little about the users, tasks, and data they deal with. DLs normally have much more knowledge of their users and tasks since they are built to satisfy the specific needs of interested communities. Lastly, the digital objects in DL collections tend to be much more structured than the information presented in the Web. Understanding the information behavior of digital library users is central to creating useful, and usable, digital libraries. One particularly fruitful area of research involves studying how users interact with the current library interface.

Present digital library systems are complex software systems, often based on a service-oriented architecture, able to manage complex and diversified collections of digital objects. One significant aspect that still relates present systems to the old ones is that the representation of the content of the digital objects that constitute the collection of interest is done by professionals. This means that the management of metadata can be based on the use of authority control rules in describing author, place names and other relevant catalogue data. A digital library system can exploit authority data that keep lists of preferred or accepted forms of names and all other relevant headings. This makes the digital library systems and search engines significantly different. In fact a search engine often becomes a specific component of a digital library system, when the digital library system faces the management and search of digital objects by content in the same manner as information retrieval systems and search engines (Agosti 2008a). In all other types of searches, either the digital library system makes use of authority data to respond to final users in a more consistent and coherent way through a search system that is a sort of a new generation of Online Public Access Catalogue (OPAC) system, or the system supports the full content search with a service that gives the final users the facilities of a search engine. For all the different categories of users of a digital library system, the quality of services and documents the digital library supplies are very important. Log data constitute a relevant aspect in the evaluation process of the quality of a digital library system and of the quality of interoperability of digital library services (Agosti 2008b; Koch et al. 2004).

The following sections analyze important research fields in the area of digital library systems log analysis: Section 3.1 faces the problem of combining different sources of log data; Section 3.2 discusses the problem of the logging format for a digital library system; and Sect. 3.3 presents the research about digital library systems usage.

3.1 Combining different streams of log data

The combination of implicitly and explicitly collected data improves understanding of behavior with respect to the understanding that can be gained by analyzing the sets of data separately. In particular for digital libraries, where the evaluation of the different services is difficult if logs are used alone, the combined sets of data provide the opportunity of reaching insights towards user personalization of digital library services.

Search query logs or Web logs alone give only a partial view of the stream of information that users produce. Teevan et al. (2006, 2007) show how to combine two different streams of data, search query logs and click-streams, in order to analyze re-finding behavior of a group of users under observation for a period of one year. Moreover, log analysis can be supported and validated by user studies which are a valuable method for understanding user behavior in different situations. User studies require a significant amount of time and effort, so an accurate design of the process has to be carried out. In general, user studies and logs are used in a separate way, since they are adopted with different aims in mind. Ingwersen and Järvelin (2005) report that it seems more scientifically informative to combine logs together with observation in naturalistic settings. Pharo and Järvelin (2004) suggest systematic use of the triangulation of different data collection techniques as a general approach in order to get better knowledge of the Web information search process.

Among all the papers reviewed in this work, only a few explicitly claim the need of a combination of different sources of evidence: Mahoui and Cunningham (2000) write that further studies, including interview-based examinations of small groups of users, are required to complete their work; Anick (2003) needs laboratory tests or questionnaires to answer more complex questions. Koch et al. (2004) found that a thorough log analysis can provide deeper understanding of the services and they may offer useful hypotheses for advanced user studies. Only Assadi et al. (2003) presents an approach which relies on a combination of three methods: user-centric data analysis, online questionnaires and interviews.

In Kim et al. (2005) an implicit rating is captured from a user activity. Each user activity has a direction, where: "rating" means the user gives some feedback to the system; "perceiving" means the user does not give feedback to the system; and regarding intention, "implicit" means the user gives feedback implicitly while "explicit" means feedback is given explicitly. Implicit ratings from 22 students at both the Ph.D. and Masters level in Computer Science at Virginia Tech were collected. Each subject was asked to perform searches using 10 queries about their research field and allowed to browse the search results to find interesting documents. All subjects were required to register into the system and so provided explicit preferences. The result of the statistical test showed that implicit ratings are good information for studies on user analysis, personalization, collaborative filtering, and recommending. The effect of different types of rating data on the performance of user cluster mining was also tested and it was found that using proposed terms performs worst, because of the sensitive overlapping ratio of appearance. This test suggests that users activities of selecting something on the interface and extracting document topics of returned documents from searches with a document clustering algorithm represent user characteristics.

In Grimes et al. (2007), Google researchers present a work which shows that three different modes of collecting data, the field study, the instrumented user panel and the raw query log, provide complementary sources of data. The main claim of the authors is that fully understanding user satisfaction and user intent requires a depth of data unavailable in search query logs but possible to acquire from other sources of data, such as one-on-one studies or instrumented panels. The three types of data collection are defined in the following way:

- Lab studies: In a field or lab study, a researcher or observer watches while a user performs tasks on the search engine.
- Instrumented panels: A periodic study where users with some form of instrumentation or browser application agree to take part in observation over a short or long period of time.
- Aggregate query logs: Records stored by a search engine containing both the query issued by the user and potentially other data about actions such as results clicked on, other queries issued by that user, or additional metadata.

The query log is the least rich source of data for individual events, but has irreplaceable information for understanding the scope of resources that a search engine needs to provide for the user.

In Agosti et al. (2009), the authors present a study about the combination of implicitly and explicitly collected data. The proposed method was envisaged during the study of the Web portal of The European Library (TEL), which provides a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage in many different languages. The analysis shows that the combination of different sources improves understanding of behavior with respect to the understanding that can be gained by analyzing the sets of data separately. In particular, the combined sets provide the opportunity of reaching insights towards user personalization of Web searching services, and also make possible results that can be generalized and applied not only to the users that have participated in the lab study.

3.2 Logging format

Log analysis is a primary source of knowledge about how digital library patrons actually use DL systems and services and how systems behave while trying to support user information seeking activities. Log recording and analysis allows evaluation assessment and opens opportunities to improvements and enhanced new services. Indeed, the benefits of logging are numerous, including improving performance by recording effective evaluation data, helping in designing and testing of user interfaces, and better allocation of resources. With this concept in mind, it is also possible to think about new and different logging formats which reflect how a generic DL system behaves (Gonçalves et al. 2002b, 2003; Klas et al. 2006).

