Information Services & Use 26 (2006) 261–269
IOS Press

Query types and search topics of German Web search engine users

Dirk Lewandowski

Heinrich-Heine-University Düsseldorf, Department of Information Science, Universitätsstraße 1, D-40225 Düsseldorf, Germany
E-mail: dirk.lewandowski@uni-duesseldorf.de

Abstract. The purpose of this study is to investigate the topics of searches in German Web search engines and the query types used, as well. Based on the query types identified by Broder and the classification of search topics developed by Spink et al., 1500 queries from German search engines Fireball, Seekport and Metager (deduced from log files and live ticker data) are assigned to a topic category and to a query type, respectively. We find that all query types are used to a large extent by our user population. The same holds true for the search topics. A combination of both shows that the distribution of query types within topics areas is uneven. This has implications on the development of more user-centred search engines.

Keywords: Web searching, search engines, query types, search topics

1. Introduction

Web search engines are an important field in current information retrieval (IR) research. What makes them so interesting are mainly the great challenges they are faced with. Search engines are unique in

- the size of the document collection,
- the number of users,
- the need to satisfy queries on all topics,
- the need to satisfy different query types (which are not specified by the users),
- the problems with unreliable or deceptive content.

For an overview of the major differences between classic Information Retrieval (IR) and Web Information Retrieval and the main problems search engines have to face when indexing the Web, see [5].

It is essential to know how users interact with IR systems in general and with search engines in particular. It is known that users search for a wide range of topics on these systems. In addition, several query types are used: Apart from informational queries, which are mainly used in traditional information retrieval systems, one can also find navigational and transactional queries on the Web [1].

In our investigation, we want to study the behaviour of German search engine users. It is currently unknown if their behaviour differs from that of the well investigated U.S. American users. In a combination of topics and query types, we want to draw a clearer picture of the needs of our user group.

2. Related work

In this section, research in three areas is discussed. First, general findings on the behaviour of German search engine users from various studies will be presented. Then, investigations on the search topics of search engine users will be reviewed and finally, the differentiation of queries based on query types will be described.

2.1. German Web search engine users

The studies investigating the behaviour of German Web search engine users mainly deal with the complexity of queries (average number of search terms, use of operators), the browsing within the result sets (number of results viewed, number of results pages viewed) and the search process as a whole (session length, reformulation of queries).

There is a comprehensive study based on a representative poll and a laboratory experiment as well [6,7]. The authors find that the users generally prefer one engine and only seldom use another one. The average query is usually short and only contains two or three words. Boolen operators and advanced search pages are known by half of all users, but not used very often.

Log file based studies confirm short query lengths in German Web search engines. For Fireball [2], the average query length is 1.66 terms, for Lycos [10] 1.7 terms. The on average shorter queries in German search engines compared with US American/international engines probably result from different rules for compound words in English and German. Both studies also show that Boolean operators are seldom used.

2.2. Search topics

Spink et al. [12] are interested in the topics that search engine users are searching for. They classified about 2500 queries from a log file of the Excite search engine into 11 categories. Further studies with data from other search engines (summarized in [3]) were conducted over the years with the newest data from 2002.

The advantage of log file based studies is that real user queries can be observed. The user does not know that an investigation takes place and therefore is not influenced in his behaviour. But the disadvantage of such studies is that one can only determine the queries with little additional information. A direct inference on the information need is not possible.

As with other qualitative analysis as well, there is a risk that the classification of topics from queries is subjective. The authors try to avoid this by having the queries classified by two people independently and then let them discuss all cases of doubt until they reach a conjoint conclusion. Although this approach cannot guarantee completely sound assignment of the search topics to the classes, we believe that such a pragmatic solution fits to a relatively simple classification scheme. For reasons of comparability, we will use the classification scheme in our study as it is and do not apply any changes to it.

The 11 classes developed in [12] are

- People, places and things
- Computers or Internet
- Commerce, travel, employment, or economy
- Entertainment or recreation
- Health or sciences

- Sex or pornography
- Government
- Education or humanities
- Society, culture, ethnicity, or religion
- Performing or fine arts
- Unknown or other

2.3. Intercultural differences in search topics

Spink et al. [11] compare the user behaviour of U.S. and European search engine users on the basis of data sets from Excite (2001) and FAST (2001), respectively. They say that "the majority of FAST users are believed by FAST to be from Europe, mostly from Germany".

For the comparison, approx. 2500 English language queries from both search engines are classified. They find that the topics searched for differ in the two search engines.

