Web Searching: A Quality Measurement Perspective

to appear in: Zimmer, M.; Spink, A. (eds.): Web Search: Interdisciplinary perspectives. Dordrecht: Spinger, 2007.

Dirk Lewandowski

Hamburg University of Applied Sciences, Department Information, Berliner Tor 5, D – 20099 Hamburg, Germany. E-Mail: dirk.lewandowski@bui.haw-hamburg.de

Nadine Höchstötter

Institute for Decision Theory and Management Science, Universitaet Karlsruhe (TH), Kaiserstrasse 12, D - 76128 Karlsruhe, Germany. E-mail: nsh@topicflux.de

Abstract

The purpose of this paper is to describe various quality measures for search engines and to ask whether these are suitable. We especially focus on user needs and their use of web search engines. The paper presents an extensive literature review and a first quality measurement model, as well. Findings include that search engine quality can not be measured by just retrieval effectiveness (the quality of the results), but should also consider index quality, the quality of the search features and search engine usability. For each of these sections, empirical results from studies conducted in the past, as well as from our own research are presented. These results have implications for the evaluation of search engines and for the development of better search systems that give the user the best possible search experience.

Introduction

Web search engines have become important for information seeking in many different contexts (e.g., personal, business, and scientific). Research questions not answered satisfactorily are, as of now, how well these engines perform regarding user expectations and what measures should be used to get an overall picture of search engine quality. It is well known that search engine quality in its entirety cannot be measured with the use of traditional retrieval measures. But the development of new, search engine specific measures, as proposed in Vaughan (2004) are not sufficient, either. Search engine quality must be defined more extensively and integrate factors beyond retrieval performance such as index quality and the quality of the search features.

One aspect neglected is the user himself. But to discuss and judge the quality of search engines, it is important to focus on the user of such systems, too. A better performance of ranking algorithms or providing additional services do not always lead to users' satisfaction and to better search results. We will focus on the search engine user behaviour to derive strategies to measure search engine quality.

Additionally, quality assurance is an important aspect to improve customer satisfaction and loyalty. This is fundamental to protect market shares and revenues from adverts. Furthermore, quality measurement helps to identify potential improvements of search engines.

We are sure that only an integrated approach to quality measurement can lead to results usable for the development of better search engines. As with information retrieval, in general, we find a paradigm shift from the more technical (document-oriented) perspective to the user-oriented perspective (cf. Ingwersen & Järvelin, 2005). Our goal in this chapter is to define the scope of our perspective in comparison to other approaches and to give a literature overview of quality measurements for search engines. We will also focus on each individual factor stated in studies dealing with user interaction with search engines and user expectations to search engines. The integrated approach of user and technical aspects shows that there are many possibilities but they are not widely adopted yet.

Our chapter first gives an overview of studies conducted to derive quality measures and to present the state of the art. The other focus in this section lies on user surveys and analyses to give an anticipation of what users really do by placing search queries. In section 3 we give a general conspectus of parameters we deduced from our literature research and explain them shortly. In section 4 we show empirical results that reflect the current quality standard by our individual measures of search engines. In the last

section we summarize our findings and give potential strategies to improve search engines.

Many of the empirical findings stem from our own research conducted over the past years. Our integrated view on search engine quality measurement is reflected by the different research areas of the authors.

Related studies

In this section, we will discuss studies dealing with search engines in the given context. The two areas relevant for extensive search engine quality measurement are the concept of information quality in general and its transfer to search engines as a technical background, and user studies to see what happens at the front-end. Each will be discussed under a separate heading.

Search engine quality

Referring to information quality, one usually appraises information on the basis of a single document or a set of documents. Two perspectives have to be differentiated: Firstly, information quality in the production of a database which means, how documents or sources have to be appropriately selected and secondly, information quality of the results retrieved by a certain IR system.

While the latter can be easily applied to Web search engines, the assurance of the quality of databases is more difficult. The approach of the major search engines is to index not only a part of the Web, but as much as possible (or as much as reasonable under economic aspects). Only certain fractions of the Web (such as Spam sites) should be willingly omitted from the database. While in the production of databases the process of selecting documents (or sources of documents) can be seen as an important quality aspect, in the context of search engines, this process is reassigned to the ranking process. Therefore, classic judgements for the selection of documents from a library context do not fit to search engines. Only specialized search engines rely on a selection of quality sources (Web sites or servers) for building their indices.

An important point is that quality measurement of search results give only limited insight into the reliability and correctness of the information presented in the document. Popular examples are documents from Wikipedia, which are often highly ranked by search engines. But there seems not to be an agreement of experts whether Wikipedia content is trustworthy or

not. For a normal user, there is only a limited chance of scrutinising these documents. In this context, perceived information quality is more a matter of trust. Within the wider context of search engine evaluation, it is possible to build models completely based on trust (Wang, Xie, & Goh, 1999), as explained later on.

When discussing quality of search results, one should also keep in mind how search engines determine relevance. They mainly focus on popularity (or *authority*) rather than on what is commonly regarded as quality. It should be emphasized that in the process of selecting documents to be indexed by engines and in the ranking process as well, no human reviews are involved. But a certain bias can be found inherent in the ranking algorithms (Lewandowski, 2004b). These rate Web pages (apart from classic IR calculations) mainly by determining their popularity based on the link structure of the Web. The basic assumption is that a link to a page is a vote for that page. But not all links should be counted the same; link-based measures take into account the popularity of the linking page itself and the number of outgoing links, as well. This holds true for both of the main link-based ranking algorithms, Google's PageRank (Page, Brin, Motwani, & Winograd, 1998) and HITS (Kleinberg, 1999).

Table 1. Query-independent ranking factors (taken from Lewandowski, 2005a)

Directory hierarchy	Documents on a higher hierarchy level are preferred.
Number of incoming links	The higher the number of incoming links, the more important the document.
Link popularity	Quality/authority of a document is measured according to its linking within the Web graph.
Click popularity	Documents visited by many users are preferred.
Up-to-dateness	Current documents are preferred to older documents.
Document length	Documents within a sudden length range are preferred.
File format	Documents written in standard HTML are preferred to documents in other formats such as PDF or DOC.
Size of the Website	Documents from larger Web sites (or within a sudden size range) are preferred.

