Standard comparators for Information Searching Behaviour in Search Engines

Nadine Hoechstoetter and Martina Koch

Institute for Decision Theory and Management Science, Universitaet Karlsruhe (TH), Germany, Correspondence to: Nadine Hoechstoetter, Sophienstr. 167, 76185 Karlsruhe, Germany, nsh@topicflux.de

Abstract

Search engines have become the most important medium for Internet users to find pages on the web. They help customers to decrease their information overload and enhance the sales of commercial web sites in different ways. For this reason, the exploration of and changes in (human) online searching behaviour has become a subject of particular importance. This paper will help search engine and web site administrators/developers to monitor online searching behaviour properly and to derive strategies from that.
We define standard comparators against which search engines could be measured and compared. They also reflect online searching behaviour of search engines users. Therefore, a literature overview of studies conducted in the last few years is given. Statistics used in different papers were compared to extract standard comparators for online searching behaviour. In the next step, search queries of four different search engines covering periods of between 10 and 13 months are compared using these comparators. For the first time different data sets over a long-time period are discussed and matched. We find out that some patterns stay stable in a number of different search engines, others were not analyzed adequately until now. We can't hold onto the assumption that web search has become more business driven. For this reason, we introduce the concept of evergreens in search queries. Implications are, e.g., that search engines should simplify web search interfaces for users since Boolean operators and special search features are commonly not used. We also present the evergreen topics in search queries.

Keywords: information searching behaviour, search engines, standard comparators

1. Introduction

Search engines are the most important tool with which people retrieve any data or pages on the Internet. Despite the convenience of these tools people still get lost in hyperspace [1] or are not able to use search engines appropriately [2]. To offer any recommendations or support in search engines, it is essential to gather sufficient knowledge of how and what people search for. Examples for such recommendations in search engines are related

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

search terms or correct search terms after a spell check. It is also interesting to get a deeper understanding of online users' preferences to develop user-friendly search engine interfaces.

'Search engine' is a very generalized expression to combine different kinds of tools to search for information on the web. General search engines such as Google, MSN, or Yahoo have their own index of documents and web pages, which is automatically generated by a web crawler. Meta search engines (e.g., Metager or Metaspinner) use a number of other search engines simultaneously. Directories like Dmoz are generated by manual input, and can be viewed as the web's 'Yellow pages'. Additionally, there are several other classes of search engines such as automated shop assistants or smart shoppers to help find pages in the web.

In this paper we focus on general and meta search engines, which are preferred for information retrieval by most Internet users. A more detailed overview of how search engines work, how to measure their quality, and which services are provided, are given in [3,4,5,6,7,8,9] and will not be repeated here.

Since 1995, online searching behaviour has been widely studied in two classes of studies. In one set of studies, subjects have been observed in laboratories or have to fill out questionnaires about their online searching processes or both [10,11,12,13,14,15,16,18,17,19,20]. In another set of studies, large search engine logs or a vast amount of search queries have been analyzed to find characteristic patterns (Table 1). Here, we focus on those analyses of search queries, which were collected without subject's awareness of observation. These studies give a more objective overview of how people interact with search engines in their life. Another argument against the use of controlled studies are the small sample sizes those studies are based on.

The numerous published studies are difficult to compare in a consistent manner. The studies are based on different data sets, lack consistent notation or do not provide adequate definitions. Our aim here is to provide an alternative analysis of those results using consistent statistical methods (see also [21]). We also seek to provide universal definitions of the most important measures we refer to as 'standard comparators'. In addition we introduce some standard expression notation, and find common statistical methods. This treatment of the previously published data renders it more meaningful as a set and allows us to derive conclusions more generally applicable than those of the particular studies.

In the next step, we evaluate these standard comparators on four largest obtainable data sets of search queries. These data sets are compared for the first time using the same statistics. We conclude that some measures remain stable over a long period of time. Others are also for a short period the same. In this case a small sample set of one day only is sufficient enough for basic analyses. Both represent core measures for online searching behaviour. More statistics and details are described in [21].

We also mention interesting additional data interpretation possibilities to visualize changes or stabilities in search queries. We can't sustain the thesis that web search topics have changed [22]. But we will demonstrate that there are 'evergreens' in search queries [23] and also in their topics. On the basis of our findings, we will give pointers for search engines to improve their indices of web pages and their user interfaces. Furthermore, we will drop some remarks on content optimization on web sites and portals. First, we will give an overview of the most important studies sorted by the search engine data sets. This will be followed by the introduction and the comparison of our data sets. We will conclude with a brief summary of our findings and conclusions.

2. State-of-the-Art and Standard Comparators for Measuring Searching Behaviour

In this chapter we summarize different studies and analyses of search queries in search engines, which were conducted and compare their results. Referring to those studies we will discuss standard comparators to measure information searching behaviour in search engines. We will also refer to special analyses conducted in different studies. Reviews of research in online searching behaviour in databases or on the web are provided in [16,21,24,25,26,27,28]. Here, we concentrate on studies concerning online searching behaviour using search engines. We do not consider specialized search engines of universities or other institutions as discussed in [29,30,31,32,33].

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

In all those studies various measures sets are given which describe those data sets. Here, we define the most cited measures, which become standard comparators for data sets. First we have to define the expressions search query, operator, search term, and search session. These are very important expressions to describe web search behaviour and are used with different meanings in the existing research.

We define a 'search query' to be the original character string formulated and typed into the search engine's interface by a search engine user to initiate an online web search.

Operators represent every command provided by a search engine, which helps users to refine their search queries. They are a part of search queries. Operators are, e.g., AND, OR, NOR, and NEAR (Boolean variables). Operators work only if they are used appropriately. Search engine users often utilize operators in a way not anticipated by the software. We will demonstrate this by discussing our results in section 4.

A 'search term' is a character string separated from other search strings by a blank or an operator. Another form of an operator, except those Boolean ones, is the usage of quotes. Terms enclosed by quotes indicate a phrase. This is again treated as a term by search engines which provide this function. We also handle phrases as terms.

Previous publications define search sessions in a number of ways. Some authors use a time-limited approach; others define a session as the entire series of queries submitted by a user during one interaction with the web search engine [34]. We suggest investigating the distribution of the lengths of user sessions to define an appropriate time frame to that purpose. We suggest using a time interval around the Median of session length.

In the next sections, we shortly introduce the different data sets studied in the past by mentioning some general aspects like the period of time during which the data was collected and we recall the references discussing different facets of online searching behaviour. We summarize the findings by repeating all standard comparators. At the end, we discuss obvious problems by comparing different data sets.

2.1. Data collection method

There are two main methods by which to gain an impression of searching behaviour in search engines. These are supervised and unsupervised methods. It is also possible to use a combination of those methods to gather different points of view on what people do with search engines.

The category of 'supervised methods' represents surveys and studies in laboratories since here, the objects feel observed and show a biased behaviour. There are several surveys and studies [10,11,12,13,14,15,18,17,19,20] concerning the interaction of users with search engines and their formulation of search queries. Search experiments also often contain an interview to gain knowledge of the demographical characteristics of participants. Most experiments are based on very few participants. Especially studies in laboratories which are very time intense and money consuming. Even though there are interesting psychological aspects of different stages in information gathering discussed, we don't feel that those studies are really representative of typical user behaviour. An extended overview is given in [21].

The examination of logged web search sessions is unsupervised, since persons are unaware of the fact that they are observed while searching the web. The analyses of transactional logs of search engines are one method of observing searching behaviour in search engines. A log is a text file consisting of a series of requests and provides different data fields. The most important entries are the timestamp measured in milliseconds from 1970, and the exact query terms submitted [34]. Details about browser and cookie information is also given.