But there are two major problems with this study:

- For the investigation on the search topics, Spink et al. only use English language queries. This appears strange when the majority of queries are from Germany and therefore, mainly in German. The study can therefore only make statements on non-German queries by (mainly) German users. For the purpose of comparing the search topics of Europeans/Germans with U.S. Americans, this seems to be of very limited worth.
- There are reasons to doubt that most of the queries in the 2001 FAST data set are from Germany. Another FAST data set (from 2002) was analysed by the same research group for the query languages: "From our analysis of the AlltheWeb.com [i.e. FAST] transaction log, nearly 90% of the query requests are in English, 6% French, 1% each Spanish, German, Italian, and a variety of other languages making up the rest" [3]. From one year to another the change seems to be too big, even though the second data set is quite larger than the first.

For a discussion on the differences of the 2001 FAST data set used by Spink et al. and other data sets see Section 5.1.

2.4. User intents in Web search

Andrei Broder [1] presents a taxonomy of Web search that focuses on the goals a user pursues with a query. The author distinguishes between navigational, informational and transactional queries.

With *informational queries* users want to find information on a certain topic. Such queries usually lead to a set of results rather than to just one suitable document. Informational queries are similar to queries sent to traditional text-based IR systems. According to Broder, such queries always target static Web pages. But the term "static" here should not refer to the technical delivery of the pages (e.g., dynamically generated pages by server side scripts like php or asp) but rather to the fact that once the page is delivered, no further interaction is needed to get the desired information.

Navigational queries are used to find a certain webpage the user already knows about or at least assumes that such a webpage exists. Typical queries in this category are searches for a homepage of a person or organization. A typical example is the search for a company (e.g., "Daimler Chrysler"). Navigational queries are usually answered by just one result; the information need is satisfied as soon as this one right result is found.

But not all queries for people are navigational. Rose and Levinson [8] are of the opinion that most queries for people are in fact not. They point out that "a search for celebrities such as Cameron Diaz or Ben Affleck typically results in a variety of fan sites, media sites, and so on: it's unlikely that a user entering the celebrity name as a query had the goal of visiting a specific site." This may be true, but these queries cannot be seen as informational because one cannot assume that the user wants to read a variety of documents. Therefore, we decided to classify queries for celebrities as navigational.

The results of *transactional queries* are Web sites where a further interaction is necessary. A transaction can be the download of a program or file, the purchase of a product or a further search in a database.

Based on a log file analysis and a user survey (both from the AltaVista search engine), Broder finds that each query type stands for a significant amount of all searches. Navigational queries account for 20–24.5 percent of all queries, informational queries for 39–48 percent and transactional queries for 22–36 percent.

The results from Rose and Levinson [8] are only in part comparable to Broder's original study, mainly because of different definitions of navigational and transactional queries. According to Rose and Levinson, informational queries account for 61 to 63 percent of all queries, transactional queries for 21 to 27 percent and navigational queries for only 11 to 15 percent.

The division of the queries into three types should be seen as seminal, because it was the first attempt to differentiate between user intents expressed in queries. Further work investigated the possibility of assigning an intention to a query automatically [4].

3. Research questions

On the one hand, we wanted to use the methods of the studies described above to find out more about the search behaviour of German search engine users. On the other hand, we wanted to combine both methods to examine if there are differences in the use of certain query types in the various topic fields. This information could be used to identify whether search engines handle queries with a certain information need adequately.

For example, there could be a topic area where the majority of queries is informational, but the majority of results from a certain search engine is transactional. This would mean that this engine supplies the user with the "wrong" results.

In summary, our research questions are:

(1) What are the main topics German Web search engine users are searching for?
(2) Are there differences in search topics and/or query types between German and U.S. American users?
(3) How are the queries distributed amongst the query types?
(4) Will a combination of query types and topics be fruitful for a better understanding of users' needs behind the queries?

4. Methods

4.1. Choice of search engines and data collection

For the purpose of this study, it would be ideal to have log files from all major search engines. But we found it very difficult to obtain these files from the engines. There are several concerns about giv-

ing such files to researchers not belonging to the company itself. The only exception was Metager <www.metager.de>, a German meta search engine run by the Search Engine Lab at the University of Hanover which gave us access to their complete log files.

With regard to the other search engines, we had to rely on "live queries" shown on the search sites. Such a functionality is not available on all search engines (mainly not on the majors such as Google, Yahoo or MSN), so we had to choose some smaller engines. Our choices were two genuine German search engines, Fireball <www.fireball.de> and Seekport <www.seekport.de>.