Link-based measures are commonly calculated query-independent, i.e., no computing power is needed to calculate these measures at the moment users place their search queries. Therefore, these measures can be applied very fast by the ranking process. Other query-independent factors are used as well (see table 1 and for a detailed discussion Lewandowski, 2005a).

Here, the important point is that the process of ranking Web pages evolved from a query-document matching, based on term frequency and similar factors, to a process where several quality measurements are also taken into account.

Link-based algorithms are of good use to push some highly relevant results to the top of the results list. This approach is oriented towards the typical user behaviour.

Users often view only a few results from the top of the list and seldom process to the second or even third page of the results list. Another problem with the calculation of appropriate result lists is the shortness of search queries. Therefore, most ranking algorithms prefer popular pages and the presence of search terms in anchor texts. Although the general user rarely uses advanced search features, this does not make them unnecessary or useless. On the one hand, there are special user groups like librarians or information professionals who conduct complex searches. On the other hand, while there is a majority of queries that can be successfully formulated without the use of advanced search syntax, one knows from his or her own searching behaviour that at least *sometimes* one needs to use operators or other advanced features. Users who have some background in the field they are searching use more often phrase searches. Users who know how search engines work also apply operators and phrase search more frequently.

With a reasonable amount of search features users are able to influence their search queries and with that the quality of returned results. When the user is able to construct more complex queries, it will be easier for the engine to return relevant pages. A discussion of features provided by different search engines can be found in Lewandowski (2004a). The topic will be discussed later in detail.

User perspective

There are two main empirical directions regarding user perspectives. One direction is represented by laboratory studies and surveys or by a combination of both. The other direction stands for the analysis of search engine transaction logs or the examination of live tickers published by search engines. Some search engines have a 'live ticker' or 'live search' enabling one to see the current search queries of other users (e.g., http.//www.lycos.de/suche/livesuche.html). This possibility is also often called 'spy function'. We will give a short overview of both regarding user behaviour to derive parameters for quality measurement. Table 2 shows the advantages and disadvantages of the different methods mentioned.

Table 2. Methods for obtaining data on search engine users' behaviour

Method	Advantages	Disadvantages
User survey	Users express themselves, demographics are available, detailed questions are possible	Users lie, they try to "look better", dependent on formulation of queries and interviewer (if present)
Laboratory studies	Detailed interactions are observable, often combined with a user survey for demographics	Very small samples, expensive, time consuming, not representative
Live ticker inquiry	Large samples of search queries, special search feature usage is also available, time-dependent analysis of search queries	No information about sessions (reformulation, topic changes, search queries per session), no demographics
Transaction log analysis	Detailed information about searching behaviour by search session analysis, time-dependent analysis of search queries	No demographics, data set is often tampered by robots

In surveys, users are sometimes directly asked which disturbing factors they notice by using Internet search engines. They also give a subjective view from the users perspective on what special search features and other offers they use and know in search engine interfaces. Another possibility is to ask questions about their knowledge of the functionality of search engines, since users with different knowledge levels show a different searching behaviour (Schmidt-Maenz & Bomhardt, 2005). In most cases, laboratory studies are only based on small samples and are for that reason not representative. It is also possible that subjects feel observed and try to search in a more professional way by using more operators or search features. One of the best and most representative ways to get user data is the analysis of transaction logs or data collected in live tickers. The problem is that there is no additional knowledge of the user himself.

The study of Machill, Neuberger, Schweiger, & Wirth (2003) consists of two parts, namely a telephone survey with 1000 participants and a laboratory study with 150 subjects. They show in their survey that 14 percent of search engine users definitely use advanced search features. Only 33 percent of respondents know that it is possible to personalize search engine interfaces. The title and the description of recommended Web pages are very important for users to evaluate the result lists. Users dislike results that have nothing in common with the search query submitted before (44 percent). Another 36 percent decline so-called dead links. Machill et al.

(2003) concluded their results with the remark that search engine users want their searches to be rewarded with success, a quick presentation of results, and clearly designed result screens. Hoelscher & Strube (2000) showed that experts and newbies show different searching behaviour. Hotchkiss found different groups of searching behaviour regarding the proceedings of the examination of result screens. Furthermore, users prefer organic results to sponsored listings.

Analyses of search engine transaction logs show a similar searching behaviour. Table 3 gives an overview. Most studies were based on the Excite search engine (Jansen, 2000; Spink, Jansen, & Ozmutlu, 2000; Spink, Wolfram, Jansen, & Saracevic, 2001; Spink & Jansen, 2004 and Spink, Ozmutlu, Ozmutlu, & Jansen, 2002). Others are conducted using logs from Altavista (Silverstein, Henzinger, Marais, & Moricz, 1999 and Beitzel, Jensen, Chowdhury, Grossman, & Frieder, 2004), and Alltheweb (Jansen & Spink, 2003 and Jansen & Spink, 2006). One log was obtained by a Spanish search engine BIWE (Buscador en Internet para la Web en Español (Cacheda & Viña, 2001)). Hoelscher & Strube (2000) analyzed a query log of Fireball, a German search engine. Zien, Meyer, Tomlin, & Liu (2000) observed the Webcrawler live ticker over a 66 days period. The year and length of observation period is given in table 3. Additionally, we extract most important results to get the users' perspective such as the number of search queries and the average length of search queries. We also analyse the percentage of complex search queries and in particular the percentage of phrase search, and the percentage of search sessions where only the first result screen is evaluated, too.

It is obvious that search queries are very short. Secondly, a remarkable part of search queries consist of only one term. With some exceptions the usage of Boolean operators is very small. The usage of phrase search is one of the most common ways to narrow search queries. Users commonly only examine the first result screen. These facts demonstrate that search engine users formulate their queries very intuitively and they do not try hard to evaluate every result in the list. The first two Excite studies (Excite 1 and Excite 2) and the BIWE log reveal that only a few users use special search features. This portion is 0.1 percent (Excite 1), 9.7 percent (Excite 2), and 0.2 percent (BIWE).