Digital library logging should offer much richer information and opportunities. For this reason Gonçalves et al. (2002b) proposes an XML-based standard digital library log format that captures a rich, detailed set of system and user behavior. The proposed standard can be implemented in a generic log component tool, which can be plugged into any digital library system to produce the specified format. A careful analysis of

common sets of problems with Web and DL logging scheme was raised. A summary of these problems is:

- Disorganization: system logs are in general poorly organized and structured.
- Complexity of analysis: lack of proper thought in recording the log information makes log analysis a hard problem.
- Incompleteness: important information that would be necessary for analysis is often omitted from some logs.
- Incompatibility: each system has its own log format, making it difficult to use the same tools to analyze logs from different systems for the same kind of study.
- Ambiguity: many log entries and their semantics are not properly and precisely specified in the log format itself, which could lead to ambiguity in analyzing them.

To be useful, the DL log format has to portray how a generic DL system behaves. In accordance with the 5S theory for digital libraries (Gonçalves and Fox 2002a), a new log structure was designed, organized and presented by Gonçalves et al. (2003). The objective of the work is the standardization of the log with a consequent standardization of log processing and analysis. The proposed log has been used in the Computing and Information Technology Interactive Digital Educational Library[13] (CITIDEL). The primary component of the log analysis tool is the log line parser that sends the content of each log line to an appropriate module which increments appropriate variables, populates files that are intermediate aggregate statistics of key log features, and performs a host of other required actions. The modules that are already developed track, browse and search requests for each resource, maintain a record of the number of accesses from each domain, keep statistics on the words used in all the search queries, record the number of hits and logons per day, month, year, etc., and keep track of the number of times that various tools and CITIDEL provided resources have been utilized.

Analysis of transaction logs is one evaluation method that has provided DL stakeholders with substantial input for making managerial decisions and establishing priorities, as well as indicating the need for system enhancements. However, the quantitative nature of this method is often criticized for its inability to provide in-depth information about user interactions with the DL being evaluated. The development of a standardized logging scheme would facilitate comparisons across DL evaluation activities and provide the means for highlighting critical events in user behavior and system performance. The logging scheme proposed by Klas et al. (2006) is framed within the DL evaluation activities of the DELOS Network of Excellence (Fuhr et al. 2007). By using this scheme, researchers will be able to extract re-usable data from the results of previous DL evaluations and identify useful benchmarks, allowing for more efficient and effective design of evaluation studies. The Daffodil[14] virtual DL was used as the platform for testing the proposed logging scheme. For searching, exploring and managing DL objects Daffodil provides information seeking patterns that can be customized by the user for searching over a federation of heterogeneous digital libraries.

[13] http://www.citidel.org/.

[14] http://www.daffodil.de/.

3.3 Digital library usage

Transaction log analysis, which examines information behavior through the search artifacts automatically recorded when a user interacts with a library search system, offers an unobtrusive means for finding out what users are doing in a digital library.

3.3.1 Online public access catalogues

Transaction log analysis has been extensively applied to the study of searching behavior in OPACs, in order to provide a descriptive summary of large numbers of user interactions, and as such it is a valuable research tool for investigating user search behavior. However, it is difficult to generalize for all users, since at all times the value of the technique lies in describing a particular system's user, searching specific collections.

In Mahoui and Cunningham (2000) and Jones et al. (2000), there is a proposal for the generalization of results by examining the logs for two digital libraries: the Computer Science Technical Reports[15] (CSTR) collection of the New Zealand Digital Library and the Computer Science Bibliographics[16] (CSBIB) collection. Results were focused on the analysis of query complexity: in both DLs over 80% of queries are simple search queries with three or fewer terms. Non-expert users tend to accept default settings no matter what these settings are.

Although log analysis cannot provide insight into the why of search behavior, the method presented in Mahoui and Cunningham (2001) and Jones et al. (2000) supports examination of very large numbers of search sessions and queries, on a scale that more qualitative studies cannot match. During the period in which the transaction logs were collected, the collection included more than 290,000 documents. The CSTR digital library also provides a full text index to computer science research material. At the time of the usage logging, the CSTR indexed nearly 46,000 technical reports, harvested from over 300 research institutions. CSBIB documents are primarily bibliographic records, rather than full text documents. The CSBIB collection at the time of the data logging included approximately 1,000,000 references. The first result from the analysis of these sessions is only about 6% of the total number of sessions started with a citation/document search query, that is, from the main search page for the digital library. One possible explanation is that many of these researchers used general purpose search engines to locate research papers more frequently than they used formal computing subject indexes. A large portion of sessions did not include a search query (46.69%).

A general study of the uses of a diversified population of remote users of a digital library is presented in Assadi et al. (2003). Here the population was made up of university researchers and postgraduate students, as well as high school teachers and students or individuals performing personal research. The study made use of innovative Internet traffic capture and analysis technologies in order to develop

[15] http://www.cnri.reston.va.us/cstr.html.

[16] http://www.nzdl.org/gsdl/collect/csbib/import/index.html.

a user-centered approach. In October 1999 the Association of Research Libraries[17] (ARL) initiated a program for research projects on statistics of electronic resources usage. The library community worked many years on the refinement of the standard services measurement and assessment guidelines to cover the electronic services. The proposed combination of approaches includes: transaction-based measures made by transaction logs, use-based measures on user activities and user satisfaction. After 6 months of traffic capture, a database containing 15,500 Web sessions and almost 1,300,000 Web pages that were visited by the 72 users of the panel was collected. Digital libraries attract a public who are not necessarily regular users, but who use the service for specific research purposes: in the interviews and through traffic analysis, it was found that digital holdings allow rapid and simple access to difficult-to-find reference documents in the context of specific research. This public seemed quite different from that of a classical library, and "professional" researchers were, comparatively-speaking, mostly absent from this group. The majority of the observed population was over forty, and for them digital libraries were principally a source of information for personal research.

Renardus[18] was a distributed Web-based service which provided integrated searching and browsing access to quality-controlled Web resources from major individual subject gateway services across Europe (Koch et al. 2004). The Renardus project did a limited human evaluation of the service because of the high cost of full usability lab studies. In this work, the authors explored to what degree a thorough log analysis capturing unsupervised usage could provide valuable insights and working hypotheses into good usage and usability studies. More than 2.3 million Renardus log entries were recorded and grouped into user sessions to study behavior log entries. Each entry was classified into one of eleven different activities offered by Renardus and these activities were then used to characterize user behavior, via a typology of usages and sequences. The most interesting finding was the clear dominance of browsing activities (80%). Possible reasons include the fact that 71% of the users reach browsing pages directly via search engines and the layout of the home page focuses on browsing. Users tend to stay in the same group of activities, whether it is browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the service. This work also raised some important points for future investigation like:

- what influences the different usage levels of browsing versus searching activities?
- to what degree is the actual design of the system influencing user behavior?
- which important changes in design are prompted by the results of such user and log studies? and
- how can we provide search strategy support and improve the support for systematic browsing of large subject structures?