We used a total amount of 1500 real user queries from three different German search engines (Fireball, Seekport and Metager). From each engine, 500 queries were intellectually classified into one topic category and one query type category, respectively. Two persons had to achieve an agreement on each judgement.

All data was collected on October 18, 2005. For Metager, we were able to obtain our queries from a complete log file. For the other engines, we had to use the "Live search" (cf. [5]) where one can see for what topics other users are searching for at the moment. We used this feature at different times of the day to avoid using only queries from a certain time. We know that this approach has limitations but it was the only way to obtain data from these search engines.

During our data collection, we found that Seekport filters some kinds of queries (especially terms in sexual context). This is a limitation to the Seekport data set and will be considered in our analysis.

4.2. Limitations

Our research is limited in the following ways:

- We had no access to queries from one or more of the larger search engines. But this is acceptable because there does not seem to be too much difference between query sets from different general-purpose search engines [2].
- There were some problems in classifying, especially with short (one word) queries. Therefore, we had a relatively large proportion of queries classified as "unknown". On the other hand, we wondered how the other researchers were able to classify most of the queries accurately under the given conditions.

5. Results

First, we present results on the query types, then on the search topics and finally, the combined results on the query types within the different topic areas.

5.1. Search topics

Table 1 shows rankings for all search topics in the different search engines of our study and the data for the two search engines in the comparison study by [11].

Regarding research question (1), we find that topics most searched for are "Commerce, travel, employment, or economy" (29 percent of all queries) and "People, places or things" (12.8 percent). All other categories are below 10 percent. "Computers or Internet", "Entertainment or recreation" and "Health or sciences" are between seven and eight percent. All other topics are between one and little more than four percent.

Table 1
Comparison of general topic categories

Rank	2001 Excite Data Set	2001 FAST Data Set	2005 Metager Data Set	2005 Fireball Data Set	2005 Seekport Data Set	Average of all 2005 Data Sets
1	24.7% Commerce, travel, employment or economy	22.5% People, places or things	25.4% Commerce, travel, employment, or economy	30.6% Commerce, travel, employment, or economy	31.6% Commerce, travel, employment, or economy	29.0% Commerce, travel, employment, or economy
2	19.7% People, places or things	21.8% Computers or Internet	10.6% People, places or things	10.2% Sex or pornography	19.0% People, places or things	12.8% People, places or things
3	9.6% Computers or Internet	12.3% Commerce, travel, employment, or economy	8.8% Health or sciences	9.0% People, places or things	7.6% Entertainment or recreation	7.7% Entertainment or recreation
4	8.5% Sex or pornography	10.8% Sex or pornography	8.6% Computers or Internet	7.8% Entertainment or recreation	6.2% Computers or Internet	7.4% Computers or Internet
5	7.5% Health or sciences	9.1% Entertainment or recreation	7.8% Entertainment or recreation	7.6% Health or sciences	5.6% Health or sciences	7.3% Health or sciences
6	6.6% Entertainment or recreation	7.8% Health or sciences	5.0% Society, culture, ethnicity or religion	7.6% Computers or Internet	5.0% Society, culture, ethnicity or religion	4.5% Sex or pornography
7	4.5% Education or humanities	4.8% Society, culture, ethnicity or religion	4.8% Government	2.0% Society, culture, ethnicity or religion	4.0% Government	4.0% Society, culture, ethnicity or religion
8	3.9% Society, culture, ethnicity or religion	4.7% Performing or fine arts	3.6% Education or humanities	1.4% Performing or fine arts	1.4% Performing or fine arts	3.4% Government
9	2% Government	2.9% Education or humanities	3.2% Sex or pornography	1.4% Education or humanities	1.2% Education or humanities	2.1% Education or humanities
10	1.1% Performing or fine arts	2.7% Government	0.8% Performing or fine arts	1.4% Government	0.2% Sex or pornography	1.2% Performing or fine arts
	11.3% Unknown	0.6% Unknown	21.6% Unknown or other	21.0% Unknown or other	18.2% Unknown or other	20.1% Unknown or other

2001 Excite and FAST data are from Spink et al. [11]. Sex and pornography related queries are filtered out in the Seekport data set.

The topics show a highly skewed distribution, but all topics are represented within the query sets. Even the least requested topic ("fine arts") accounts for every hundredth query.

The amount of queries for "Sex or pornography" is around 4.5 percent on average, but we found that Seekport filters out such queries in its "Live Search". So the total amount should be higher, but is quite different for the observed search engines. In Metager 3.2 percent of all queries are on this topic, while in Fireball they account for 10.2 percent.