Table 3. Overview of studies based on logs and some results

Search engine	Excite 1	Excite 2	Fireball	Altavista 1	Excite 3	Web-crawler	BIWE	Allthe web 1	Excite 4	Allthewe b 2	Altavista 2
Year of observation	1997	1997	1998	1998	1999	2000	2000	2001	2001	2002	2002
Length of obeservation period	1	1	31	43	1	66	16	1	1	1	1
Number Search Queries	51,473	1,025,908	16,252,902	993,208,159	1,025,910	50,538,653	105,786	451,551	1,025,910	957,303	1,073,388
Average length of queries	38750	38750	38899	38809	38809	38779	38869	38809	38870	38778	38962
Percentage of one term queries	-	62.6%	-	25.8%	29.8%	22.5%	-	25.0%	29.6%	33.0%	20.4%
Complex queries	15.9%	9.3%	2.6%	20.4%	10.9%	35.6%	8.6%	4.3%	11.3%	4.6%	27.3%
Phrase Search	6.0%	5.1%	-	-	5.9%	10.4%	3.6%	0.0%	5.9%	0.0%	12.1%
Only 1st result screen (%)	58.0%	66.3%	-	85.2%	69.9%	-	67.9%	54.1%	84.6%	76.3%	72.8%

These extractions from user surveys and studies show that search engine users definitely have factors which disturb them and that they do not adopt all offered services such as special search features, possibilities to personalize search engines, or operators. Surveys are a good way to ask the user directly what he likes or dislikes while interacting with search engines. But surveys can become problematical when users get the illusion of a perfect search engine. For that reason the interpretation of search engine transactions logs is a objective way to see defective and non-adopted features or services. This helps to derive strategies for a user-friendly design or to design services that will be adopted by the user. With this in mind, we will give examples of interaction points between the user and search engines that could cause users' disconfirmation. Additionally, we give examples of how to evaluate these interaction points and already realized improvements.

Search engine quality measurement

In this section, we focus on the quality indicators for search engines. We are aware of the fact that more factors exist than we describe in each subsection. But we regard the selected factors as the most important ones. Other factors could be considered in further studies while they are omitted, here, for the clarity of the overview.

Retrieval measures

Retrieval measures are used to measure the performance of IR systems and to compare them to one another. The main goal for search engine evaluation is to develop individual measures (or a set of measures) that are useful for describing the quality of search engines. Retrieval measures have been developed for some 50 years. We will give an overview of the main retrieval measures used in IR evaluation. It will be shown that these measures can also be used for search engine evaluation, but are of limited use in this context. Therefore, web-specific retrieval measures were developed. But a set of measures that can be used for getting a complete picture of the quality of search engines is still missing.

General retrieval measures

The retrieval performance of the IR system is usually measured by the "two classics", precision and recall.

Precision measures the ability of an IR system to produce only relevant results. Precision is the ratio between the number of relevant documents retrieved by the system and the total number of documents retrieved. An ideal system would produce a precision score of 1, i.e. every document retrieved by the system is judged relevant.

Precision is relatively easy to calculate, which mainly accounts for its popularity. But a problem with precision in the search engine context is the number of results usually given back in response to typical queries. In many cases, search engines return thousands of results. In an evaluation scenario, it is not feasible to judge so many results. Therefore, cut-off rates (e.g. 20 for the first 20 hits) are used in retrieval tests.

The other popular measure, the so-called recall, measures the ability of an IR system to find the complete set of relevant results from a collection of documents. Recall is the ratio of the number of relevant documents retrieved by the system to the total number of relevant documents for the given query. In the search engine context the total number of relevant documents refers to all relevant documents on the Web. As one can easily see, recall cannot be measured, in this context. A proposed solution for this problem is the method of pooling results from different engines and then measuring the relative recall of each engine.

Precision and recall are not mathematically dependent on each other, but as a rule of thumb, the higher the precision of a result set, the lower the recall and vice versa. For example, a system only retrieving one relevant result receives a precision score of 1, but usually a low recall. Another system that returns the complete database as a result (maybe thousands or even millions of documents) will get the highest recall but a very low precision.

Other "classic" retrieval measures are fallout and generality (for a good overview of retrieval measures see Korfhage, 1997). Newer approaches to measure the goodness of search results are

- Median Measure (Greisdorf & Spink, 2001), which takes into account the total number of results retrieved. With median measure, it cannot only be measured how positive the given results are, but also how positive they are in relation to all negative results.
- Importance of completeness of search results and Importance of precision of the search to the user (Su, 1998). These two measures try to employ typical user needs into the evaluation process. It is taken into ac-

count whether the user just needs a few precise results or maybe a complete result set (while accepting a lower precision rate). For the purpose of search engine evaluation that focuses on the user, these two measures seem highly promising.
- Value of Search Results as a Whole (Su, 1998), which is a measure that seems to correlate well with other retrieval measures regarded as important. Therefore, it can be used to shorten the evaluation process and make it less time and cost consuming.

In the information science community, there is an ongoing and lively debate on the best retrieval measures. But unfortunately, there is a lack of current and continuous evaluation of search engines testing different measures.

Web-specific retrieval measures

Quite early in the history of search engines, it became obvious that for the evaluation of these systems, Web-specific retrieval measures should be applied. In this section, we present the most important ones. They all have in common that they are used in experimental research and they are not widely used in real evaluations. Some empirical tests were applied in the development of these measures, but there are no larger evaluations, yet, that compare their use to that of other measures.

- **Salience** is the sum of ratings for all hits for *each* service out of the sum of ratings for *all* services investigated (Ding & Marchionini, 1996). This measure takes into account how well all search engines studied perform on a certain query.
- **Relevance concentration** measures the number of items with ratings of 4 or 5 [from a five-point relevance scale] in the first 10 or 20 hits (Ding & Marchionini, 1996).
- **CBC ratio** (MacCall & Cleveland, 1999) measures the number of content-bearing clicks (CBC) in relation to the number of other clicks in the search process. A CBC is "any hypertext click that is used to retrieve possibly relevant information as opposed to a hypertext click that is used for other reasons, such as the 'search' click that begins a database search or a 'navigation' click that is used to traverse a WWW-based information resource" (p. 764).
- **Quality of result ranking** takes into account the correlation between search engine ranking and human ranking (Vaughan, 2004), p. 681).
- **Ability to retrieve top ranked pages** combines the results retrieved by all search engines considered and lets them be ranked by humans. The "ability to retrieve top ranked pages" measures the ratio of the top 75

percent of documents in the results list of a certain search engine (Vaughan, 2004).