[17] http://www.arl.org/.

[18] http://www.renardus.org/.

3.3.2 Advanced digital library systems

As for the Web, the new trends of DLS services involve multimedia documents and mobility. Video retrieval, audio retrieval and online book reading are some of the recent killer applications of today's digital library services.

In Hopfgartner et al. (2008) a study about video retrieval interfaces is presented, in particular simulating users interacting with a facet-based video retrieval interface by exploiting log files from a user-study. Most interactive video retrieval systems are benchmarked in laboratory based user experiments. However, to make a robust measurement, the evaluation must be based on a large user population, which is very expensive, and a possible alternative of evaluating such systems is the use of simulations. The proposed retrieval model was simple: users enter textual search queries in each facet and the system returns a list of shots which are represented by a keyframe in the result list of the facet. After performing the initial user study, researchers analyzed the resulting log files and extracted user behavior patterns, such as when did users start a search query and when did they mark results as relevant. In Hopfgartner (2008), the authors propose to further adapt the retrieval model based on the users interaction with the retrieval systems interface.

In the text retrieval domain, the approach of interpreting the users action as implicit indicator of relevance turned out to be an effective method to increase retrieval performance. In the video retrieval domain, however, rarely anything is known about which implicit feedback can be used as implicit indicators of relevance. Community based implicit feedback mined from the interactions of previous users to aid users in their search tasks was exploited. Christel et al. (2009) reports an advanced use of transaction logs for aligning the metadata of a digital video library of over 900 h of video and 18,000 stories from The HistoryMakers.[19] The objective was to apply automated techniques and generate time-aligned metadata for use in accessing video narratives. The study found that first time sessions were characterized by more text queries with less precision (larger answer sizes). Return users still spent a majority of their time playing video, but experimented more with advanced search options and the various views into video sets. Around 83% of self-described students never moved beyond the default view, using no optional views. In general, users then spent the majority of their time selecting video stories from the answer sets and playing them.

Online reading behavior has increasingly become an area of empirical and theoretical exploration by researchers from a wide range of disciplines, such as psychology, education, literacy studies, information science and computer science. Different disciplines have diverse ways of probing these questions. For these kinds of studies, it would be ideal to actually observe individual reading online, since any lab effort to do so would lose the realism of how people actually use web-based digital libraries. In Chen et al. (2009), the authors analyze and visualize Web log data by focusing on book-centered reading behavior with the actual logs from the online digital library of childrens books, the International Childrens Digital Library (ICDL).[20]

[19] http://www.thehistorymakers.com/.

[20] http://en.childrenslibrary.org/.

4 Verifiability and repeatability of log analysis experiments

Computer science researchers have been building a case for search log access to allow them to study and analyze new information retrieval algorithms via a common benchmark search log, as well as to learn about user information needs and query formulation approaches. Social scientists could investigate the use of language in queries as well as discrepancies between user interests as revealed by their queries versus user interests as revealed by face-to-face surveys. Advertisers could use the logs to understand how users navigate to their pages, gain a better understanding of their competitors, and improve keyword advertising campaigns (Korolova et al. 2009).

At the end of May 2009, a TrebleCLEF[21] workshop entitled "Query Log Analysis: From Research to Best Practice" was held at the British Computer Science Offices in London. The goal of the workshop was to provide a forum in which invited speakers from multiple disciplines could share and discuss their experiences with query and server log analysis (Clough and Berendt 2009). The workshop was seen as a starting point in addressing the wider goals of clarifying current research collating standardized procedures and resources commonly used. One of the main points of discussion was the availability and use of log data: considerations included how log files should be made publicly available to researchers, whether log data should be gathered for specific tasks, whether there is value in general log data, how additional information can be gathered and correlated with query log data. The current lack of recent and long-term data makes the verifiability and repeatability of experiments very limited. In Table 1, we summarized the source of log data of many of the works we reviewed. It is practically impossible to find two works on the same dataset unless by the same author, or where at least one of the authors worked for a commercial search engine company. This is not only a question of the same data source, but also a problem of using the same period of time for the analysis if the analysis has to be comparable with other works. The period of the recorded logs for each work (also shown in Table 1) clearly demonstrates this problem: some works record hours of logging activity, others days, weeks, and in some cases a year.

In Murray and Teevan (2007), the authors pose a number of unanswered questions regarding technology and policy. In particular, the problem of capturing information about user preferences and needs by public institutions and private companies is discussed. For example, retailers offer deep discounts to customers who allow their purchasing habits to be tracked. Meanwhile other institutions like public libraries make concerted efforts to purge their data caches frequently and to effectively clean away traces of patron behaviors. Search engines must figure out an appropriate balance within this space. The recorded search behaviors of individuals can compromise their identity, and institutions holding such data must be careful what they record and with whom they share it. However, while the release of logs might prove to be a major advance in the research field, their release to the broader public would be problematic from a privacy perspective (Korolova et al. 2009). Users communicate with a search engine in an uninhibited manner, leaving behind an electronic trail of confidential

[21] TrebleCLEF Coordination Action, Web site at the URL: http://www.treble-clef.eu/.

Table 1 Log data sources and time span of the logged data sets

Data source	Time span
Dreamer search engine (Chuang et al. 2000)	3 months in 1998
GAIS search engine (Chuang et al. 2000)	2 weeks in 1999
Computer Science Technical Reports (Mahoui and Cunningham 2000, 2001)	2 months in 1999
Computer Science Bibliographics (Mahoui and Cunningham 2000, 2001)	16 months from 1996
Wharton Business School (Srikant and Yang 2001)	6 days
Encarta Web site (Cui et al. 2002)	2 months
National Library of France (Assadi et al. 2003)	6 months
Renardus system (Koch et al. 2004)	More than a year
America Online (Beitzel et al. 2004)	7 days in 2003/2004
Yandex (Buzikashvili 2006; Maslov et al. 2006)	1 week in 2005
Excite (Buzikashvili 2006)	1 day 2001
SINA corp. (Zhang and Nasraoui 2006)	1 week/3 months in 2005
Tianwang search engine (Shi and Yang 2006)	4 months in 2003
Yahoo! (Teevan et al. 2006, 2007)	1 year
AOL query log (Carman et al. 2009)	3 months in 2006
Galaxy of Words (Sakai and Nogami 2009)	1 month in 2008
Google search interface (Kamvar et al. 2009)	35 days in 2008
Informedia Digital Video Library (Christel et al. 2009)	234 h of use in 2008
International Childrens Digital Library (Chen et al. 2009)	1 week in 2008
MSN search engine (Gao et al. 2007; White et al. 2007; Sun et al. 2007; Wang and Zhai 2007; Sekine and Suzuki 2007; Hu et al. 2008; Cao et al. 2009)	(a wide range of timespans)