Regarding research question (2) (the comparison with other studies), it is interesting to see that the top category is "Commerce, travel, employment, or economy" for all data sets except the FAST 2001 data set (the "German" data). The same holds true for the top 2 position except for the Fireball data set, where "Sex or pornography" accounts for more than 10 percent of all queries. For the discussion, we used the rankings instead of absolute numbers because of the relatively high amount of queries classified as "unknown" in the 2005 data sets.

This relative similarity between the Excite (U.S.) data set and the 2005 (German) data sets is interesting because there seems to be a greater similarity between the actual German data sets with the U.S. set than the European set (see the columns on Excite and the average of German data sets in Table 1). But if one views the results from other data sets (from the U.S. and Europe as well) in [3], one can see large differences from data set to data set, regardless whether they are from U.S. or European search engines. The differences between continental sets seem as great as the differences between the sets from different continents.

5.2. Query types

Our next research question (3) was how the queries were distributed amongst the three query types. As expected, the informational queries form the largest group (with 45 percent of all queries). Nearly as much queries are navigational with an average of 40 percent. With the transactional queries, we find larger deviations between the search engines. They account for about 11 to 18 percent. The exact results for the different engines can be found in Table 2.

We find that all query types account for a noteworthy portion of all queries. But our study shows significant discrepancies to the results of Broder's study from 2002. While informational queries account for nearly the same amount of all queries, there are significant discrepancies with the navigational and transactional queries. We found a much larger proportion of navigational but less transactional queries.

We cannot say for sure why this is the case. It could result from changed user behaviour since Broder's study, but also from the specific behaviour of German Web search engine users or from differences between the AltaVista users and the users of other search engines. And finally, the smaller amount of transactional queries could result from the increasing use of special search engines for product-related searches (e.g., froogle.com).

Table 2
Distribution of query types in the different search engines

Data set	Informational	Navigational	Transactional
Fireball 2005	47%	35%	18%
Metager 2005	42%	43%	15%
Seekport 2005	47%	42%	11%
Average 2005	45%	40%	15%
AltaVista 2002 [1]	39–48%	20–24.5%	22–36%
AltaVista 2004 [8]	61–63%	11–15%	21–27%

5.3. Query types within topic areas

Research question (4) enquired for the combination of query types and query topics. The distribution of query types within the different topic areas is shown in Fig. 1. A large amount of informational queries (more than 70 percent) can be found in the topic areas

- Health or sciences,
- Society, culture, ethnicity, or religion.

For navigational queries, the largest groups (with more than 40 percent of all queries) are

- People, places and things,
- Commerce, travel, employment, or economy,
- Entertainment or recreation,
- Government,
- Performing or fine arts.

Navigational queries account for a notably high amount (more than 25 percent of all queries) for

- Computers or Internet,
- Commerce, travel, employment, or economy,
- Sex or pornography,
- Entertainment or recreation.

In the "Computers or Internet" category, user mainly want to download a specific software. In the "Sex or pornography" category, as well as in the "Entertainment or recreation" category, transactional queries