With the domain and browser used, it is possible to identify users if they have cookies disabled. In Study [34] most users permitted cookies. With knowledge of this information, we can define sessions. The user begins a search session by sending his first request, e.g., a search query to the search engine. By analyzing sessions, inferences of how many search queries were submitted, how many results or search engines result pages were viewed can be made on basis of individual users' behaviour.

Another possible unsupervised method is observation of live tickers. Only some search engines provide a live ticker. By visiting a live ticker's web page, one can see the actual searches of other users of this search engine.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

This circumstance is used to observe those tickers automatically by a program. At the end, a list with all search queries, which appeared in the live ticker, is generated.

The transaction log analysis is the most precise method to gain a profound knowledge of what users do in search engines. It is possible to make statements on a global level as well as to analyze patterns on a session and user basis.

All in all, the method most often used is the transactional log analysis of web search engines (Table 1). The results of former studies presented in this context account for the search engines Alltheweb, Altavista, Excite, Fireball, and Webcrawler. Biwe (Buscador en Internet para la Web en Español) is a Spanish directory. Most studies are based on a very short time period of one day only.

Search engine	ID	References	Method	Period	Days
Excite	E1	[35,36,37,38,39,40]	Logs	March 9, 1997	1
Excite	E2	[22,26,27,36,41,42]	Logs	September 16, 1997	1
Fireball	F	[25,26,43,44]	Logs	July 1-31, 1998	31
Altavista	AV1	[25,26,34]	Logs	August 2 – September 13, 1998	43
Excite	E3	[18,22,26,27,36,45,42]	Logs	December 1, 1999	1
Webcrawler	W	[46]	Ticker	March 22 – May 26, 2000	66
Biwe	B	[26,47]	Logs	May 3-18, 2000	16
Alltheweb	A1	[18,26,27,48,49,50,51]	Logs	February 6, 2001	1
Excite	E4	[22,26,27,51]	Logs	April 30, 2001	1
Alltheweb	A2	[26,27]	Logs	May 28, 2002	1
Altavista	AV2	[26,27,52]	Logs	September 8, 2002	1

Table 1. Overview of former studies: data collection method

Besides the description of the data collection method, also the length of observation period should be mentioned. There are only three studies, which include a longer observation period than one day. One day is a very short period to establish standard comparators in online searching behaviour or to discuss changes in the topics searched for. We will show and explain that fact later.

2.2. Size of data sets

The description of the size of data sets doesn't follow the same standards. We introduce five figures to represent the size.

First of all, the number of sessions will be given. This number reflects how many search queries were produced or reformulated during each user session. It is impossible to estimate this number if the analysis is based on data gathered by observation of live tickers.

In most papers the total number of search queries in the collection is reported. We refer to this as the gross number of search queries. Another measure is the total number of terms. Here we use our definition of terms, but by in- or excluding operators and phrases other publications in this area are based on different definitions. We define the gross number of search terms as the number of terms including repetitions minus correctly used operators.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

We refer to net numbers of search queries when we only take the number of search queries into account, which appeared at least once in our data set. The case is the same with net numbers of search terms, still neglecting correctly used operators.

The numbers reported in former studies are shown in Table 2. Most studies report the number of sessions, if it is possible, and the gross number of search queries and terms. The net numbers of queries and terms are seldom given. But these numbers are also interesting since they show how often a term or a query reappears on average in the data set. The division of the gross by the net numbers receives the average appearances of search queries or terms.

ID	Sessions	Gross Number Search Queries	Net Number Search Queries	Avg. Appearance Search Queries	Gross Number Search Terms	Net Number Search Terms	Avg. Appearance Search Terms
E1	18,098	51,473	×	×	113,793	21,862	5.2
E2	211,063	1,025,908	531,416	1.9	2,216,986	140,279	15.8
F	×	16,252,902	×	×	×	×	×
AV1	285,474,117	993,208,159	153,645,050	6.5	×	×	×
E3	325,711	1,025,910	×	×	1,500,500	×	×
W	×	50,538,653	×	×	165,763,490	×	×
B	57,259	105,786	35,518	3.0	173,128	23,707	7.3
A1	153,297	451,551	×	×	1,350,619	180,998	7.4
E4	262,025	1,025,910	×	×	1,538,120	×	×
A2	345,093	957,303	×	×	2,225,141	340,711	7.5
AV2	369,350	1,073,388	×	×	3,132,106	297,528	10.5

Table 2. Overview of former studies: size of data set

The largest study was the analysis of the Altavista log in 1998. This log contains the most sessions, search queries and terms. However the numbers about the amount of terms aren't reported there. The number of queries per day could also give an impression of the number of visitors per day. Since these data sets are collected on different days of different years they would not be comparable due to the progressive growth of the number of internet users between 1998 and 2002 during the internet hype. It is important to compare the length of observation period, too, which is neglected in most studies.

The size of data set leaves an impression about the number and the appearance of terms and queries during these observation periods. By observing only one search engine from time to time those numbers demonstrate de- or increasing numbers of users, but it is not meaningful to compare these numbers of different search engines. Since market shares are different, i.e. the volume of search queries per day, these figures should only be seen as a benchmark how frequently search engines are visited. It is more important that the data set is complete and represents a longer time period than a single day. The largest amount of search queries in a data set does not stand for the most completed data set.

2.3. Query Length

The calculation of the average length of a query can be based likewise on gross or net numbers of queries. The Average lengths based on gross numbers are calculated by the division of the sum of the lengths, i.e. the number of terms in queries, of all search queries by the gross number of search queries.

The average lengths based on net numbers of search queries are calculated by the division of the sum of the length of every search query appearing at least once by the net number of search queries. Normally, the average

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

lengths based on the net number is a little bit higher than the average based on the gross number since one-term queries are very popular. For this, it is interesting to give the percentage of all queries containing only one term. We refer to this as the number of one-term queries.

ID	E1	E2	F	AV1	E3	W	B	A1	E4	A2	AV2
Query Length (by Gross Number)	2.2	2.2	1.7	2.4	2.4	3.3	1.6	2.4	2.6	2.3	2.9
Query Length (by Net Number)	×	2.4	×	×	×	×	×	×	×	×	×
One-term queries	×	26.6%	54.6%	25.8%	28.4%	22.5%	×	25.1%	25.0%	33.1%	20.4%

Table 3. Overview of former studies: query length

The lengths in Table 3 show that a query, on average, contains between 1.6 and 3.3 terms, only. The average query length based on net numbers of search queries is reported for E2, only. Overall, independent of the year and data set, nearly a quarter of all search queries submitted contains one term. That is also the reason why search queries, on average, are very short. This comparator is also the same after a short observation time, so for this comparator no long-time data collections are required.

2.4. Distribution of Queries and Terms

A possibility to describe the distribution of queries or terms is the percentage of search queries or terms, which appeared only once, so-called one-time queries. One-time queries can also be misspelled queries that give a hint for spell checking possibilities. On the one hand, this number is the division of the number of queries or terms by the gross number; on the other hand, the net numbers divides it. It is interesting to evaluate the division by the net number, too, since this comparator gives an idea about the real proportionality of one-time queries and terms. Certainly, it is also possible to mention the percentage of search queries and terms, which appeared only twice, three times and so on.

In [38] and [29], more complex methods for the analysis of the distribution of search queries are described. They apply to the Heap's and Zipf's law [53], respectively. The outcome of the estimation of Heap's law is that only a few search queries and terms appear often. The conclusion of the approximation of Zipf's law is that gross numbers of search queries versus their net numbers don't grow as rapidly as gross numbers of search queries versus their net numbers (pro rata temporis). We don't go deeper in this, since one sees this circumstance by looking at the gross and net number of search queries/terms and at the percentage of search queries/terms, which appeared only once.