thoughts and painfully identifiable information. For example, users search for their own name or the names of their friends, home address, their health concerns, as well as names of their family and friends. Users even enter their credit card number and social security number just to find out what information is present on the web. It would be irresponsible for a search engine company to release this data without modification. The open question to date is whether there even exists a way to publish search logs in an altered fashion that is simultaneously useful and private. Replacing usernames with random identifiers, as in the case of the AOL data release in 2006, is not sufficient to guarantee the privacy of innocent citizens.[22] The issue of confidentiality protection from the perspective of privacy preservation is an important matter that requires attention and it is out of the scope of this overview. We suggest the reader two recent surveys which delve deeper into the matter (Poblete et al. 2010; Cooper 2008).

A first attempt to release a collection of log data with the aim of verifiability and repeatability was done within the Cross-Language Evaluation Forum

[22] http://www.nytimes.com/2006/08/09/technology/09aol.html.

Table 2 Log file resources at LogCLEF

Year	Origin	Size	Type
2009	Tumba!	350,000 queries	Query log
2009	TEL	1,900,000 records	Query and activity log
2010	TEL	760,000 records	Query and activity log
2010	TEL	1.5 GB (zipped)	Web server log
2010	DBS	5 GB	Web server log
2011	TEL	950,000 records	Query and activity log
2011	Sogou	730 MB (zipped)	Query log

(Agosti et al. 2010) (CLEF)[23] in 2009 in a track named LogCLEF.[24] Since 2000, CLEF promotes research in multilingual information access by supporting the development of tools for testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and creating test-suites of reusable data which can be employed by system developers for benchmarking purpose. LogCLEF is an evaluation initiative for the analysis of queries and other logged activities used as an expression of user behavior (Mandl et al. 2010; Di Nunzio et al. 2011). Since 2009, the main goal of LogCLEF has been the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems. Another important long-term aim of this track is to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data. In the three years of LogCLEF editions, different data sets have been distributed to the participants: search engine query and server logs from the Portuguese search engine Tumba![25] and from the German EduServer[26] (Deutscher Bildungsserver: DBS); digital library systems query and server logs from The European Library[27] (TEL); and Web search engine query logs of the Sogou[28] Chinese search engine. Table 2 summarizes the log resources and the relative sizes.

In each edition of LogCLEF, participants are required to:

- Process the complete logs;
- Make publicly available any resources created based on these logs;
- Find out interesting issues about the user behavior as exhibited in the logs; and
- Submit results in a structured file.

The public distribution of the datasets as well as the results and the exchange of system components aim at creating of a community in order to advance the state of the art in

[23] http://www.clef-campaign.org/.

[24] http://www.promise-noe.eu/mining-user-preference.

[25] http://www.tumba.pt/(notworking).

[26] http://www.eduserver.de/.

[27] http://www.theeuropeanlibrary.org/.

[28] http://www.sogou.com/.

this research area. The LogCLEF 2011 Lab presents four different tasks which tackle some of the issues presented in this work:

- Language identification task: participants are required to recognize the actual language of the query submitted (see Sect. 2.1.3).
- Query classification: participants are required to annotate each query with a label which represents a category of interest (see Sect. 2.1.4).
- Success of a query: participants are required to study the trend of the success of a search. The success can be defined in terms of time spent on a page, number of clicked items, actions performed during the browsing of the result list (see Sect. 2.3.2).
- Query re-finding, when a user clicks an item following a search, and then later clicks on the same item via another search; Query refinement, when a user starts with a query and then the following queries in the same session are a generalization/specification/shift of the original one (see Sects. 2.2.1, 2.2.2).

5 Future trends in log analysis research

The more recent research on log analysis clearly reflects the new trends of the new technologies and applications such as social networks, mobile devices and search advertisement.

Social bookmarking systems like Delicious[29] and Bibsonomy[30] offer a rich source of information about popular Web pages and research articles. In these systems, users annotate resources with a small set of unstructured terms called tags that they choose from an unlimited vocabulary. Despite this freedom users generally choose to annotate Web pages with common terms from natural language that best describe the content, purpose or functionality of the Web site. Popular tags for a resource over a population of users can then be assumed representative of the consensus opinion. In Carman et al. (2009) the authors investigate the similarity between query logs and tag data from the perspective of the terms used to search for and annotate individual resources. The problem is whether the tags that people use to annotate resources are the same or similar to those used to search for resources. This question is critical for determining if and how social bookmarking data can be used to improve Web search. The study made use of the AOL query log, over a period of 3 months from March to May, 2006. A large amount of overlap was observed and demonstrated that query and tag vocabularies are indeed very similar. Future work in this context could be the analysis of the correlation between tag distribution or query distribution, and the content of the selected document.

Serendipitous search is a new research area which studies how users consider relevant information which was not part of the initial information need (Sakai and Nogami 2009). In 2008, a click-oriented Japanese search engine called Galaxies of Words[31] was released. This search engine utilizes the link structures of Wikipedia for

[29] http://delicious.com/.

[30] http://www.bibsonomy.org/.

[31] http://kotochu.fresheye.com/.

encouraging the user to change his information need and to perform repeated, serendipitous, exploratory searches. The aim of this search engine is to provide users with a lot of useful information that may not be relevant to their initial information need. The objective of the work was to analyze the query log for the entire month of October 2008 of Galaxies of Words to see how users move from one query to the next. The result of the analysis showed that users tend to make transitions within the same query type: from person names to person names, from place names to place names, and so on.

Search engine advertisers create an advertisement by providing a Web search engine the title, the description, and the URL of the advertisement to be displayed together with a set of keywords to bid for the advertisement. These keywords are usually related to the product or service the advertiser wishes to promote. Precisely identifying the keywords which can introduce high user interests and traffic is a hard task. Moreover top advertisers would be more interested in gaining advantage over the competitors and boost their own advertisements (Wang et al. 2009). In general search engines, if a query q leads users to two different advertiser Web sites w_a and w_b, then w_a and w_b are in competition with each other on query q. Thus, the search click-through data, which is the log data generated by the Web search engine, is a good resource for discovering the competitive relations.