But every quality measurement dealing with web-specific retrieval measures has to be combined with user strategies. In reality, users only examine the first result screens (see table 3), they do not even use search features or operators to really interact with search engines. (Hotchkiss et al., 2004) defined different search types. The normal search engine user corresponds to the "Scan and Clickers". They only watch the top results, sometimes also paid listings. They decide very quickly to visit a page after reading the short description texts and URLs. Machill et al. (2003) also observe subjects who try to get good answers after very short questions. Regarding these annotations, it is important to think about retrieval measures that deal with this user specific searching behaviour. If a user always watched the first three results, only, the best search engine would be the one returning the most appropriate pages within those first results. How do retrieval measures comply with the search engine users' search strategies?

Towards a framework for search engine quality

As already can be seen from the web-specific retrieval measures, search engine quality goes well beyond the pure classification of results in relevant or non-relevant ones. The relevance judgements may be the most important point in the evaluation of search engines, but surely not the only one.

A framework for measuring search engine quality was proposed in Xie, Wang, & Goh (1998) and further developed in Wang et al. (1999). The authors base their model on the application of the SERVQUAL (Service and Quality) model (Parasuraman, Zeithaml, & Berry, 1988) on search engines. As this is a completely user-centred model, only the *user perceived* quality can be measured. The authors apply gap analysis to make a comparison between expectations and perceived performance of the search engines, but do not weight the factors observed.

The model clearly lacks the system centred model of IR evaluation. It is interesting to see that according to this investigation, one of the main points in search engine evaluation ("Search results are relevant to the query") does not differ greatly from engine to engine.

Contrary to such user-centred approaches is the "classic" system approach, which tries to measure the performance of information retrieval systems from a more "objective" point of view. Saracevic (1995) divides

the evaluation of IR systems into two broad categories with three levels each:

- System-centred evaluation levels: Engineering level (e.g., hardware or software performance), input level (coverage of the designated area), and processing level (e.g., performance of algorithms).
- User-centred evaluation levels: Output level (interaction with the system, feedback), use and user level (where questions of application to given problems and tasks are raised), and social level (which takes into account the impact on the environment).

Saracevic concludes that results from one level of evaluation do not say anything about the performance of the same system on the other levels of evaluation and that "this isolation of levels of evaluation could be considered a basic shortcoming of all IR evaluations" (p. 141).

In our opinion, this also applies to the evaluation of search engines. Only a combination of both, system and user-centred approach can lead to a clearer picture of the overall search engine quality.

There are several points of contact between users and search engines that can cause user discontent. The first and obvious point is the front-end of search engines. Next will be additional services that should help users to perform their search sessions. As shown above, special search features, personalization possibilities and operator usage are possible to control over transaction logs. Geoghegan (2004) gives five measures to compare search engine usability. He compares five major search engines by relevance of results, speed of result list calculation, the look of the input window and result list, and the performance of results based on a natural question. We suggest four main measures to check search engine quality out of the users' perspective.

- Interface design: structure of search engine Web pages and the presentation of the results. The input window should be structured in a clear way without overwhelming advertising. The result lists have to separate organic results from sponsored links. A different colour will be helpful.
- Acceptance of search features and operators: Which functions are accepted by users? Do they use operators? Do users personalize their preferred search engine?
- Performance of search engines: The speediness of result list presentation is one important point. Also intuitive and very short search queries should yield serious results. So-called dead links and spam have to be avoided.
- User guidance: Newbies need help to formulate adequate search queries, phrase searches, or complex searches. It is also helpful to give users

some hints how search features work and what to do with them. A short introduction in search engine technology is recommended, too.

Taking both into account, the system approach and the user-centred approach, we propose another quality framework that considers more objective measures as well as the user perspective. Therefore, we expand the quality framework first proposed in Lewandowski (2006c) to four sections as follows:

- Index Quality: This points out the importance of the search engines' databases for retrieving relevant and comprehensive results. Measures applied in this section include Web coverage, country bias, and up-to-dateness.
- Quality of the results: This is the part where derivates of classic retrieval tests are applied. As can be seen from the discussion on retrieval measures above, it should be asked which measures should be applied and if new measures are needed to satisfy the unique character of the search engines and their users. An additional measure that should be applied is, for example, the uniqueness of search results in comparison to other search engines. It is worth mentioning that users are pretty satisfied by finding what they search for. The subjects in the laboratory study conducted by Machill et al. (2003) admit that they are very pleased with search results and also with their favorite search engine. In the survey conducted by Schmidt-Maenz & Bomhardt (2005), 43.0 percent of 6723 respondents very often found what they wanted and another 50.1 percent often. The question is if users could really evaluate the quality of results. Users are not able to compare all recommended web pages. Sometimes 1,000,000 results are listed. It is more probable that they only think they find what they want since they do not even know what they could find in other results.
- Quality of search features: A good set of search features (such as advanced search), and a sophisticated query language is offered and works reliable.
- Search engine usability: This gives a feedback of user behaviour and is evaluated by user surveys or transaction log analyses. This will give comparable parameters concerning interface design. Is it possible for users to interact with search engines in an efficient and effective way? Is the number of search queries and of reformulations in different search engines lower? It is also of importance which features are given to assist users regardless if they are beginners or professionals in using search engines. Users search in a very intuitive way (Schmidt-Maenz & Koch, 2006).

All in all, the user should feel comfortable using search engines. Since users currently have not developed all necessary skills to handle search engines in the best way their usage should be intuitive and simple. In addition, users should get every support whenever it is useful or required. It has to be possible that users enhance their searching behaviour by using additional services and features to get the best recommendations of web pages as possible.

Empirical results

In this section, we will present studies dealing with search engine quality and the behaviour of search engines users. The combination of these two research areas shows that there is a research gap in the user-centred evaluation of search engines. While there are a lot of studies dealing with single points, there is no study (or series of studies) focussing on an overall picture of search engine quality from the user perspective.