ID	E1	E2	F	AV1	E3	W	B	A1	E4	A2	AV2
One-time Search Queries (by Gross Number)	×	×	×	×	×	×	20.3%	×	×	×	×
One-time Search Terms (by Gross Number)	8.6%	6.8%	×	×	7.3%	×	6.3%	7.0%	7.4%	10.0%	5.6%
One-time Search Queries (by Net Number)	×	×	×	63.7%	×	×	60.4%	×	×	×	×
One-time Search Terms (by Net Number)	44.8%	×	×	×	×	×	46.3%	×	×	×	×

Table 4. Overview of former studies: distribution of queries and terms

The figures in Table 4 show the percentage of search queries and terms, which occurred only once based on gross and net numbers. Based on net numbers (AV1 and B) nearly two thirds of search queries appear only once.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

According to net number of terms (E1 and B), a term is reoccurring more often than search queries and there are fewer terms occurring only once.

2.5. Complexity of Search Queries

The complexity of search queries is indicated by the existence of operators in search queries. It's calculated by the number of search queries, which contain correctly typed operators divided by the gross number of search queries. Normally the portion of complex search queries is very small. In some cases, also the percentage of certain operators is calculated. Here, the percentage of searches after phrases is also given since this is the most often used operator in search queries. In some studies, the usage of quotes for phrase search is not seen as an operator but as a modifier.

It is obvious that there are differences in operator usage depending on the search engine and time period the study was conducted (Table 5). In [46] (W), it is reported that the frequent operator usage does not result from more sophisticated search engine users, but rather, it is due to a flexible search engine user interface containing a pull-down menu. The user was allowed to choose an operator on the menu to refine his search.

ID	E1	E2	F	AV1	E3	W	B	A1	E4	A2	AV2
Complexity	15.9%	9.3%	36,6%	20.4%	10.9%	35.6%	12.1%	4.3%	11.3%	4.6%	27.3%
Phrase	6.0%	5.1%	8,6%	×	5.9%	10.4%	2.5%	0.0%	5.9%	0.0%	12.1%

Table 5. Overview of former studies: complexity of search queries

In most studies, nothing is mentioned about incorrectly used operators. An incorrectly typed operator is only a term and does not help to refine a web search. In [26], it is discussed that in European-based search engines such as Alltheweb, Biwe and Fireball, fewer operators were used than in US-based web search engines. The difference is that, in the analysis of, e.g., German-based search engines, web search queries containing 'und' were not considered which corresponds to the incorrectly used operator AND in German. In the Biwe study, the according Spanish Boolean operators were explicitly analyzed, but only those, which were correctly used. This means that operators written in lower case aren't treated as operators since search engines won't handle them as operators, either, but it would be also interesting to analyze wrong usage of operators to see whether this is a language problem that in European countries less operators are used, since they were used in the local language.

2.6. Content Analyses of Search Queries

In most studies, also content analyses have been conducted, as well. By content analysis, we summarize aspects about search queries, which are formulated in a natural way, and the most popular search queries and terms. Another aspect is the restriction of the area in which the search ought to be placed. One possibility is to search only for German web pages, or another possibility is to conduct a search for images only.

Table 6 shows the result of search queries, which are formulated in a natural way. This is the percentage of queries beginning with interrogatives such as who, where, when, see [18]. The percentage in the Webcrawler study is very high, since at that time the Askjeeves Meta search engine, where users conduct search queries in their respective natural languages, used this.

It is also obvious that only in a small amount of searches conducted special search features, offered by search engines, are used.

In most studies, the most popular search terms or queries are referred, too. Since these lists are not very meaningful, we will not show them, in this context. The most popular search queries and terms are regarding sexual aspects and consist in most cases of only one term.

It is also possible to calculate the correlation of terms, which reflects terms being used very likely with other terms, see [34]. In [27,54,55] a small sample of search queries with different topics (ecommerce, multimedia, and

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

sex) was compared by different comparators such as query length or complexity of search queries. An Interesting research question is whether there are significant differences between the formulations of these search queries regarding different topics. We will extend these content analyses by deeper topic analyses with help of our long-time data sets.

ID	Natural	Features	Top Queries	Top Terms
E1	×	0.1%	×	×
E2	×	9.7%	×	✓
F	×	×	×	×
AV1	×	×	✓	×
E3	1.0%	×	×	✓
W	17.9%	×	×	×
B	×	0.2%	✓	✓
A1	×	×	✓	✓
E4	0.3%	×	×	✓
A2	×	×	✓	✓
AV2	×	×	×	✓

Table 6. Overview of former studies: content analysis of search queries

2.7. Temporal Analysis of Search Queries

This category comprises general aspects of daily rush hours of Internet search or high traffic days of the week. To calculate the average percentage of search queries per hour or day one divides the number of search queries, which appeared on a certain day of the week or during certain hours of the day by the gross number of search queries in the data set.

Since most studies are based on a transaction log of one day only, temporal analysis is not possible or reasonable. In [56], a day in the life of web search is explored to detect dependencies in regard to the daytime and different search topics. However, one can easily understand that such a small sample can be highly infected by seasonal topics or upcoming news. Another aspect is the content analysis based on different data sets to detect assumed topic shifts over several years. But regarding the shortage of those time periods, we think that these analyses are not representative. In those studies based on a longer observation period about one month or longer (AV1, F and W), no general aspects about temporal patterns were regarded. We will give examples for temporal analyses by using our data sets.

2.8. Navigational Patterns

It is interesting how many search engine result pages will be regarded by searching persons. How many results does a person click on? These questions can only be answered if detailed click-through data is available. Based on live ticker data, it is not possible to make assumptions about documents retrieved and viewed on a session basis (see, e.g., W in Table 7).

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

ID	E1	E2	F	AV1	E3	W	B	A1	E4	A2	AV2
Only 1st result page viewed	58.0%	66.3%	59.5%	85.2%	69.9%	×	67.9%	54.1	84.6%	76.3%	72.8%
Only one query per session	67.0%	48.5%	×	77.6%	60,4%	×	×	53.0	55.4%	58.7%	47.6%
Queries per session	2.8	2.3	×	2.0	1.9	×	×	2.9	2.3	2.8	2.9

Table 7. Overview of former studies: navigational patterns

Table 7 shows that independent of data basis and time frames more than half the search engine users only regard the first result page. It is also remarkable that about half the users submit only one query per session. That indicates the occurrence of need for information followed by a search query. A possible hypothesis is that people indeed find what they search for on the first result page. But it doesn't mean that on the first result page only high quality pages are located. But regarding those results, this is a very clear pattern which is evident after a short observation time. The sessions are in average very short and contain only up to three search queries.

2.9. Summary

Above, we introduced shortly standard comparators to get key data of searching behaviour in search engines based on transaction logs or on observations of live tickers. We juxtapose different studies to visualize what statistics are given in those. Every publication has analyses departing from these standard comparators or going beyond those basic statistics. We do not refer to all those special analyses in publications, but rather, to stable standard comparators which it is possible to compare data sets with.

One problem with the research of web searching behaviour is the different method of analysis. We already mentioned the different definitions of what is counted as an operator. There is also a lack of standard statistics. Some studies calculate most of the standard comparators we discussed above; others only give some of them. There is no study, which is completely comparable with others. This is mainly caused by the shortness of those data sets and analyses of different search engines. Additionally, there are huge temporal gaps between every observation periods. For example, it is troublesome to compare a data set from the year 1998 of any search engine with any other from the year 2002. In most cases, only one day was the basis for analysis and interpretation. Maybe, there have been influences existed on searching behaviour on this single day that caused significant differences. In 1998 innovators used search engines; in 2002 a broader spectrum and 'normal' people get involved in the usage of the Internet and search engines, too.