As more mobile devices support rich access to the Web, we need to understand if and how user information needs and search patterns vary from each device. The study presented in Kamvar et al. (2009) compares user search behavior on three different types of devices: the computer, the iPhone and the conventional mobile phone. The objectives are: firstly, to provide a direct comparison of user search patterns on multiple search mediums; and secondly, to present an extensive and controlled comparison of user search patterns rather than the aggregate analyses of search queries presented in prior studies. The analysis was carried out on search patterns on three different Google search interfaces, during a month-long period in 2008. The main findings were: query length is very similar between computer and iPhone searches, but it is significantly shorter for mobile phone searches; the distribution of query categories was similar between iPhone and computer searches. Users search for local content within an application that can provide a richer experience (such as the iPhone maps application) if it is available. In the absence of a dedicated maps application on mobile phones, we see an increase in queries for local information, relative to computer-based searches. On a per search session basis, computer users had the greatest number of queries per session, followed by iPhone, and then conventional mobile phones. This may be indicative of the nature of the information needs exhibited by users on different devices (e.g. users may be more likely to search for quick factual information on mobile phones). For conventional mobile phone users, the difficulty of text entry may also discourage them from issuing more queries. Users on mobile phones may be more likely to browse multiple results in place of issuing query refinement. The biggest difference discovered between computer and iPhone users was that frequent computer-based searchers had a much higher rate of return than frequent iPhone or mobile phone searchers.

In Subasic and Berendt (2010), the authors present a new problem in the area of temporal text mining: the tracking of story evolution. A Web user who misses several days or who wants to gain an overview of major events and developments in a story

that lies in the past, is today faced with a situation that easily becomes unmanageable given the number of results returned by search engines. The same problem arises in other areas with high publication intensity and readers who aim to gain, refresh and/or extend overviews of topical developments—scholarly publications are a prime example. Regardless of the domain, users have two main goals when dealing with such corpora: story understanding and story search. The authors discuss the problem named ETP3 (Evolutionary Theme Pattern discovery, summary and exploration) which consists of: identifying topical sub-structure in a set of documents constrained by being about a common topic, showing how these substructures emerge, change, and disappear (and maybe re-appear) over time, and giving users intuitive interfaces for interactively exploring the topic landscape and at the same time the underlying documents. The STORIES algorithm presented in the paper is mainly based on frequencies of the co-occurrences of all pairs of content-bearing terms.

Map search engines, such as Google Maps,[32] Yahoo! Maps,[33] and Microsoft Live Maps,[34] allow users to explicitly specify a target geographic location, either in keywords or on the map, and to search businesses, people, and other information of that location. In Xiao et al. (2010), the authors analyze the kind of interaction with these maps and identify the following issues: map search users write shorter queries and modify queries more often in a session than general and mobile search users; map search users view fewer result pages than mobile search users; a small number of queries, target cities and user cities, have high search frequencies; the popular query topics in map search are different from those in general search; the popularity of target states and user states are positively correlated to the populations of states. The analysis of the distance between the query and the locations are computed by means of the nearest neighbor searching algorithm, in particular the Approximate Nearest Neighbor algorithm (Arya et al. 2009). One interesting finding is that 33.3% of queries have a target location within 50 km from the user location. Another finding is that the distribution of a query over target locations seems to follow the geographic location of the queried entity. For entities that are located in only a few places, for example, Disneyland, their query distributions over target locations are highly skewed. In addition, popular target cities differ in the set of hot queries.

6 Conclusions

The last 10 years of log analysis have raised important points of discussions and have indicated many interesting lines of research. Research in this area can be divided in two main categories: the analysis of Web search engine logs, and the analysis of Digital Library Systems logs. In the first case, research focuses on aspects regarding the charaterization of the user information need in terms of how the user formulates his request to the search engine, how the user interacts with the search engine and how the search engine organizes the results. In the second case, gaining understanding of

[32] http://maps.google.com/.

[33] http://maps.yahoo.com/.

[34] http://www.bing.com/maps/.

the behavior of digital library users is central to creating useful, and usable, digital libraries. The study of the usage, the logging format and the advanced services of a digital library are fruitful areas of research.

Log data alone give only a partial view of the usage of the system and the behavior of the user. Fully understanding user satisfaction and user intent requires a depth of data unavailable in search query logs but possible to acquire from other sources of data, such as one-on-one studies or instrumented panels. It is more scientifically informative to combine logs together with observation in naturalistic settings, and to use the triangulation of different data collection techniques as a general approach in order to get better knowledge of the Web information search process.

The current lack of recent and long-term data makes the verifiability and repeatability of experiments very limited. There is still a lot of work to do on the availability and use of log data: how log files should be made publicly available to researchers, whether log data should be gathered for specific tasks, whether there is value in general log data, how additional information can be gathered and correlated with query log data. A first attempt in this direction has been made since 2009 under the Cross-Language Evaluation Forum in a track named LogCLEF which aims to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data.

The new trends of log analysis research focus on new technologies and applications such as social networks, mobile devices and search advertisement. Social bookmarking systems offer a rich source of information as well as encyclopedic information like Wikipedia. This short-term way of tagging and judging documents can be studied to understand how users changes their information need. This change is very important, for example, for search engine advertisers and to create personalized advertisements through Web search engines. Mobile devices are giving new possibilities to researchers to study how the user interacts with the new technology and how they issue queries according to their geographical position. Advanced multimedia services for mobile devices such as video retrieval, audio retrieval, online book reading are some of the recent killer applications of today's digital library services. For this reason, we need to understand if and how users information needs and search patterns vary from each device.

Acknowledgments The paper reports on work which originated in the context of the DELOS Network of Excellence on Digital Libraries (http://www.delos.info/). The work has been partially supported by the TELplus Targeted Project for digital libraries, as part of the eContent*plus* Program of the European Commission (Contract ECP-2006-DILI-510003), by the TrebleCLEF Coordination Action, as part of the 7th Framework Program of the European Commission, Theme ICT-1-4-1 Digital libraries and technology-enhanced learning (Grant agreement: 215231), and by the PROMISE (http://www.promise-noe.eu/) network of excellence (contract n. 258191) project, as part of the 7th Framework Program of the European Commission.