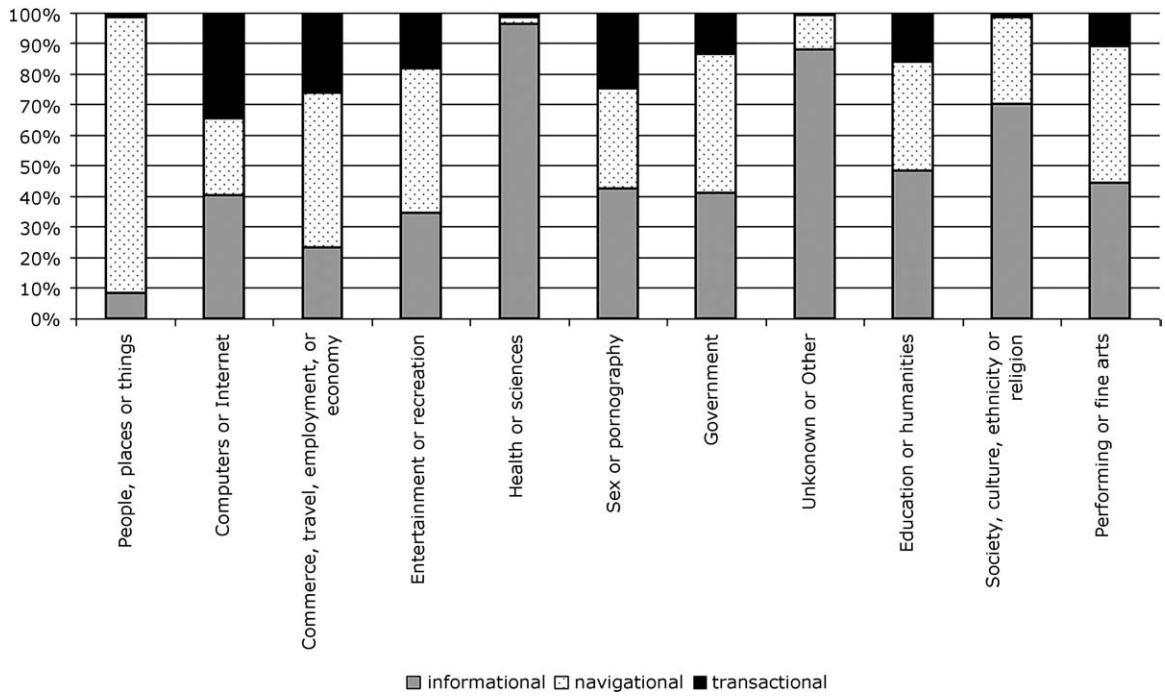

Fig. 1. Distribution of query types within topic areas.

are to a large amount for the download of movies, but in entertainment, also for music and other audio files.

6. Discussion and conclusion

In this study, we found that there is a specific behaviour of German Web search engine users, which (for our data sets) is comparable to one set of queries from U.S. users, but differs largely from others. Whether this is time or search-engine-dependent cannot be inferred from our investigation.

The percentages of individual topics in our study differs from the older studies mentioned, as well as the amounts of certain query types do. But we find that for topics as well as for types, all possible categories account for a certain amount of queries.

There are significant differences in the ratio of the several query types. In our investigation, we were not able to explicitly find the reasons why. This requires further research that directly compares German users with US users or with users from other countries.

Our research is a good expansion of the described studies on search topics and query types. The combination of both approaches is promising: We were able to show that the distribution of topics within query types is unequal. Therefore, search engines should focus on enabling better searching through the use of different modalities for different topics.

In further studies, the intents of the users within the different topic areas should be compared to the actual performance of search engines for such queries. With regard to the methods used, the classification for the search topics should be improved, as we found it sometimes difficult to assign queries to the topic categories.

References

[1] A. Broder, A taxonomy of web search, *SIGIR Forum* **36**(2) (2002), 3–10.
[2] C. Hölscher, Die Rolle des Wissens im Internet. Gezielt suchen und kompetent auswählen, Klett-Cotta, Stuttgart, 2002.
[3] B.J. Jansen and A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, *Information Processing & Management* **42**(1) (2006), 248–263.
[4] I.-H. Kang and G. Kim, *Query Type Classification for Web Document Retrieval, SIGIR'03*, ACM Press, Toronto, Canada, 2003, pp. 64-71.
[5] D. Lewandowski, Web searching, search engines and Information Retrieval, *Information Services and Use* **18**(3) (2005), 137–147.
[6] M. Machill, C. Neuberger, W. Schweiger and W. Wirth, Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen, in: *Wegweiser im Netz*, M. Machill and C. Welp, eds, Bertelsmann Stiftung, Gütersloh, 2003.
[7] M. Machill, C. Neuberger, W. Schweiger and W. Wirth, Navigating the Internet: A study of German-language search engines, *European Journal of Communication* **19**(3) (2004), 321–347.
[8] D.E. Rose and D. Levinson, Understanding user goals in Web search, in: *Thirteenth International World Wide Web Conference Proceedings, WWW2004*, 2004, pp. 13–19.
[9] N. Schmidt-Maenz and M. Koch, Patterns in search queries, in: *Data Analysis and Decision Support*, D. Baier, R. Decker and L. Schmidt-Thieme, eds, Springer, Heidelberg, 2005, pp. 122–129.
[10] N. Schmidt-Maenz and M. Koch, A general classification of (search) queries and terms, in: *3rd International Conference on Information Technologies: Next Generations*, Las Vegas, NV, USA, 2006.
[11] A. Spink, S. Ozmutlu, J.C. Ozmutlu and B.J. Jansen, U.S. versus European Web searching trends, *SIGIR Forum* **36**(2) (2002), 32–38.
[12] A. Spink, D. Wolfram, B.J. Jansen and T. Saracevic, Searching the Web: The public and their queries, *Journal of the American Society for Information Science and Technology* **53**(3) (2001), 226–234.