These are some reasons why it is important to have longitudinal and simultaneously collected observation data based on different search engines. We have observed four different search engines since summer 2004. We think, this is the first step toward introducing standard comparators to compare search query data. We will show that some standard comparators are the same for every search engine. For some competitors a data set based on a single day is sufficient, as shown above, while for others a long observation period of a month or longer is suggestive. We also regard two meta and two general search engines to compare these standard comparators to see whether there are differences in meta and general search engine usage.

In table 8, the total overview of all standard comparators is given. The numbers in the first column are correspondent to the subchapters in which the results of this top level comparator category is described. We will extend this overview by some more comparators which will come up with long-time data sets.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

1. Data collection method	• Supervised	• Surveys
		• Studies in laboratories
	• Unsupervised	• Log files
		• Observation of live tickers
2. Size of data set	• Number of	• Days • Hours
	• Gross number of • Net number of	• Search queries • Search terms
	• Average appearance of	• Search queries • Search terms
3. Query length	• Based on gross number of • Based on net number of	• search queries
	• One-time queries based on	• Gross number of search queries • Net number of search queries
4. Distribution	• Gross number of one-time • Net number of one-time	• Search queries • Search terms
5. Complexity	• In percent	• e.g., phrase search
6. Content Analyses	• Usage of search features • Usage of natural language • Top 10 of search queries • Top 10 of search queries	
7. Temporal Analyses	• Volume per	• Day • Hour • Week • etc.
8. Navigational Patterns	• Percentage of	• Only first result page viewed • Only one query per session
	• Average number of queries per session	

Table 8. Standard Comparators

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

3. Results and Comparison of Our Data Sets

In this chapter we will compare our empirical data sets of four different search engines by the introduced standard comparators.

3.1. Method of Data Collection

We observed live tickers of three different search engines. In our live tickers observed, the list could be updated automatically by refreshing those pages by use of a program. With this, we collected a nearly complete list of search queries performed on these engines during our observation period.

For every request generated by our program, we saved the time and the search queries shown. The frequency of refreshing and saving these lists was, e.g., 2.5 times per second for Fireball (Lycos: 0.9, Metaspinner: 1.3). These different time interval lengths are caused by the different response times and numbers of search queries shown by those tickers.

With those very short fundamental frequencies of saving the updated live ticker lists, we intentionally caused redundancies in our data saved. We developed an algorithm to put several redundant lists of one observed live ticker together in one representative stack with search queries sorted by the observation time.

We performed 100,000 unique search queries to revise the completeness of our observation method. In the data collection of Fireball, 98,068 of these search queries were retrieved (Lycos: 99,968, Metaspinner: 99,968). In the worst case, we have a missing data rate of 2%. After this performance test, those unique search queries were deleted again for not tampering the following analyses.

Metager sent us a list of the top 4,000 search queries per day. For this, search queries were sorted by their frequency of appearance on a single day. In the next step, search queries, which appeared only once, were sorted alphabetically. Of this daily list, only the first 4,000 search queries were sent to us by email.

Table 9 shows the observation periods and number of days. With respect to the size of data sets and the length of observation period, we are able to conduct a longitudinal study of search queries. Since we collected the data nearly simultaneously, we are, now, also able to compare our results. Based on these data sets, we can also regard temporal patterns in search queries.

Search engine	ID	Method	Period	Number of days
Fireball	FB	Ticker	August 18, 2004 – September 20, 2005	399
Lycos	LY	Ticker	August 14, 2004 – September 20, 2005	403
Metager	MG	Top 4,000	November 11, 2004 – September 20, 2005	314
Metaspinner	MS	Ticker	September 28, 2004 – September 20, 2005	358

Table 9. Our data sets: data collection method

3.2. Size of data sets

Table 10 shows the size of data sets. One sees that general search engines are frequented more often than meta search engines. Regarding gross and net numbers of terms, and the average appearance of a term, it is obvious that search terms occur more often than search queries. Fireball's search terms reoccur five times more often than

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

search queries. With the Lycos search engine, this proportion is four. Concerning the meta search engines this proportion is smaller and lies between two and three. It is assumable that search queries in meta search engines are more diverse and specialized than in general search engines. It is definitely independent of search engines' type that search terms reappear more frequently than search queries.

ID	Sessions	Gross Number Search Queries	Net Number Search Queries	Avg. Appearance Search Queries	Gross Number Search Terms	Net Number Search	Avg. Appearance Search Terms
FB	×	132,833,007	17,992,069	7.4	241,833,877	6,296,833	29.4
LY	×	189,930,859	29,322,366	6.5	344,242,099	11,232,710	30.6
MG	×	4,407,566	678,655	6.5	7,333,343	430,338	17.0
MS	×	4,089,731	1,287,417	6.2	7,853,501	627,507	12.5

Table 10. Our data sets: size of data sets

3.3. Query Length

The average lengths based on gross numbers of search queries are similarly short (Table 11). There are also no significant differences between meta and general search engines.

ID	FB	LY	MG	MS
Query length (by Gross Number)	1.8	1.7	1.6	1.8
Query length (by Net Number)	2.7	2.5	1.4	2.0
One-term queries (by Gross Numbers)	50.1%	51.9%	58.7%	48.4%

Table 11. Our data sets: query length

In comparison to other studies, German-based search queries are a little bit shorter since in German language compounds are often used. Regarding the average length based on net numbers, the measures are bigger. This phenomenon occurs since many search queries consist of only one term and, in addition, are very frequent. In our data sets, half of the top queries consist of only one term. Based on net numbers, the absolute number of occurrences does not carry any weight. That's why in total the average length becomes longer. Metager is the odd one, because here, only the top 4000 terms were regarded to calculate the average length. In this data set there are less one-term queries, since this long tail has been cut, so there is nearly no difference.

3.4. Distribution of Queries and Terms

The distribution of the repeated occurrences of search queries and terms is likewise of interest. From this, it can be derived that, queries are very brief and the average length based on net quantities is larger than the ones based on gross quantities. Additionally, differences between the distribution of search queries and terms are recognizable.

For this, the absolute number of appearances of search queries and terms which turn up exactly 1, 2, 3, 4, 5 and more than 5 times during the observation period are computed. In the next step, the percentage based on gross and net numbers are calculated. The percentage based on gross numbers is calculated by the sum of search queries or terms which appear exactly n times, multiplied by n, and divided by the gross number of terms. The percentage based on net numbers is calculated by the division of the sum of search queries or terms which appear exactly n times by the net numbers.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

	ID	Search queries, which appeared exactly…					
		1	2	3	4	5	>5
Absolut	FB	10,480,377	3,024,799	1,330,798	738,817	461,185	1,956,093
	LY	17,618,682	4,727,513	2,022,780	1,124,878	773,026	3,055,487
	MG	6,580	370,674	168,791	49,597	19,716	63,297
	MS	732,429	224,171	107,354	65,866	42,021	115,021
Percentage of Gross Number	FB	7.9%	4.6%	3.0%	2.2%	1.7%	80.6%
	LY	9.3%	5.0%	3.2%	2.4%	2.1%	78.2%
	MG	0.2%	16.8%	11.5%	4.5%	2.2%	64.8%
	MS	17.9%	11.0%	7.9%	6.4%	5.1%	51.7%
Percentage of Net Number	FB	58.3%	16.8%	7.4%	4.1%	2.56%	10.9%
	LY	60.1%	16.1%	6.9%	3.8%	2.6%	10.4%
	MG	1.0%	54.6%	24.9%	7.3%	2.9%	9.3%
	MS	56.9%	17.4%	8.3%	5.1%	3.3%	9.0%

Table 12. Our data sets: search queries, which appeared exactly…

The figures in Table 12 show that most search queries occurred only once. If only the net number of search queries is considered, then more than half the disjunctive search queries occur only once. With Metager, only 0.2% of these search queries arise once due to the restricted usage of the 4000 most frequent search queries per day.