References

Agosti M (ed) (2008a) Information access through search engines and digital libraries. Springer, Berlin

Agosti M (2008b) Log data in digital libraries. In: Agosti M, Esposito F, Thanos C (eds) Post-proceedings of the fourth Italian research conference on digital library systems (IRCDL 2008). DELOS: an Association for Digital Libraries, pp 115–122

Agosti M, Di Nunzio GM (2007) Gathering and mining information from web log files. In: Thanos C, Borri F, Candela L (eds) DELOS conference. Lecture notes in computer science, vol 4877. Springer, pp 104–113

Agosti M, Crivellari F, Di Nunzio GM (2009) A method for combining and analyzing implicit interaction data and explicit preferences of users. In: Doan BL, Jose JM, Melucci M, Tamine-Lechani L (eds) Proceedings of the workshop on contextual information access, seeking and retrieval evaluation (held in conjunction with the 31st European conference on information retrieval—ECIR 2009), pp 13–16

Agosti M, Ferro N, Peters C, de Rijke M, Smeaton AF (eds) (2010) Multilingual and multimodal information access evaluation, international conference of the cross-language evaluation forum, CLEF 2010, Padua, Italy, September 20–23, 2010. Proceedings, lecture notes in computer science, vol 6360. Springer

Anick PG (2003) Using terminological feedback for web search refinement: a log-based study. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, July 28–August 1, 2003, Toronto, Canada. ACM, pp 88–95

Arya S, Malamatos T, Mount DM (2009) Space-time tradeoffs for approximate nearest neighbor searching. JACM 57:1–54

Assadi H, Beauvisage T, Lupovici C, Cloarec T (2003) Users and uses of online digital libraries in France. In: Koch T, Sølvberg I (eds) Proceedings of the 7th European conference on research and advanced technology for digital libraries (ECDL 2003). Lecture notes in computer science, vol 2769. Springer, pp 1–12

Baeza-Yates RA, Ribeiro-Neto B (1999) Modern information retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston

Bar-Yossef Z, Gurevich M (2008) Mining search engine query logs via suggestion sampling. PVLDB 1(1):54–65

Bar-Yossef Z, Gurevich M (2009) Estimating the impressionrank of web pages. In: Quemada J, León G, Maarek YS, Nejdl W (eds) Proceedings of the 18th international conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009. ACM, pp 41–50

Beitzel SM, Jensen EC, Chowdhury A, Grossman DA, Frieder O (2004) Hourly analysis of a very large topically categorized web query log. In: Sanderson M, Järvelin K, Allan J, Bruza P (eds) Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 321–328

Beitzel SM, Jensen EC, Chowdhury A, Frieder O, Grossman DA (2007a) Temporal analysis of a very large topically categorized web query log. JASIST 58(2):166–178

Beitzel SM, Jensen EC, Lewis DD, Chowdhury A, Frieder O (2007b) Automatic classification of web queries using very large unlabeled query logs. ACM Trans Inf Syst 25(2):9

Buzikashvili N (2006) An exploratory web log study of multitasking. In: Efthimiadis EN, Dumais ST, Hawking D, Järvelin K (eds) SIGIR 2006: proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, Washington, USA, 2006. ACM, pp 623–624

Buzikashvili N (2007) Sliding window technique for the web log analysis. In: Williamson CL, Zurko ME, Patel-Schneider PF, Shenoy PJ (eds) Proceedings of the 16th international conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 2007. ACM, pp 1213–1214

Cao H, Jiang D, Pei J, Chen E, Li H (2009) Towards context-aware search by learning a very large variable length hidden markov model from search logs. In: Quemada J, León G, Maarek YS, Nejdl W (eds) Proceedings of the 18th international conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009. ACM, pp 191–200

Carman MJ, Baillie M, Gwadera R, Crestani F (2009) A statistical comparison of tag and query logs. In: Allan J, Aslam JA, Sanderson M, Zhai C, Zobel J (eds) Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2009, Boston, MA, USA, July 19–23, 2009. ACM, pp 123–130

Chen R, Rose A, Bederson BB (2009) How people read books online: mining and visualizing web logs for use information. In: Agosti M, Borbinha JL, Kapidakis S, Papatheodorou C, Tsakonas G (eds) Proceedings of the 13th European conference on Research and advanced technology for digital libraries. Lecture notes in computer science, vol 5714. Springer, pp 364–369

Christel MG, Maher B, Li H (2009) Analysis of transaction logs for insights into use of life oral histories. In: Heath F, Rice-Lively ML, Furuta R (eds) Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, JCDL '09. ACM, pp 371–372

