How does the Library Searcher behave?

A contrastive study of library search against ad-hoc search

Suzan Verberne, Max Hinne, Maarten van der Heijden,
Eduard Hoenkamp, Wessel Kraaij, Theo van der Weide

Information Foraging Lab, Radboud University Nijmegen

Abstract. We investigated the search behaviour of the library searcher using the log data from The European Library (TEL). We were especially interested in how this behaviour compares to the search behaviour of ad-hoc searchers, represented by log data in the MSN search engine. At first sight, the two data sets mainly differ in the topics of the queries entered and their multi-lingual vs. mono-lingual content.
When studying user behaviour, session information is very important: how does the user navigate through the engine's interface? We visualized the TEL users' interactions with the system by creating a transition network for the users' intra-session actions. In general, we think that research into user behaviour on the basis of search engine logs can be very informative for the evaluation of search engine interfaces.

1 Introduction

For the logCLEF 2010 workshop[1], the organization released a log file of user activities from the The European Library (TEL). TEL provides access to a number of national libraries in Europe through a search interface.[2]

The aim of the current paper is to model the search behaviour of the library searcher in contrast with the ad-hoc searcher. To this end, we collect a number of statistics of the TEL data and a log file from a general web search engine: the Microsoft 2006 RFP dataset that was distributed for the WSCD 2009 workshop[3]. We assume that the TEL data are representative for library search and the MS click data are representative for ad-hoc web search.

In Section 2, we present a number of statistics about both data sets. In Section 3, we discuss the main differences between them. In Section 4, we focus on one specific characteristic from the TEL data: the behaviour of the user within one session. In Section 5, we answer our main question: "How does the Library Searcher behave?"

2 Descriptions of the two data sets

The ad-hoc web search data consists of approximately 12 million clicks from US users entered into the Microsoft MSN search engine during the spring of 2006. For each query, the following details are available: a query ID, the query itself, the user session ID (more on session information later), a time-stamp, the URL of the clicked document, the rank of the URL and the number of results.

The library search data consists of a little under 2 million records collected between January 1st of 2007 and the 30th of June of 2008. For these records, the following information is included: a record ID, a user ID, an obfuscated user IP, a session ID, the chosen interface language, the query, the user's action (e.g. simple search or search within search), a document collection ID, the number of results, the rank of the clicked result, a search box ID, the URL of the object being viewed and a time-stamp. Some records in the data do not contain a query: these are interactions such as choosing a specific collection for the next search.

A number of key statistics of both data sets are shown in Table 1 and Table 2.

[1] http://www.uni-hildesheim.de/logclef/
[2] http://search.theeuropeanlibrary.org/
[3] http://research.microsoft.com/en-us/um/people/nickcr/wscd09/

Table 1. Basic statistics of the two data sets

TEL	MSN
1 866 330 interactions	12 251 067 clicks
1 345 508 queries issued	8 831 280 queries issued
220 409 unique queries	3 875 427 unique queries
2.2 average query length	2.5 average query length

Table 2. The most frequent queries with their frequency.

TEL		MSN	
Query	frequency	Query	frequency
mozart	16 605	google	123 392
van gogh	3 788	yahoo	111 724
meisje met de parel	3 481	myspace	79 955
harry potter	3 451	yahoo.com	60 494
einstein	1 925	myspace.com	49 327
pink floyd	1 664	ebay	48 691
rembrandt	1 618	mapquest	31 062
shakespeare	1 579	aol.com	27 345
nuremberg	1 552	google.com	25 304

3 Differences between the search engines

A quick inspection of the TEL data immediately shows that the search topics are different from the general-domain web search queries. Some example queries are 'il vocabolario degli accademici della crusca', 'twain mark', 'christie', 'maumet', 'vvedenskij', 'daylight' and 'dubois cg'. These examples confirm that the database is multi-lingual.

Some topics have a special form related to the metadata contained in the TEL index, for example '(subject all experiential learning)'. This is only a very small proportion of the queries: 0.13%. 6.8% of the queries contain boolean operators that combine atomic queries, for example '(title all keywords) and (creator all keywords)'. Note that all queries have been lowercased.

3.1 Language selection

In contrast to the MSN search engine, the TEL interface gives the user the option to select the interface language. Not all users use this option, resulting in relatively many number of records that use the default language setting, English. In total, the TEL click data contain 41 different interface languages. The most frequent are in shown Table 3.

Table 3. The most frequent languages selected in the TEL interface with their frequency (number of queries).

Language	frequency
English	1,600 514
French	60,108
Polish	36,438
German	27,384
Italian	24,068
Spanish	20,074

3.2 Sessions

A more striking difference between the data sets is the way the records are obtained. In the MSN data set, a record is stored each time a user clicks on a result in the result list. As a consequence, information about queries that did not result in a click is lost. On the other hand, in the TEL data set a record is stored each time the user interacts with the engine. Such an interaction can be the execution of a query (the most common interaction), but also viewing a document, saving a session or a switch in document collection.