	ID	Search terms, which appeared exactly…					
		1	2	3	4	5	>5
Absolut	FB	3,093,723	1,077,359	490,277	297,831	197,102	1,140,541
	LY	6,053,824	1,756,677	790,300	462,476	352,419	1,817,014
	MG	2,832	177,288	86,404	39,731	20,823	103,260
	MS	270,024	100,731	55,424	37,453	26,580	137,295
Percentage of Gross Number	FB	1.3%	0.9%	0.6%	0.5%	0.4%	96.32%
	LY	1.8%	1.0%	0.7%	0.5%	0.5%	95.5%
	MG	0.0%	4.8%	3.5%	2.2%	1.4%	88.0%
	MS	3.4%	2.6%	2.1%	1.9%	1.7%	88.3%
Percentage of Net Number	FB	49.1%	17.1%	7.8%	4.7%	3.1%	18.1%
	LY	53.9%	15.6%	7.0%	4.1%	3.1%	16.2%
	MG	0.7%	41.2%	20.1%	9.2%	4.8%	24.0%
	MS	43.0%	16.1%	8.8%	6.0%	4.2%	21.9%

Table 13. Our data sets: search terms which appeared exactly…

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

In Table 13, the values for search terms are similarly in the high rate of terms, which occur only once. Here, less than half of the unique terms occurred only once. The values of Metager are unique, since most queries, which occurred seldom are missing by the exclusive view of the Top-4000 terms.

Overall, the distributions of search queries opposite to those of terms show clear differences. Terms reoccur definitely more often than search queries.

3.5. Complexity of Search Queries

For investigation of search queries' complexity, only those search operators were considered, that were correctly used, as discussed in chapter 1. Wrong inputs which will be ignored by search engines did not become subject of this investigation. The final results are presented in table 14.

ID	FB	LY	MG	MS
Phrase	2.1%	2.4%	×	2.5%

Table 15. Our data sets: complexity of search queries

The results of those studies mentioned in chapter 2 could not be confirmed, here. Operators which are located at the beginning of search terms such as '+', '-', or the phrase search were used relatively frequently. Altogether, the usage of operators accounted for less than 3% of all search queries observed. Phrase search was the most frequent form to arrange search queries in a complex way. Here, the portion of search queries with phrases is 2.1% for Fireball, 2.4% for Lycos, and 2.5% for Metaspinner. This investigation was not accomplished for Metager, since the lists of the Top-4000 search queries contained only search terms in lower case, which would not correspond to a correct operator input. Hence, we didn't consider these measures here.

3.6. Content Analyses of Search Queries

In the following subsections we extended the content analysis by different aspects such as natural language formulation and advanced search features in search engines to show changes and stabilities in search queries and user behaviour. This is important to show that a simple snapshot of topics or a single view on the top 10 of search queries will cause mistakes in interpretation.

3.6.1 Natural language Queries

The relative large number of stop words (and, in, that, or for) in combination with the most frequently used search terms gives the idea that search queries are frequently formed in a way similar to natural language. Thus, not only keywords are entered in order to start a search, but also the suitable fillers ('Bars in Karlsruhe', 'gifts for Christmas'). Moreover, the stop word 'and' in the most frequent terms can point at operators frequently used in a wrong way ('Bars and Karlsruhe', instead of 'Bars AND Karlsruhe'). In most cases, such stop words do not affect search results since search engines filter those out and the results are based on a search query formulated without these words.

ID	Natural	Features	Top Queries	Top Terms
FB	0.1%	65.8	✓	✓
LY	0.1%	×	✓	✓
MG	0.1%	×	✓	✓
MS	0.2%	87.9	✓	✓

Table 16. Our data sets: content analysis of search queries

Due to the fact that the natural language plays a dominant role in the formulation of search queries, it is worthwhile to examine to what extent search queries are formulated alike natural questions. A main indicator for

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

such queries is the presence of question words within the query. The first column in Table 16 shows the percentage of search queries in our data set containing question words.

All in all, the portion is very small regarding Fireball 145,881 queries contained question words and regarding Lycos 244,496 did. The Metager data set contains 4,522 natural questions and the Metaspinner set only 6,789. Concerning Fireball, Lycos and Metager, the proportional frequency of questions is approximately 0.1%. Metaspinner has a portion of approximately 0.2%. Therefore, formulations of search queries such as 'hotels in Berlin' were frequently used, but rarely naturally formulated questions.

3.6.2 Additional Content Information

In their Live Ticker, Fireball and Metaspinner also show beside current search queries the search area. Table 16 shows the proportions of search areas mostly used. The category 'German search' is selected most frequently. This results from the fact that this attitude is a pre-adjustment standard in both search engines. In more than two thirds of all search queries, the basic adjustment will not be changed. However, it is interesting that Fireball users, nevertheless, use the option 'image search' (24.3%). Regarding Metaspinner, all options except the pre-setting 'German search' are strongly underrepresented. These results show that users do not adapt search engines to their own needs or personal preferences.

3.6.3 The top 10 of Search Queries

We also analyzed the top 10 search queries and terms in table 17 and 18. It is recognizable that under the most frequent queries, only one-term queries are located. In addition, it is evident that in general search engines such as Fireball and Lycos, queries dominate which are to be categorized as erotic. In the meta search engines, however, more general search queries are placed, whereas queries such as 'ebay', 'google' or 'freenet' give several possibilities of interpretation. It is conceivable that search engine optimizers manipulate these search queries or that these are navigational queries [57, 58].

Position	FB	LY	MG	MS
1	sex	lycos	ebay	link
2	porno	sex	google	sex
3	hentai	link:http.//www.	routenplaner	routenplaner
4	lycos	hentai	gmx	gebrauchtwagen
5	fkk	porno	telefonbuch	hotel
6	erotik	ebay	www.ebay.de	london
7	lack	erotik	web.de	jahreswagen
8	anal	google	sex	autobewertung
9	titten	fkk	www.google.de	versicherung
10	bondage	christina aguilera	wetter	werkswagen

Table 19. Our data sets: Top ten of search queries

It is likewise possible that the top 10 search queries are consciously falsified in order to mislead persons, who use those lists as a guideline. This could be done via automated queries, so that the possibility for web page administrators of aligning their Internet advertising campaigns at these lists is taken. By the manipulation of those lists, no courses of action are derivable since they do not show the original search behaviour of users. There are further manipulation possibilities, but due to the high potential of interpretation we do not go deeper in this.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

Many users get also used to type in a known name of a company or parts of URLs of web sites to receive the whole link as a top result in search engines' result pages. Broder and Lewandowski refer to this as navigational queries [57, 58].

It is interesting that in the listings of the top 10 search terms similar positions emerge as in the top 10 lists of search queries. There are many queries, which consist only of a single word. Therefore, the most frequent terms correspond, in many cases, to the most frequent search queries.

Position	FB	LY	MG	MS
1	sex	lycos	ebay	in
2	download	sex	google	download
3	in	in	der	link
4	free	der	routenplaner	der
5	porno	und	download	berlin
6	dvd	hotel	und	hotel
7	hotel	von	in	und
8	kostenlos	für	für	für
9	der	porno	berlin	dvd
10	bilder	free	sex	von

Table 20. Our data sets: Top ten of search terms

By comparison of the top 10 of the terms from two different periods, it is also evident that no substantial shifts occur. The only difference between different search engines is that meta search engines exhibit fewer erotic terms in their listings. All in all, neither the top 10 of the search queries nor those of the terms seem to give a representative overview of fluctuations or trends in search queries. For this, more meaningful evaluations must be performed in order to be able to represent changes in behaviour.