Index quality

Search engines are unique in the way they build up their databases. While traditional IR systems are usually based on databases manually built by human indexers from selected sources (e.g., from journals or books within a certain subject area), search engines have to make use of the link structure of the Web to find their documents by crawling it. It is a big challenge to build up and maintain an index generated by Web robots. A good overview is given in Risvik & Michelsen (2002).

The quality of the index of an individual search engine can be regarded in several ways. At first, the index should be comprehensive (i.e. cover a large portion of the Web). While the overall comprehensiveness is important, a search engine with a good overall coverage is not necessarily the best for every area of the Web. For example, a user searching for German language content will not be satisfied if the search engine with a general Web coverage of maybe 80 percent does not include German documents at all or just to a small degree. Therefore, country bias in search engine databases is an important point in research.

The third important index quality factor is the up-to-dateness of the databases. The Web is in constant flux, new documents are added, older documents disappear and other documents change in content. As can be seen from Schmidt-Maenz & Koch (2006), to a large amount, users search for current events and actual news stories. In addition, the number of in-

coming links changes in a similar manner. Web pages concerning a current topic will achieve more incoming links, when this page is of importance. When the event will not longer be of interest anymore, the number of inbounds decreases again (Schmidt-Maenz & Gaul, 2005). Such queries (to give one example) can only be "answered" by search engines with an up-to-date index.

Index sizes and Web coverage

An ideal search engine would keep a complete copy of the Web in its database. But for various reasons, this is impossible. Many searches return lots of results, often thousands or even millions. Keeping this in mind, one could ask why a search engine should take the effort to build indices as large as possible and not just smaller ones that would fit the general users' queries.

A large index is needed for two purposes. The first case is when the user wants a comprehensive list of results, e.g., to become familiar with a topic. The second case is obscure queries that produce just a few results. Here, the engine with a bigger index is likely to find more results.

In table 4 the distribution of search terms is listed. Independent of search engines observed most search queries appear only once. Around 60 percent of all unique search queries appeared only once. Regarding all search queries including their recurrences, only 7.9 percent appeared once. With this in mind, it is maybe not important to have the largest but the most specialized index. It is also of interest to have the possibility to calculate results for very specialized and seldom queries rather than for those that are very popular. We have to stress that users only view the first two or three pages. For popular search queries, it is sufficient to list the most popular pages on the first result page. Search engines like Google already prefer pages such as the ones from Wikipedia.

But the index sizes do not seem to be as important as reported for example in the general media. What makes them such a popular measure is the simplicity of comparison. But the mere sizes don't reveal that much about the quality of the index. A search engine could have, e.g., a large amount of spam pages in its index. Index size is just one measure that is only of importance in relation to other measures.

Search engine sizes are sometimes compared with one another on an absolute basis. But that says nothing about how big they are in relation to the total of the Web. Therefore, Web coverage should be taken into account. Studies dealing with the size of the Web often also investigate on the ratio covered by the search engine. Therefore, both types of studies are discussed together in this section.

Table 4. Appearance of search queries (Schmidt-Maenz & Koch, 2006)

ID		Search queries which appeared exactly…					
		once	twice	3 times	4 times	5 times	>5 times
Fireball	absolute	10,480,377	3,024,799	1,330,798	738,817	461,185	1,956,093
	Percentage of all SQ	7.9%	4.6%	3.0%	2.2%	1.7%	80.6%
	Percentage of unique	58.3%	16.8%	7.4%	4.1%	2.56%	10.9%
Lycos	absolute	17,618,682	4,727,513	2,022,780	1,124,878	773,026	3,055,487
	Percentage GN	9.3%	5.0%	3.2%	2.4%	2.1%	78.2%
	Percentage NN	60.1%	16.12%	6.9%	3.8%	2.6%	10.4%
Metaspin-ner	absolute	732,429	224,171	107,354	65,866	42,021	115,021
	Percentage GN	17.9%	11.0%	7.9%	6.4%	5.1%	51.7%
	Percentage NN	56.9%	17.4%	8.3%	5.1%	3.3%	9.0%

There are three ways to get numbers for the discussion about the size of the Web and search engine coverage:

- **Self-reported numbers**. Search engines sometimes report their index sizes to show that they increased in size and/or have the largest index.
- **Overlap measures**. Based on the overlap of search engines, the total size of the Web indexed by all search engines is measured. A limitation of this method is that it omits all pages that are found by none of the search engines under investigation.
- **Random sampling**. Random samples are taken and tested for availability. A total number of available Web pages is calculated from the sample and all pages available are tested against the search engines.

The following paragraphs will give an overview of the most important studies using the different methods.

A comparison based on the self-reported numbers can be found on the SearchEngineWatch.com Web site (Sullivan, 2005). The site offers information on the evolution of search engine sizes from the early days on until 2005. Unfortunately, the major search engines do not report their index sizes anymore. Furthermore, while such a comparison is nice to have, it does not say anything about the Web coverage of the indices. In addition, for such comparisons, one has to trust the search engines in giving the correct number. As some studies showed, self-reported numbers can be trusted from some search engines, while others are highly exaggerated (Lewandowski, 2005b).

The most important studies determining the Web size and the coverage by search engines on the basis of overlap are Bharat & Broder (1998) and Lawrence & Giles (1998).

Bharat & Broder use a crawl of a part of the Web to build a vocabulary from which queries are selected and sent to four major search engines. From each result set (with up to 100 hits), one page is selected at random. For each Web page found, a "strong query" is built. Such a "strong query" consists of eight terms that should describe the individual documents. These queries are sent to the search engines studied. Ideally, only one result should be retrieved for each query. But there could be more results for various reasons: The same page could be reached under different URLs, and there could be near-identical versions of the same page. The method proposed can deal with this problem and should find the page searched for even if it is indexed by one search engine under a different URL than in the other search engine. From all pages found, the authors calculate the coverage ratio for all search engines. The results show that search engines in general have a moderate coverage of the Web with the best engine in-

dexing 62 percent of the total of all pages, while the overlap of all engines is extremely low with just 1.4 percent at the end of 1997. Based on the data, the total size of the Web is estimated at 200 million pages.