Chuang SL, Chien LF (2003a) Automatic query taxonomy generation for information retrieval applications. Online Inf Rev 27(4):243–255
Chuang SL, Chien LF (2003b) Enriching web taxonomies through subject categorization of query terms from search engine logs. Decis Support Syst 35(1):113–127
Chuang SL, Pu HT, Lu WH, Chien LF (2000) Auto-construction of a live thesaurus from search term logs for interactive web search. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '00, pp 334–336
Clough P, Berendt B (2009) Report on the TrebleCLEF query log analysis workshop 2009. SIGIR Forum 43(2):71–77. http://doi.acm.org/10.1145/1670564.1670578
Cooper A (2008) A survey of query log privacy-enhancing techniques from a policy perspective. ACM TWEB 2:19:1–19:27. doi:10.1145/1409220.1409222
Cui H, Wen JR, Nie JY, Ma WY (2002) Probabilistic query expansion using query logs. In: Proceedings of the 11th international conference on World Wide Web, WWW 2002, WWW '02, pp 325–332
Cui H, Wen JR, Nie JY, Ma WY (2003) Query expansion by mining user logs. IEEE Trans Knowl Data Eng 15(4):829–839
Di Nunzio GM, Leveling J, Mandl T (2011) Multilingual log analysis: Logclef. In: Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11. Springer-Verlag, Berlin, Heidelberg, pp. 675–678
Freyne J, Smyth B, Coyle M, Balfe E, Briggs P (2004) Further experiments on collaborative ranking in community-based web search. Artif Intell Rev 21(3–4):229–252
Fuhr N, Tsakonas G, Aalberg T, Agosti M, Hansen P, Kapidakis S, Klas CP, Kovács L, Landoni M, Micsik A, Papatheodorou C, Peters C, Sølvberg I (2007) Evaluation of digital libraries. Int J Digit Libr 8(1):21–38
Gao W, Niu C, Nie JY, Zhou M, Hu J, Wong KF, Hon HW (2007) Cross-lingual query suggestion using query logs of different languages. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07. ACM, New York, pp 463–470. doi:10.1145/1277741.1277821
Gao W, Niu C, Nie JY, Zhou M, Wong KF, Hon HW (2010) Exploiting query logs for cross-lingual query suggestions. ACM Trans Inf Syst 28(2):1–33
Gonçalves MA, Fox EA (2002) 5sl: a language for declarative specification and generation of digital libraries. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, JCDL '02. ACM, pp 263–272
Gonçalves MA, Luo M, Shen R, Ali MF, Fox EA (2002) An xml log standard and tool for digital library logging analysis. In: Agosti M, Thanos C (eds) Proceedings of the 6th European conference on research and advanced technology for digital libraries. Lecture notes in computer science, vol 2458. Springer, pp 129–143
Gonçalves MA, Panchanathan G, Ravindranathan U, Krowne A, Fox EA, Jagodzinski F, Cassel LN (2003) The XML log standard for digital libraries: analysis, evolution, and deployment. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, JCDL '03. IEEE Computer Society, pp 312–314
Grimes C, Tang D, Russell DM (2007) Query logs alone are not enough. In: Amiaty E, Teevan J, Murray GC (eds) Query log analysis: social and technological challenges. A workshop at the 16th international World Wide Web Conference (WWW 2007)
Hopfgartner F (2008) Studying interaction methodologies in video retrieval. PVLDB 1(2):1604–1608
Hopfgartner F, Urruty T, Villa R, Gildea N, Jose JM (2008) Exploiting log files in video retrieval. In: Larsen RL, Paepcke A, Borbinha JL, Naaman M (eds) Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries. ACM, p 454
Hu R, Chen W, Bai P, Lu Y, Chen Z, Yang Q (2008) Web query translation via web log mining. In: Myaeng SH, Oard DW, Sebastiani F, Chua TS, Leong MK (eds) Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08. ACM, pp 749–750
Huang CC, Chuang SL, Chien LF (2004) Using a web-based categorization approach to generate thematic metadata from texts. ACM Trans Asian Lang Inf Process 3(3):190–212
Hung CM, Chien LF (2007) Web-based text classification in the absence of manually labeled training documents. JASIST 58(1):88–96
Ingwersen P, Järvelin K (2005) The turn. Springer, The Netherlands
Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Jansen BJ (2006) Search log analysis: what it is, what's been done, how to do it. Libr Inf Sci Res 28(3):407–432. doi:10.1016/j.lisr.2006.06.005. http://www.sciencedirect.com/science/article/B6W5R-4KSD84F-1/2/d131e3e1b135d3533787b301a941f893

Jones R, Diaz F (2007) Temporal profiles of queries. ACM Trans Inf Syst 25:14:1–14:31. doi:10.1145/1247715.1247720

Jones S, Cunningham SJ, McNab RJ, Boddie SJ (2000) A transaction log analysis of a digital library. Int J Digit Libr 3(2):152–169

Jones R, Bartz K, Subasic P, Rey B (2006) Automatically generating related queries in japanese. Lang Resour Eval 40(3–4):219–232

Jones R, Zhang WV, Rey B, Jhala P, Stipp E (2008) Geographic intention and modification in web search. Int J Geogr Inf Sci 22(3):229–246

Kamvar M, Kellar M, Patel R, Xu Y (2009) Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In: Quemada J, León G, Maarek YS, Nejdl W (eds) Proceedings of the 18th international conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009. ACM, pp 801–810

Kim S, Murthy U, Ahuja K, Vasile S, Fox EA (2005) Effectiveness of implicit rating data on characterizing users in complex information systems. In: Rauber A, Christodoulakis S, Tjoa AM (eds) Proceedings of Research and Advanced Technology for Digital Libraries, 9th European Conference. Lecture notes in computer science, vol 3652. Springer, pp 186–194

Klas CP, Albrechtsen H, Fuhr N, Hansen P, Kapidakis S, Kovács L, Kriewel S, Micsik A, Papatheodorou C, Tsakonas G, Jacob EK (2006) A logging scheme for comparative digital library evaluation. In: Gonzalo J, Thanos C, Verdejo MF, Carrasco RC (eds) Proceedings of Research and Advanced Technology for Digital Libraries, 10th European Conference. Lecture notes in computer science, vol 4172. Springer, pp 267–278

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. JACM 46(5):604–632. doi:10.1145/324133.324140

Koch T, Ardö A, Golub K (2004) Browsing and searching behavior in the renardus web service a study based on log analysis. In: Chen H, Wactlar HD, chih Chen C, Lim EP, Christel MG (eds) Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, JCDL '04. ACM, p 378

Korolova A, Kenthapadi K, Mishra N, Ntoulas A (2009) Releasing search queries and clicks privately. In: Quemada J, León G, Maarek YS, Nejdl W (eds) Proceedings of the 18th international conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009. ACM, pp 171–180

Krauth W, Mezard M (1987) Learning algorithms with optimal stability in neural networks. J Phys A Math Gen 20(11):L745. http://stacks.iop.org/0305-4470/20/i=11/a=013

Lavrenko V, Croft WB (2001) Relevance-based language models. In: Croft WB, Harper DJ, Kraft DH, Zobel J (eds) SIGIR 2001: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, September 9–13, 2001, New Orleans, LA, USA. ACM, pp 120–127

Levene M (2010) An introduction to search engines and web navigation, 2nd edn. Wiley, UK

Levenshtein V (1966) Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Doklady 10(8):707–710. http://sascha.geekheim.de/wp-content/uploads/2006/04/levenshtein.pdf

Mahoui M, Cunningham SJ (2000) A comparative transaction log analysis of two computing collections. In: Borbinha JL, Baker T (eds) Proceedings of research and advanced technology for digital libraries, 4th European conference. Lecture notes in computer science, vol 1923. Springer, pp 418–423

Mahoui M, Cunningham SJ (2001) Search behavior in a research-oriented digital library. In: Constantopoulos P, Sølvberg I (eds) Proceedings of research and advanced technology for digital libraries, 5th European conference. Lecture notes in computer science, vol 2163. Springer, pp 13–24