3.6.4 Evergreens in Search Queries

As the evaluation of search queries and terms revealed, the pure top 10 lists did not give an interesting perspective on topics searched by Internet users. We will, here, shortly introduce the concept of Evergreens that gives better possibilities to interpret time-dependent changes in topics. **Evergreens** are topics which are over a long-time period in the users' mind. Here, only terms are considered, since knowledge of evergreen terms give implications for evergreen search queries, too [21].

In [46], a procedure has been reported to visualize the distribution of search terms in different time slots. Evergreens have to appear in nearly every time slot defined. Here, we define a time slot as a day, and a term, which appeared in 90% of all days observed (403, e.g., for Lycos) as evergreens. Additionally, we define a threshold f to consider only days where a term is occurring frequently enough. If a search term is requested less than f times, this term does not appear in our analysis on that day. All evergreens of the Lycos search engine with f=400 are shown in table 21. One sees that most of the evergreens are very short search queries, too. Evergreens give a better insight in search topics and terms.

For search engines, knowledge of frequently returning terms is of interest, since recommendations or adapted caches in the background can be built up. Search queries and terms which come up nearly every day can be saved with their result pages. Knowledge of evergreens is also helpful to name categories of web sites or in web directories properly. In [21] instable patterns in search queries and terms are reported which occur after breaking news or upcoming seasonal events.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

	term			term			term			term			term	
(o)	+	401	(e)	porno	400	(e)	sexy	398	(f)	mit	394	(e)	girl	386
(e)	sex	401	(f)	the	400	(f)	von	398	(e)	nude	393	(sh)	kostenlos	386
(t)	berlin	400	(o)	und	400	(f)	de	397	(e)	porn	393	(m)	video	386
(m)	bilder	400	(o)	-	399	(t)	hotel	397	(e)	gay	392	(t)	hamburg	383
(f)	der	400	(m)	download	399	(f)	of	397	(f)	im	392	(sh)	kostenlose	381
(f)	die	400	(e)	erotik	398	(o)	and	396	(e)	pics	391	(e,t)	hilton	376
(sh)	free	400	(e)	frauen	398	(e,t)	paris	396	(se)	ebay	390	(e)	hardcore	375
(e)	hentai	400	(e)	girls	398	(m)	dvd	395	(f)	2	389	(sh)	shop	374
(f)	in	400	(e)	nackt	398	(f)	für	395	(se)	google	389	(m)	software	374
(se)	lycos	400	(m)	online	398	(e)	fkk	394	(o)	or	388	(e)	titten	372

Table 22. Our data sets: evergreening search terms in the Lycos search engine

Categories, which the words are divided into, are indicated in parentheses in front of the terms. The topic areas are erotic (e), fillers (f), multimedia and music (m), operators (o), travel (r), shopping (sh) and search engines (se). The terms 'Hilton' and 'Paris' were mostly combined in search queries. From this fact it can be concluded that 'Paris Hilton' is meant, even though two topic areas are indicated. Altogether, 17 of 50 identified evergreens fall in the category erotic (e) and another eleven are fillers. Regarding this, topics in search queries do not shift from eSex to eCommerce [22]: they are still erotic related. Besides, it is interesting that such a high portion consists of fillers, which do not affect the search restriction, since they are not considered by most search engines. Therefore, natural speech is firmly embodied with the search so that people do not work with keywords only.

This concept shows also that lists of top search queries and terms that are reported in most papers published are not a meaningful basis for further topic analyses. Also the representation and interpretation of changes in topics over the last years of search engine usage by data analyses of a single day is not sufficient.

3.7. Temporal Analyses of Search Queries

In Fig. 1, the results of the average quantities of search queries during certain hours of the day are shown. The daily height lies between 1 and 3 pm. The absolute minimum is reached in the early morning hours between 5 and 6 am. While the minimum corresponds to the results in [56], the maximum is observed in the afternoon and not in the late evening hours for all search engines. Here, we only show the result for Lycos (LY) and Fireball (FB) to achieve an easily identifiable figure. From Metager, only the list of the Top-4000 retrieval queries is present, why the hourly temporal information was lost. The Metaspinner curve has same characteristics.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

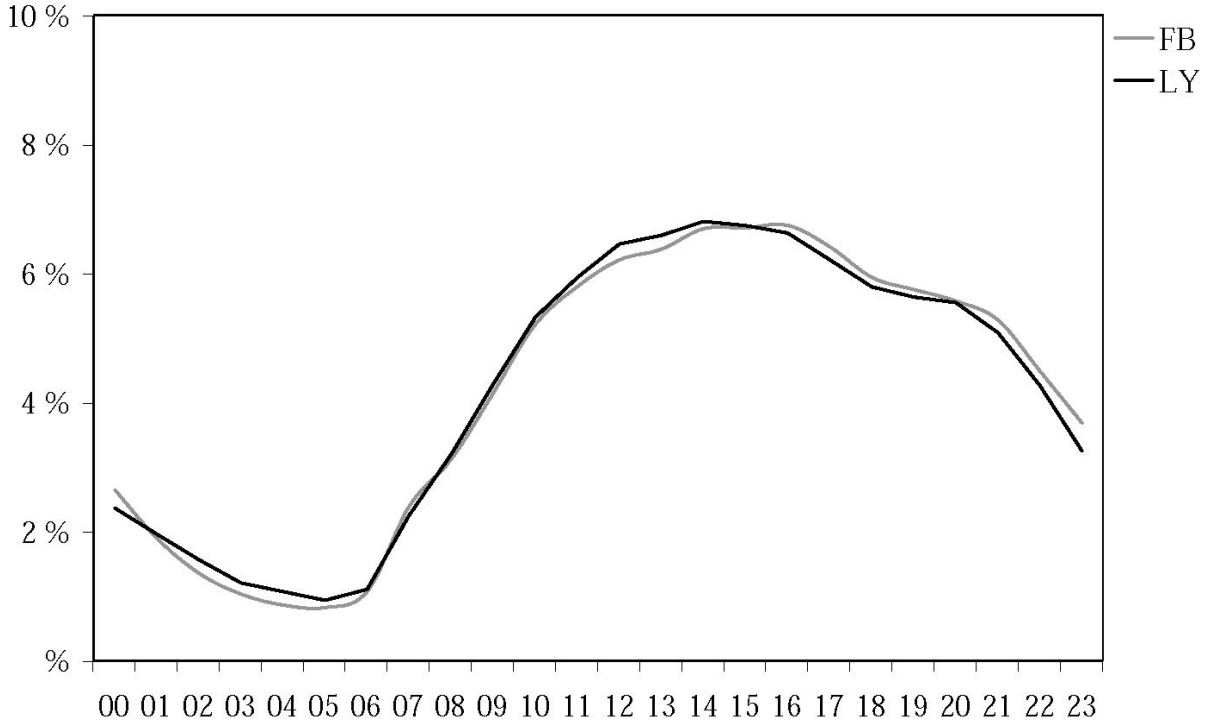

Fig. 1. Our data sets: Per cental distribution of search queries during the day

It is also possible to analyze the distribution of the average quantity over the weekdays in percent. It is obvious that on Mondays, more search queries are submitted than on other weekdays. On Mondays 14% of all search queries is submitted. The proportion decreases until Saturdays. Here, only 10% of all search queries are submitted, on Sundays 12%). This effect shifts however when a Monday is a holiday (e.g., Easter Monday). Then the following Tuesday shows this arising quantity of search queries. We refer to this effect as the **Monday effect**.

3.8. Navigational Patterns

We do not have any detailed session data, since our observation data is not based on transactional logs of search engines. This is the reason we do not have any possibility to discuss navigational patterns. The same is true with the result pages viewed and session lengths.