The study from Lawrence & Giles (1999) is based on 575 queries from scientists at the NEC Research Institute. From the result sets, the intersection of two search engines under consideration is calculated. The total size of the Web is calculated based on the overlap between the known total index size of one search engine (HotBot with 110 million pages) and the search engine with the second-biggest index, AltaVista. The result is an estimate of the total size of the Web of 320 million pages and coverages of search engines from three to 34 percent.

While the total size estimates and the ratio of Web coverages differ in both studies presented, both show that (at least in 1997/1998) search engines were nowhere near complete coverage of the Web and that the overlap between the engines is rather small. This leads to the conclusion that meta search engines and/or the use of another search engine in case of failure could be useful.

The most current overlap study is from Gulli & Signorini (2005). They use an extended version of Bharat and Broder's methodology and find that the indexable Web in 2005 contains at least 11.5 billion pages. Search engine coverage of the data set (which consists of all pages found by at least one engine) lies between 57 to 76 percent for the four big search engines (Google, Yahoo, MSN, Ask).

The most prominent study using random sampling to determine the total size of the Web is the second study from Lawrence & Giles (1999). The basis is a set of random generated IP addresses which are tested for availability. For each of these IPs generally available, it is tested whether it is used by a public server (i.e., a server that hosts pages indexable by a search engine). Based on 3.6 million IP addresses, 2.8 million servers respond in the intended way. From these, 2500 are randomly chosen and their contents are crawled. From the average number of pages per server of 289, the authors determine the size of the indexable Web to about 800 million pages. Search engine coverage is tested with 1050 queries. NorthernLight, the search engine performing best, covers only 16 percent of the indexable Web. All engines under investigation cover only 42 percent.

All Web size and search engine coverage studies reported have in common that they focus on the indexable part of the Web, or *Surface Web*. But this is just a part of the Web in its entirety, the rest consisting of the so-called *Invisible Web* or *Deep Web*.

In short, the Invisible Web is the part of the web that search engines do not add to their indices. There are several reasons for this, mainly limited storage space and the inability to index certain kinds of content.

There are two main definitions of the Invisible Web, and in this chapter, we do not need to distinguish between the terms Invisible Web and the Deep Web. Both terms are widely used for the same concept and using one or the other is just a matter of preference. We use the established term Invisible Web.

Sherman and Price give the following definition for the Invisible Web: "Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages" (Sherman & Price, 2001), p. 57).

This is a relatively wide definition as it takes into account all file types and includes the *inability* of search engines to index certain content as well as their *choice* not to index certain types of contents. In this definition, for example, Spam pages are part of the Invisible Web because search engines choose not to add them to their indices.

Bergman (2001) defines the Invisible Web (or in his words, the Deep Web) much more narrowly, focusing on databases available via the web, he writes: "Traditional search engines cannot "see" or retrieve content in the deep Web – those pages do not exist until they are created dynamically as the result of a specific search."

Bergman estimates the size of the Invisible Web to be 550 times larger than the surface Web. Given that the size of the surface Web was estimated to one billion pages at the time the study was conducted, Bergman says the Deep Web consists of about 550 billion documents.

But, as Lewandowski & Mayr (2006) found, these size estimates are far too high, because of two fundamental errors. First the statistical error of using the mean instead of the median calculation and second his misleading projection from the database size in GB. When using the 85 billion documents from his Top 60 (which forms the basis of all further calculations), one can assume that the total number of documents will not exceed 100 billion because of the highly skewed distribution (for details, see Lewandowski & Mayr, 2006). Even though this estimation is based on data from 2001, it seems that the typical growth rate of database sizes (cf. Williams, 2005) will not affect the total size to a large extent.

Further research is needed for the distinction between the Visible and the Invisible Web. In the past years, we saw the conversion of large databases into HTML pages for the purpose of getting indexed in the main Web search engines. Although this is mainly done in the commercial con-

text, other vendors such as libraries followed this approach with varying degrees of success (cf. Lewandowski, 2006b). Further research on this topic is needed because today nobody knows to what extent database content is already available on the surface web.

The interest of the search engines in indexing the Invisible Web seems just moderate. There is an attempt from Yahoo to index parts of the commercial Invisible Web (Yahoo Subscriptions; http://search.yahoo.com/subscription) as well as some specialised search engines for Invisible Web content (e.g., http://turbo10.com/). But as of yet, no real integration of larger parts of IW content into general search engines was achieved.

Country bias

In the process of crawling the Web, there is a certain index due to the starting points chosen and the structure of the Web, as well. Highly linked pages have a better chance to be found by the engines than pages from the "periphery" of the Web. The Web was modelled as having a "bow-tie" structure by Broder et al. (2000). But pages in the centre of the Web (the "Strongly Connected Core") are of a higher probability to be older and – regarding the growth structure of the Web – from the U.S. (Vaughan & Thelwall, 2004).

But for users not from the U.S. it is important that content in their native languages and from their native countries can be found in the search engines. It is astonishing that there is (at least to our knowledge) just one study dealing with country bias. Especially in the European context with the many languages spoken across Europe, there should be a focus on this topic.

Vaughan & Thelwall (2004) ask for the coverage of Web sites from different countries in three major search engines. Countries investigated are the U.S.A., China, Singapore, and Taiwan. The countries are chosen in a way that it can be differentiated between bias due to language factors and "real" country bias. Selected sites both from the U.S. and from Singapore are in English, while sites both from China and Taiwan are in Chinese. The search engines chosen are Google, All the Web and AltaVista.

There are two main research questions: 1. What ratio of the Web sites is indexed in the search engines? 2. What ratio of documents within these Web sites is indexed by the search engines?

While the first question asks for the ratio of servers from a certain country known by a search engine, the second question asks how deep a certain search engines digs within the sites of a certain country.

All sites chosen for investigation are commercial sites from a random sample (based on IP numbers) from the chosen countries. A research crawler was used to index all sites as deeply as possible. Each page found was checked with the chosen search engines for availability in the indices.

The main result was that the coverage of the sites differs enormously between countries and search engines, as well. As expected, the U.S. sites received the best coverage with 80 to 87 percent according to the search engine. Sites from China had a coverage from 52 to 70 percent, while the ones from Singapore reached between 41 and 56 percent, and the ones from China between four and 75 percent.