An important observation is the difference in the grouping of records per session. The MSN data set considers a new session to start each time a user issues a new query. The TEL data set uses a more intuitive notion of sessions, where a session contains all subsequent user interactions within a certain time limit. Unfortunately, because of the way MSN sessions are stored, it is impossible to directly compare MSN and TEL sessions, for example with respect to query reformulation.

More exploratively, we considered the distribution of number of queries per query session as a basic descriptor of user search behaviour [1] for just the TEL data set. We plotted the cumulative distribution of the number of queries per session, as well as the unique number of queries per session, as shown in Figure 1. The plot shows that there is a significant number of sessions with a large number of queries, but that the amount of sessions with a large number of *unique* queries is much lower.

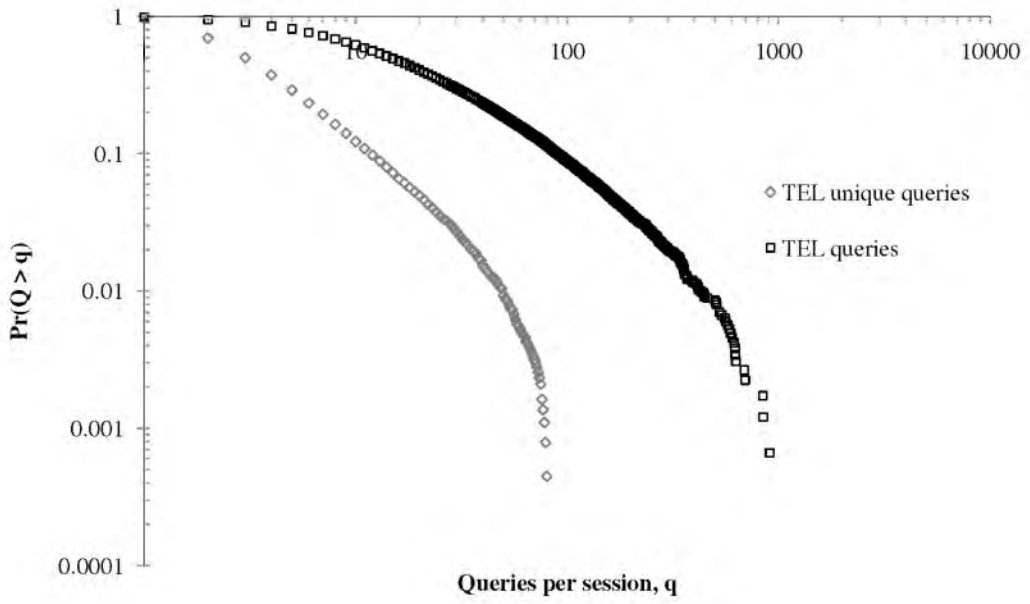

Fig. 1. The cumulative distribution $\Pr(Q > q)$ for the number of queries per session q.

We found that 16% of the queries are repeated three or more times by the same user in the same session. We suppose that the user retried the query with different settings in the TEL interface, or repeatedly interacted with the search engine for accessing a known document. We further investigate this phenomenon below.

4 Intra-session user behaviour in TEL

Ad-hoc web search generally consists for a large part of navigational and transactional queries [2]. For the library searcher, these types of queries are much less relevant. This warrants the question:

why are there so many repeating queries in the TEL log, if they are not used for navigation or transactional queries?

As discussed earlier, the TEL search engine provides the user with several options to refine a search. The language options have already been mentioned, but the search engine also offers several actions alternative to basic search, such as:

- to use either a simple or an advanced search form,
- to search within results,
- to continue searching with a specific record as a starting point,
- to start searching with a URL as query string,
- to view a short or a long description (title) of a result,
- to view a retrieved object in the interface of the library, or to see it online.

Some of these actions are not available in web search interfaces. To see how users work with this functionality, we proceeded as follows. Each time a query was repeated within a session, we kept track of the particular transition between the previously selected action and the newly selected action.

After normalizing so that for each action all possible transitions sum to 1, this resulted in a transition matrix representing the users' selection of actions, which is displayed as a directed graph in Figure 2. The sizes of the circles represent the overall behaviour of the searcher on the long term: what is the probability that at a given time, a user is involved in each action?

5 Conclusion: How does the library searcher behave?

We have investigated the search behaviour of the library searcher using the log data from The European Library (TEL). We were especially interested in how this behaviour compares to the search behaviour of ad-hoc searchers, represented by log data from the MSN search engine.

Although the average length of the queries does not differ much between the two search engines, the topics of the queries are very different. In web search there are many navigational and transactional queries, which can be seen from the most frequent queries "Google", "Yahoo" and "Myspace" in the MSN data. Among the most frequent queries in the TEL data are many named entities such as "mozart", "van gogh" and "meisje met de parel". Moreover, the TEL data contain queries in multiple languages.

When studying user behaviour, session information is very important: how does a searcher formulate and reformulate queries, and how does he navigate through the engine's interface? Unfortunately, the MSN data does not contain the type of session information that we need for such an analysis, which makes the two data collections not comparable in this respect.