4. Summary and Implications

We give a guideline for further studies and research of online searching behaviour how to analyze data sets of search queries and which analyses have to be provided by researches to produce comparable results. Our standard comparators are a platform on which studies can be conducted. All these standard comparators should be calculated and mentioned in further studies to see differences and similarities in between different data sets. This is also important for data sets which are achieved from different search engines and years.

We show by comparing data of long but similar observation periods that some conclusions in papers published so far are not correct. Some comparators, such as the length of queries, operator usage, or result pages viewed, do not require a long-time data basis, but as soon as further analyses will be done, such as content and topic analyses, a long-time data basis is necessary. We extended the content analyses by introducing the concept

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

of evergreens. With this it is possible to show stabilities in search queries and their topics. Additional content analyses are presented in [21,23]. There are also fluctuations in search queries discussed.

It is important to compute these comparators from time to time, to notice changes in user behaviour to readjust search engine interfaces and features. Since operators are rarely used a drop down menu in search engines should be considered. Only pre-adjustment features are used in most cases while other features except image search are not adopted by users.

We filled also the knowledge gap with respect to German-speaking users by collecting search queries in live tickers of four different German (meta) search engines. Summarizing, search queries are very short and are not complex. The most favoured operator used by a searching person is the phrase search. The topic, which appears most frequently in the top 10 of general search engines, is 'erotic' whereas in the top 10 of the meta, more general topics are important. The 'erotic' topic is represented only once by the query 'sex'. Additionally, fillers and wrongly used operators dominate the lists. Most search engine users conduct searches with the standard configuration settings.

By regarding the functionality between distinct terms/search queries and gross number of search queries per time, we see that the number of terms increases much slower than the number of search queries.

The 'quiet period' of search traffic is in the early morning and the 'rush hour' at noon with a peak after lunchtime. On Mondays the volume of search queries conducted is higher than on other days with the minimum on Saturdays.

Since search engine users do not accept operators or special configuration settings, search engine interfaces should be designed in a very simple way with a special announcement for experts. A discussion of quality measurement aspects of search engines, which also focus on users, is given in [57].

The preferred topic is 'erotic', so it would be advisable to build an 'erotic' cache to serve such searches directly without affecting the 'normal' cache. Additionally, it is possible to cache the most frequently used terms to recommend the top web sites corresponding to these searches. In consideration of 'rush hours' and 'high frequency' days, administrative work on search engines should be conducted early Saturday morning. Special events or search features such as 'news' should be realized on Mondays. This is also interesting for selling things on Ebay. Biddings should end on Mondays at noon.

Leveraging the identification of evergreens, caching strategies can be developed. Search engines can build an exclusive index, which is only based on pages that deal with evergreen (search) topics and another, which deals with non-evergreen topics. Such a divided index increases the capacity for new queries and decreases the response time, which is of particular importance when a brand search topic occurs. This insight is of high relevance for the design of online portals as well, since two main evergreens are erotic and travel subjects. It is also important to know evergreen topics for marketing campaigns or for generating increased traffic on pages. Another possibility is not to operate with those common terms only, but to combine common evergreen terms with exotic ones.

Most important is the fact that longitudinal data as we introduced with our data sets provide us with a deep glimpse into information interests of searching persons [21,23]. It is difficult and problematical to analyze shifts and changing of human searching behaviour and topics on the basis of very short data sets of different years. Internet user demographics and the social structure have changed during the Internet hype. It is advisable to compare different data sets concerning the same observation period and comparators.

We will go on with our research on query topics and time series analyses of search queries. Our next step will be to predict trends and upcoming new topics in search queries to advise web site publishers of new trends and keywords to optimize their pages. Web site optimization due to search engines could cause the exclusion of pages by search engines since they are regarded as spam pages. A better and fair way to optimize pages would be the user and searching behaviour focused approach. For this, it is very important to have an insight into searching behaviour. Search engines should also use the information of search queries to improve ranking algorithms and to detect spam queries by reoccurring patterns.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

5. References

[1] Y. Theng, M. Jones, and H. Thimbleby , Lost in Hyperspace - Psychological Problem or Bad Design?, *Proceedings of the first Asia Pacific Conference on Computer Human Interaction*, (Singapore, June 1996) 387-396.

[2] A. Pollock and A. Hockley, *What's Wrong with Internet Searching* (1997). Available at: D-Lib Magazine, http://www.dlib.org/dlib/march97/bt/03pollock.html (accessed 13 June 2007).

[3] M. Berry (Ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval* (Springer, 2003).

[4] M. Berry and M. Browne, *Understanding Search Engines - Mathematical Modeling and Text Retrieval, Software-Environments-Tools* (SIAM, 1999).

[5] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of the 7th International World Wide Web Conference*, (Brisbane, Australia, 1998) 107-117.

[6] T. Calishain and R. Dornfest, *Google Hacks*, (O'Reilly & Associates, 2003).

[7] D. Lewandowski, *Web Information Retrieval: Technologien zur Informationssuche im Internet*, Informationswissenschaft 7, (Frankfurt am Main, DGI, 2005).

[8] D. Lewandowski, Web Searching, Search Engines, and Information Retrieval, *Information Services & Use* 25 (2005) 137-147.

[9] L. Vaughan and M. Thelwall, Search Engine Coverage Bias: Evidence and Possible Causes, *Information Processing & Management* 40(4) (2004) 693-707.

[10] M. Bates, The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *Online Review*, 13 (1998) 407 – 424. Available at http://www.gseis.ucla.edu/faculty/bates/berrypicking.html (accessed 13 June 2007).

[11] L. Catledge and J. Pitkow, Characterizing Browsing Strategies in the World Wide Web, *Proceedings of the 3rd International World Wide Web Conference*, (Darmstadt, Germany, 1995). Available at http://www.igd.fhg.de/archive/1995_www95/proceedings/papers/80/userpatterns/UserPatterns.Paper4.formatted.html (accessed 13 June 2007).

[12] Ch. Choo, B. Detlor, and D. Turnbull, *Information Seeking on the Web: An Integrated Model of Browsing and Searching*, First Monday 5(2) (2000). Available at http://firstmonday.org/issues/issue5_2/ (accessed 13 June 2007).

[13] Ch. Choo, B. Detlor, and D. Turnbull, Working the Web: An Empirical Model of Web Use, *Proceedings of HICSS 33 Knowledge: Creation, Organization and Use*, 7 (Maui, Hawaii, 2000).

[14] Ch. Choo and Ch. Marton, Information Seeking on the Web by Women in IT Professions, *Internet Research* 13(4) (2003) 267-280.

[15] D. Hawkins, Hunting, Grazing, Browsing: A Model for Online Information Retrieval, *Online*, 20(1) (1996) 71-73.

[16] G. Marchionini, *Information Seeking in Electronic Environments* (Cambridge, Cambridge University Press, 1995).

[17] R. Navarro-Prieto, M. Scaife, and Y. Rogers, *Cognitive Strategies in Web Searching*, Proceedings of 5th Conference on Human Factors and the Web, (Gaithersburg, Maryland, USA, 1999). Available at http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto (accessed 13 June 2007).

[18] S. Ozmutlu, and H. Ozmutlu, and A. Spink, Are People Asking Questions of General Web Search Engines?, *Online Information Review*, 27(6) (2003) 396-406.

[19] A. Spink, J. Bateman, and B. Jansen, Searching heterogeneous collections on the Web: Behaviour of Excite Users, *Information Research*, 4(2) (1998) 317-328.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

[20] A. Spink, J. Bateman,and B. Jansen, Searching the Web: Survey of Excite Users, *Internet Research: Electronic Networking Applications and Policy*, 9(2) (1999) 117-128.