There were large differences in the depth of indexing, too. From U.S. sites, on average, 89 percent of the pages were indexed, while this number was only 22 percent for China and only three percent for Taiwan.

Regarding these results, the assumption that Chinese language Web sites are not indexed as well as English language Web sites due to properties of the Chinese language must be rejected. The same low indexing ratio is shown for English language sites from Singapore. The authors come to the conclusion that disadvantage for these sites must stem from the link structure of the Web.

This study gives indication of a heavy country bias in the search engines indices. We see it as important that similar studies should be conducted because of two reasons: Firstly, the results are now some years old and it can only be guessed that they are still valid today. Secondly, a larger country basis should be investigated. Keeping in mind the discussion in Europe whether a genuine European search engine should be built in competition to the dominating U.S. search engines and the discussion about the usefulness of country-specific search engines, we see an urgent need for studies investigating the country bias for at least a selection of European countries.

Up-to-dateness

Up-to-dateness is a threefold problem for search engines. Firstly, up-to-dateness is important in keeping the index fresh. Secondly, up-to-dateness factors are used in the ranking of Web pages (Acharya et al., 2005; Lewandowski, 2006a). And thirdly, up-to-dateness factors could play an important role in Web based research (Lewandowski, 2004c). This section only deals with the first aspect, while the last one will be discussed later.

Ke, Deng, Ng, & Lee (2006) give a good overview of the problems for search engines resulting from Web dynamics. Crawling and indexing problems resulting from Web dynamics from a commercial search engine's point of can be found in Risvik & Michelsen (2002).

A study by Ntoulas, Cho, & Olston (2004) found that a large amount of Web pages is changing on a regular basis. Estimating the results of the study for the whole Web, the authors find that there are about 320 million new pages every week. About 20 percent of the Web pages of today will disappear within a year. About 50 percent of all contents will be changed within the same period. The link structure will change even faster: About 80 percent of all links will have changed or be new within a year. These results show how important it is for the search engines to keep their databases up to date.

But there are just two (series) of studies discussing the actual up-to-dateness behaviour of the major search engines.

Notess conducts studies on the average age of Web pages in the search engines' indices. In the latest instalment, Notess (2003) uses six queries to analyse the freshness of eight different search engines (MSN, HotBot, Google, AlltheWeb, AltaVista, Gigablast, Teoma, and Wisenut). Unfortunately the author gives no detailed information on how the queries were selected. For each query all URLs in the result list are analysed which meet the following criteria: First, they need to be updated daily. Second, they need to have the reported update information in their text. For every Web page, its age is put down. Results show the age of the newest page found, the age of the oldest page found and a rough average per search engine. In the most recent test (Notess, 2003), the bigger search engines such as MSN, HotBot, Google, AlltheWeb, and AltaVista have all some pages in their databases that are current or one day old. The databases of the smaller engines such as Gigablast, Teoma, and Wisenut contain pages that are quite older, at least 40 days.

When looking for the oldest pages, results differ a lot more and range from 51 days (MSN and HotBot) to 599 days (AlltheWeb). This shows that a regular update cycle of 30 days, as usually assumed for all the engines, is not used. All tested search engines have older pages in their databases.

For all search engines, a rough average in freshness is calculated, which ranges from four weeks to seven months. The bigger ones obtain an average of about one month except for AltaVista of which the index with an average of about three months is older.

Notess' studies have several shortcomings, which mainly lie in the insufficient disclosure of the methods. It is neither described how the queries are selected, nor how the rough averages were calculated. The methods used in the described study were used in several similar investigations from 2001 and 2002. Results show that search engines are performing better in indexing current pages, but they do not seem to be able to improve

their intervals for a complete update. All engines have quite outdated pages in their index.

Lewandowski, Wahlig, & Meyer-Bautor (2006) use a selection of 38 German language Web sites that are updated on a daily basis for their analysis of the update frequencies of the major Web search engines. Therefore, the cache copies of the pages were checked every day within a time span of six weeks. The search engines investigated were Google, Yahoo and MSN. Only sites that display their latest update date or another currently updated date information were used because Yahoo doesn't display the date the cache copy was taken.

The analysis is based on a total of 1558 results for every search engine. The authors measure how many of these records are no older than 1 or even 0 days. It was not possible to differentiate between these two values because the search engines were queried only once a day. If there had been a search engine that updated pages at a certain time of the day it would have been preferred to the others. Therefore, it was assumed that a page that was indexed yesterday or even today is up-to-date in the cache.

Google handed back most of the results with the value 1 (or 0). The total number of 1291 records shows that 82.86 percent of the Google results were no older than one day. MSN follows with 748 (48.01 percent). Yahoo contains 652 (41.85 percent) one or zero days old pages in its index.

Also, the arithmetic mean up-to-dateness of all web pages was calculated. Again, Google hands back the best results with an average age of 3.1 days, closely followed by MSN with 3.5 days and Yahoo is behind with 9.8 days. The use of the median instead of the arithmetic mean presents a different picture in which the competitors are closer together: Google and MSN have a median of 1 while Yahoo has a median of 4 days.

Another important point is the age of the oldest pages in the indices. While Google as well as Yahoo have several pages in their indices that were not updated for quite a long time, only MSN seems to be able to completely update its index within a time-span of less than 20 days. Since the research only focussed on Web pages that are updated on a daily basis, this cannot be proved for the complete index. Further research is needed to answer this question. But on the basis of the findings it can be conjectured that Google and Yahoo, which both have outdated pages in their indices, will perform even worse for pages that are not updated on a daily basis.

To summarise the findings, Google is the fastest search engine in terms of index quality, because many of the sites were updated daily. In some cases there are outliers that were not updated within the whole time of the research or show some noticeable breaks in their updating frequency. In contrast to that, MSN updates the index in a very clear frequency. Many of

the sites were updated very constantly. Taking a closer look at the results of Yahoo, it can be said that this engine has the worst update policy.

Retrieval effectiveness

As already mentioned before, there are several difficulties measuring retrieval effectiveness. The studies discussed below follow a system approach to evaluation. Therefore, the real user behaviour is not represented adequately in the settings. Users only use short search queries and place in average only 2.1 queries per session (Ozmutlu, Spink, & Ozmutlu, 2003). More than 40 percent of sessions exist only of one search query (Spink & Jansen, 2004). Machill et al. (2003) show that users only place search queries consisting of only one term and they are possibly as effective as users who formulate and reformulate longer and complex queries. In consideration of these facts, it is inevitable to measure retrieval effectiveness with user searching behaviour in mind.