Therefore, we investigated the interaction of the user with the search interface of TEL only. The finding that 16% of the queries are repeated three of more times shows that the library searcher makes an effort to get relevant results for his information need, often even without reformulating his query.

For research into user behaviour on the basis of click data, visualizing the users' interactions with the system can be very informative. We did this for the TEL data by creating a transition network for the users' intra-session actions. Naturally, the searcher ends his session with viewing the object he was searching for ('view full' or 'view brief'). But the graph also shows that users spend more time on simple search than on the other, more advanced, search options.

We think that research into user behaviour on the basis of search engine logs can be very informative for the evaluation of search engine interfaces.

References

1. Baeza-Yates, R., Hurtado, C., Mendoza, M., Dupret, G.: Modeling user search behavior. In: LA-WEB '05: Proceedings of the Third Latin American Web Congress. p. 242. IEEE Computer Society, Washington, DC, USA (2005)
2. Broder, A.: A taxonomy of web search. SIGIR FORUM 36(2), 3–10 (2002)

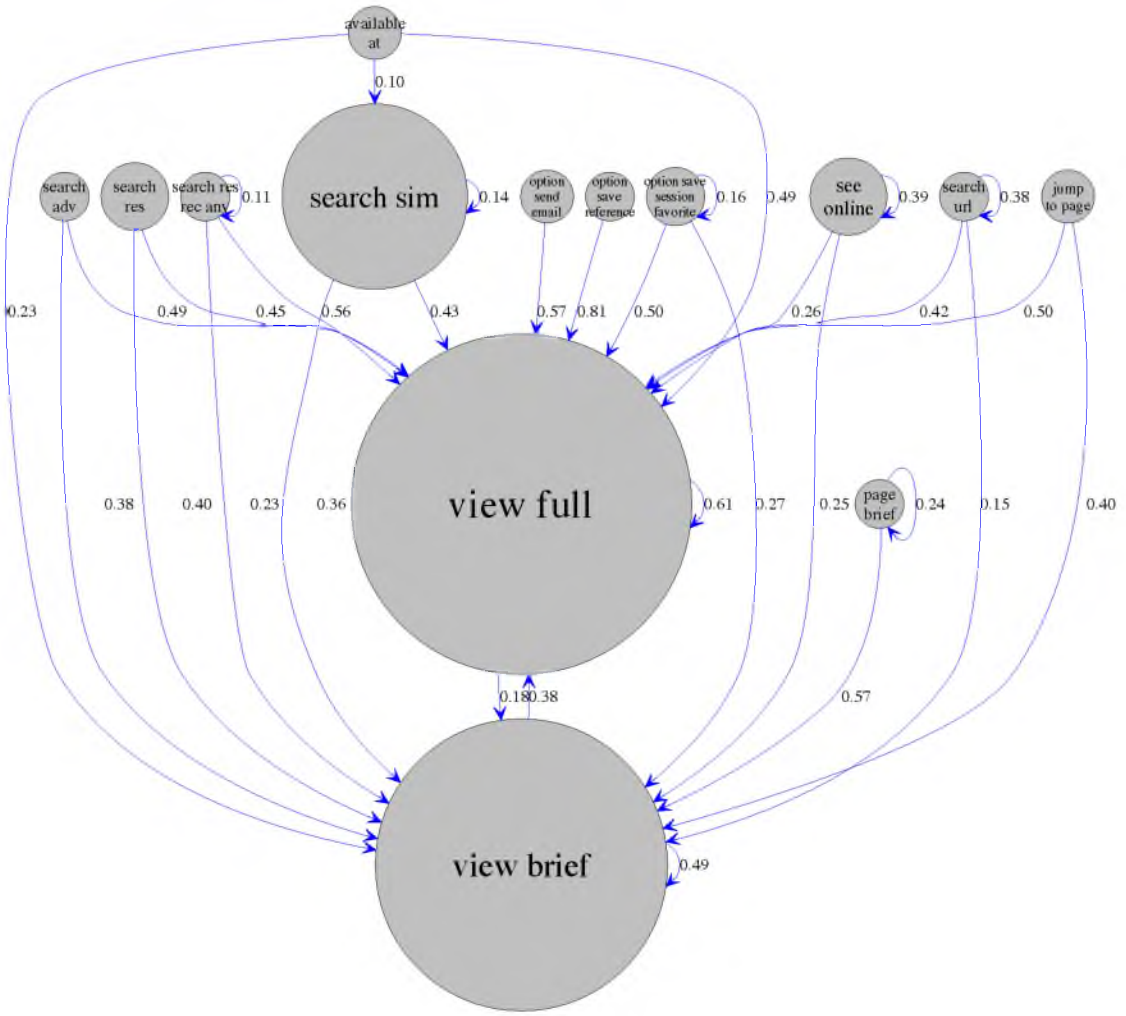

Fig. 2. The transition network of user search action selection. The sizes of the circles represent the probabilities for an asymptotic distribution, i.e. the overall behaviour of the searcher on the long term. Transitions with probability $p < 0.2$ have been excluded to avoid cluttering. Consequently, not all transitions sum to 1.