[21] N. Schmidt-Maenz, *Untersuchung des Suchverhaltens im Web - Interaktion von Internetnutzern mit Suchmaschinen*, (Dr. Kovac Verlag, Hamburg, 2007).

[22] A. Spink, B. Jansen, D. Wolfram, and T. Saracevic, From E-Sex to E-Commerce: Web Search Changes, *IEEE Computer*, 35(3) (2002) 107-109.

[23] N. Schmidt-Maenz and M. Koch, A General Classification of (Search) Queries and Terms, *Proceedings of the third International Conference on Information Technology: New Generations (ITNG' 2006)*, (Las Vegas, Nevada, USA, IEEE Computer Society, 2006), 122-129.

[24] I. Hsieh-Yee, Effects of Search Experience and Subject Knowledge on Online Search Behaviour: Measuring the Search Tactics of Novice and Experienced Searchers, *Journal of the American Society for Information Science*, 44 (1993) 161-174.

[25] B. Jansen and U. Pooch, Web User Studies: A Review and Framework for Future Work, *Journal of the American Society of Information Science and Technology*, 52(3) (2001) 10073-1074.

[26] B. Jansen and A. Spink, How Are we Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs, *Information Processing and Management*, 52(1) (2006) 248-263.

[27] A. Spink and B. Jansen, B., *Web Search: Public Searching of the Web*, (Kluwer Academic Publishers, Dordtrecht, 2004).

[28] J. Bar-Ilan, The Use of Web Search Engines in Information Science Research, *Annual Review of Information Science and Technology*, 38 (2004) 231-288.

[29] P. Wang, J. Bownas, and M. Berry, Trend and Behaviour Detection from Web Queries. In: M. Berry (Ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval* (Springer, 2003).

[30] P. Wang, J. Bownas, and M. Berry, Mining Longitudinal Web Queries: Trends and Patterns, *Journal of the American Society for Information Science and Technology: JASIST*, 54(8) (2003) 743-758.

[31] P. Huntington and D. Nicholas, Improving the Relvance of Web Menus Using Search Logs: a BBCi Case Study, *Aslib Proceedings*, 58 (1/2) (2006) 118-128.

[32] M. Thelwall, L. Vaughan, and L. Bjoerneborn, Webometrics, *Annual Review of Information Science and Technology*, 39 (2005) 81-135.

[33] D. Nicholas, H.R. Jamali, and A. Watkinson, The Information Seeking Behaviour of Users of Digital Scholary Journals, *Information Processing & Management*, 42(5) (2006) 1345-1365.

[34] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, *Analysis of a Very Large Web Search Engine Query Log*, ACM SIGIR Forum, 33(1) (1999) 6-12.

[35] B. Jansen, An Investigation Into the Use of Simple Queries on Web IR Systems, *Information Research*, 6(1) (2000).

[36] B. Jansen and A. Spink, Methodological Approach in Discovering User Search Patterns Through Web Log Analysis, *Bulletin of the American Society for Information Science and Technology*, 27(1) (2000) 15-17.

[37] B. Jansen, A. Spink, J. Bateman, T. Saracevic, *Searchers, the Subjects they Search, and Sufficiency: A Study of a Large Sample of Excite Searchers*, Proceedings of WebNet-98, World Conference on the WWW and Internet, (Orlando, USA, 1998) 472-477.

[38] B. Jansen, A. Spink, and A. Pfiff, *A Linguistical Analysis of World Wide Web Queries*, American Society of Information Science and Technology, Annual Meeting: Knowledge Innovations: Celebrating Our Heritage, Designing Our Future (2000).

[39] B. Jansen, A. Spink, and T. Saracevic, Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, *Information Processing and Management*, 36(2) (2000) 2007-2027.

[40] A. Spink, B. Jansen, and H. Ozmutlu, Use of query reformulation and Relevance Feedback by Excite Users, *Internet Research: Electronic Networking Applications and Policy*, 19(4) (2000) 317-328.

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007

[41] A. Spink, D. Wolfram, B. Jansen, and T. Saracevic, Searching the Web: The Public and Their Queries, *Journal of the American Society fot Information Science and Technology*, 52(3) (2001) 226-234.

[42] D. Wolfram, A. Spink, B. Jansen, and T. Saracevic, Vox Populi: The Public Searching of the Web, *Journal of the American Society for Information Science and Technology*, 52(23) (2001) 1073-1074.

[43] Ch. Hoelscher and G. Strube: The Role of Knowledge in WWW Search: Differences in IR Skills and Domain Expertise, *Multi-Modalities and Interactions*, (Brest, France, HCP'99, 1999), 425-430.

[44] Ch. Hoelscher and G. Strube, Web Search Behaviour of Internet Experts and Newbies, *Proceedings of the 9th International World Wide Web Conference*, (Amsterdam, the Netherlands, 2000) 337-346.

[45] H. Ozmutlu, A. Spink, and S. Ozmutlu, Analysis of Large Data Logs: An Application of Poisson Sampling on Excite Web Queries, *Information Processing and Management*, 38 (2002) 473-490.

[46] J. Zien, J. Meyer, J. Tomlin, and J. Liu, *Web Query Characteristics and their Implications on Search Engines*, (Almaden Research Center, Research Report, RJ 10199 (95073), 2000).

[47] F. Chacheda and Á. Viña, Understanding How People Use Search Engines: A Statistical Analysis For E-Business, Proceedings of the e-2001 (e-Business and e-Work Conference and Exhibition), 1 (2001) 319-325.

[48] B. Jansen and A. Spink, An Analysis of Web Documents Retrieved and Viewed, *Proceedings of the 4th International Conference on Internet Computing*, (Las Vegas, USA, 2003) 65-69.

[49] S. Ozmutlu, H. Ozmutlu, and A. Spink, A Study of Multitasking Web Search, *Proceedings of the International Conference on Information Technology: Computers and Communications*, (ITCC, 2003), 145-148.

[50] S. Ozmutlu, A. Spink, and H. Ozmutlu, A Day in the Life of Web Searching: An Exploratory Study, *Information Processing and Management*, 40 (2004) 319-345.

[51] A. Spink, S. Ozmutlu, H. Ozmutlu, and B. Jansen, U.S. Versus European Web Searching Trends, *ACM SIGIR Forum*, 36(2) (2002) 32-38.

[52] A. Spink, M. Park, B. Jansen, and J. Pedersen, Multitasking Web Search on AltaVista, *Proceedings of the International Conference on Information Technology: Coding and Computing*, 2 (2004) 309-313.

[53] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, (Addison Wesley Publishing Company New York, USA, 1999).

[54] A. Spink, and O. Gunar, E-Commerce Web Queries: Excite and Ask Jeeves Study, *First Monday, Peer-Reviewed Journal on the Internet*, 6(7) (2001) 4.

[55] A. Spink, H. Ozmutlu, and D. Lorence, Web Searching for Sexual Information: An Exploratory Study, *Information Processing and Management*, 40 (2004) 113-123.

[56] S. Beitzel, C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, Hourly Analysis of a Very Large Topically Categorized Web Query Log, *Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM-SIGIR, Sheffield, UK, 2004).

[57] A. Broder, *A Taxonomy of Web Search*, SIGIR Forum, 36(2) (2002) 3-10.

[58] D. Lewandowski, Query types and search topics of German Web search engine users, *Information Services & Use*, 26(4) (2006) 261-269.

[59] D. Lewandowski, N. Hoechstoetter, Web Searching: A Quality Measurement Perspective. To appear in: A. Spink, M. Zimmer (Eds.), *Web Search: Interdisciplinary Perspectives*, Series of Information Science and Information Management, (Springer, Heidelberg, 2008).

JIS-0523 - v5 Received: 20th March 2007 Revised: 24th October 2007