Mandl T, Agosti M, Di Nunzio G, Yeh A, Mani I, Doran C, Schulz JM (2010) LogCLEF 2009: the CLEF 2009 cross-language logfile analysis track overview. In: Peters C, Di Nunzio G, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) Multilingual information access evaluation. Vol I Text retrieval experiments: proceedings 10th workshop of the cross-language evaluation forum, CLEF 2009, Corfu, Greece. LNCS. Springer

Maslov M, Golovko A, Segalovich I, Braslavski P (2006) Extracting news-related queries from web query log. In: Carr L, Roure DD, Iyengar A, Goble CA, Dahlin M (eds) Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23–26, 2006. ACM, pp 931–932

Miller JC, Rae G, Schaefer F (2001) Modifications of kleinberg's hits algorithm using matrix exponentiation and weblog records. In: Croft WB, Harper DJ, Kraft DH, Zobel J (eds) SIGIR 2001: proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, September 9–13, 2001, New Orleans, LA, USA. ACM, pp 444–445

Murray GC, Teevan J (2007) Query log analysis: social and technological challenges. SIGIR Forum 41(2):112–120

Parikh J, Kapur S (2006) Unity: relevance feedback using user query logs. In: Efthimiadis EN, Dumais ST, Hawking D, Järvelin K (eds) SIGIR 2006: proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, WA, USA, 2006. ACM, pp 689–690

Pharo N, Järvelin K (2004) The SST method: a tool for analysing Web information search processes. Inf Process Manag 40(4):633–654

Poblete B, Spiliopoulou M, Baeza-Yates R (2010) Privacy-preserving query log mining for business confidentiality protection. ACM TWEB 4:10:1–10:26. doi:10.1145/1806916.1806919

Pu HT, Chuang SL, Yang C (2002) Subject categorization of query terms for exploring web users' search interests. JASIST 53(8):617–630

Sakai T, Nogami K (2009) Serendipitous search via Wikipedia: a query log analysis. In: Allan J, Aslam JA, Sanderson M, Zhai C, Zobel J (eds) Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2009, Boston, MA, USA, July 19–23, 2009. ACM, pp 780–781

Sekine S, Suzuki H (2007) Acquiring ontological knowledge from query logs. In: Williamson CL, Zurko ME, Patel-Schneider PF, Shenoy PJ (eds) Proceedings of the 16th international conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 2007. ACM, pp 1223–1224

Shi X, Yang CC (2006) Mining related queries from search engine query logs. In: Carr L, Roure DD, Iyengar A, Goble CA, Dahlin M (eds) Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23–26, 2006. ACM, pp 943–944

Shi X, Yang CC (2007) Mining related queries from web search engine query logs using an improved association rule mining model. JASIST 58(12):1871–1883

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222

Smyth B, Balfe E (2006) Anonymous personalization in collaborative web search. Inf Retr 9(2):165–190

Smyth B, Balfe E, Freyne J, Briggs P, Coyle M, Boydell O (2004) Exploiting query repetition and regularity in an adaptive community-based web search engine. User Model User-Adapt Interact 14(5):383–423

Srikant R, Yang Y (2001) Mining web logs to improve website organization. In: WWW 2001, pp 430–437

Subasic I, Berendt B (2010) Discovery of interactive graphs for understanding and searching time-indexed corpora. Knowl Inf Syst 23(3):293–319

Sun Y, Xie K, Liu N, Yan S, Zhang B, Chen Z (2007) Causal relation of queries from temporal logs. In: Williamson CL, Zurko ME, Patel-Schneider PF, Shenoy PJ (eds) Proceedings of the 16th international conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 2007. ACM, pp 1141–1142

Teevan J (2008) How people recall, recognize, and reuse search results. ACM TOIS 26:19:1–19:27. doi:10.1145/1402256.1402258

Teevan J, Adar E, Jones R, Potts MAS (2006) History repeats itself: repeat queries in yahoo's logs. In: Efthimiadis EN, Dumais ST, Hawking D, Järvelin K (eds) SIGIR 2006: proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, WA, USA, 2006. ACM, pp 703–704

Teevan J, Adar E, Jones R, Potts MAS (2007) Information re-retrieval: repeat queries in yahoo's logs. In: Kraaij W, de Vries AP, Clarke CLA, Fuhr N, Kando N (eds) SIGIR 2007: proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, The Netherlands, July 23–27, 2007. ACM, pp 151–158

Tolle J (1983) Transactional log analysis: online catalogs. In: Kuehn JJ (ed) Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '83. ACM, pp 147–160

Wang X, Zhai C (2007) Learn from web search logs to organize search results. In: Kraaij W, de Vries AP, Clarke CLA, Fuhr N, Kando N (eds) SIGIR 2007: proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, The Netherlands, July 23–27, 2007. ACM, pp 87–94

Wang JH, Teng JW, Lu WH, Chien LF (2006) Exploiting the web as the multilingual corpus for unknown query translation. JASIST 57(5):660–670

Wang G, Hu J, Zhu Y, Li H, Chen Z (2009) Competitive analysis from click-through log. In: Quemada J, León G, Maarek YS, Nejdl W (eds) Proceedings of the 18th international conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009. ACM, pp 1051–1052

White R, Ruthven I, Jose JM (2002) The use of implicit evidence for relevance feedback in web retrieval. In: Crestani F, Girolami M, van Rijsbergen CJ (eds) ECIR. Lecture notes in computer science, vol 2291. Springer, pp 93–109

White RW, Ruthven I, Jose JM, van Rijsbergen CJ (2005) Evaluating implicit feedback models using searcher simulations. ACM Trans Inf Syst 23(3):325–361

White RW, Clarke CLA, Cucerzan S (2007) Comparing query logs and pseudo-relevance feedback for web-search query refinement. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07. ACM, New York, pp 831–832. doi:10.1145/1277741.1277931

Wu LC, Horng JT, Liu BJ, Wang CY, Chen GD (2000) Indexing semistructured data using patricia tree. In: Ibrahim MT, Küng J, Revell N (eds) DEXA. Lecture notes in computer science, vol 1873. Springer, pp 859–868

Xiao X, Luo Q, Li Z, Xie X, Ma WY (2010) A large-scale study on map search logs. TWEB 4(3):8:1–8:33

Zhang Z, Nasraoui O (2006) Mining search engine query logs for query recommendation. In: Carr L, Roure DD, Iyengar A, Goble CA, Dahlin M (eds) Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23–26, 2006. ACM, pp 1039–1040

Zhang Z, Nasraoui O (2008) Mining search engine query logs for social filtering-based query recommendation. Appl Soft Comput 8(4):1326–1334