Furthermore, the different query types used in search engines are not taken into account. From the now classic distinction between navigational, informational and transactional queries (Broder, 2002), usually, only informational queries are used for evaluation purposes.

According to Broder, with *informational queries*, users want to find information on a certain topic. Such queries usually lead to a set of results rather than 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 Web page 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. Navigational queries are usually answered by just one result; the informational need is satisfied as soon as this one right result is found.

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.

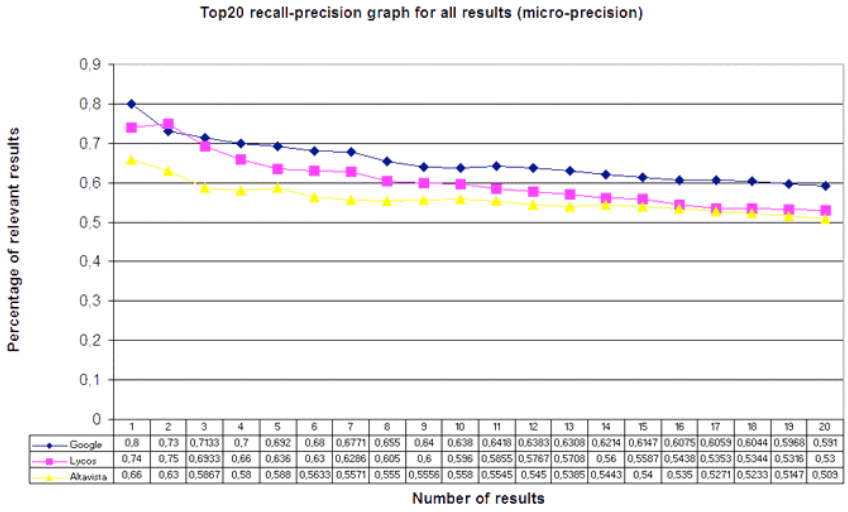
For the further discussion on retrieval tests, one should keep in mind that these only present results for a certain kind of queries, whereas the ranking approaches of some search engines are explicitly developed to better serve navigational queries (Brin & Page, 1998), also see Lewandowski (2004b).

With respect to quality of the results, there is a vast amount of literature on the evaluation of the retrieval effectiveness of search engines (e.g., Ford, Miller, & Moss, 2002; Griesbaum, Rittberger, & Bekavac, 2002; Leighton & Srivastava, 1999; Machill, Neuberger, Schweiger, & Wirth, 2004; Singhal & Kaszkiel, 2001; Wolff, 2000). Because of the constantly changing search engine landscape, older studies are mainly interesting for their methods, but provide only limited use in their results for the different search engines.

For the purpose of this chapter, we will discuss two newer studies (Griesbaum, 2004; Véronis, 2006), from which we will derive our demand for expanded tests on retrieval effectiveness. The most interesting result from these studies, in our opinion, is that the results of the different engines have converged within the last years. This supports our demand for a more extensive model for quality measurements.

Griesbaum's 2004 study (Griesbaum, 2004) continues research begun and uses methods developed in Griesbaum et al. (2002). Three search engines (Google, Lycos and AltaVista) are tested for relevance on a three-point scale. Results are judged either as relevant, not relevant or not relevant but leading (through a hyperlink) to a relevant document.

The study uses 50 queries and the first 20 results are judged for each query and search engine. Results show that the differences between the three engines investigated are quite low. Google reaches a mean average precision of 0.65, while Lycos reaches 0.60 and AltaVista 0.56, respectively. The complete precision-recall graph is plotted in fig. 1. These results are out-dated in that they do not describe the search engine landscape as of 2006. Major changes have occurred since the accomplishment of the study. But what the results clearly show is that the relevancy scores of the different engines tend to converge.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Google	0,8	0,73	0,7133	0,7	0,692	0,68	0,6771	0,655	0,64	0,638	0,6418	0,6383	0,6308	0,6214	0,6147	0,6075	0,6059	0,6044	0,5968	0,591
Lycos	0,74	0,75	0,6933	0,66	0,636	0,63	0,6286	0,605	0,6	0,596	0,5855	0,5767	0,5708	0,56	0,5587	0,5438	0,5353	0,5344	0,5316	0,53
Alkavista	0,66	0,63	0,5867	0,58	0,588	0,5633	0,5571	0,555	0,5556	0,558	0,5545	0,545	0,5385	0,5443	0,54	0,535	0,5271	0,5233	0,5147	0,509

Fig. 1. Top 20 recall-precision graph for all results (taken from Griesbaum, 2004)

Véronis (2006) measures the retrieval effectiveness of six search engines (Google, Yahoo, MSN, Exalead, Voila, Dir.com) as of December, 2005. Here, these queries concern 14 topic areas with five queries each selected by student evaluators. Results are limited to the French language. For each query and search engine, the first ten results are evaluated. A six-point relevance scale (from 0=worst to 5=best) is used and some additional criteria are recorded.

Results show that neither of the engines tested receives a good overall relevance score. The author concludes that "the overall grades are extremely low, with no search engine achieving the 'pass' grade of 2.5" (Véronis, 2006). The best search engines are Yahoo and Google (both 2.3), followed by MSN (2.0). The other (French) search engines perform worse with 1.8 for Exalead, 1.4 for Dir.com and 1.2 for Voila.

Looking at the relevance graph for the top 10 results (fig. 2 one finds confirmation for the convergence of the results at least of the three major search engines.

The convergence of the relevance based on the precision measure leads us to the conclusion that, at least, the major search engines perform comparable on standard informational search queries. Other query types were not tested in either of the studies reported.

We think that there are differences between the retrieval effectiveness of the different search engines. But it seems that the precision of the first X

results is not the best measure to compare search engines with one another. Therefore, retrieval tests applying other/new and web-specific measures should be developed. Unfortunately, such retrieval measures are only developed on an experimental basis (see above) and there is no larger initiative working on this topic yet.

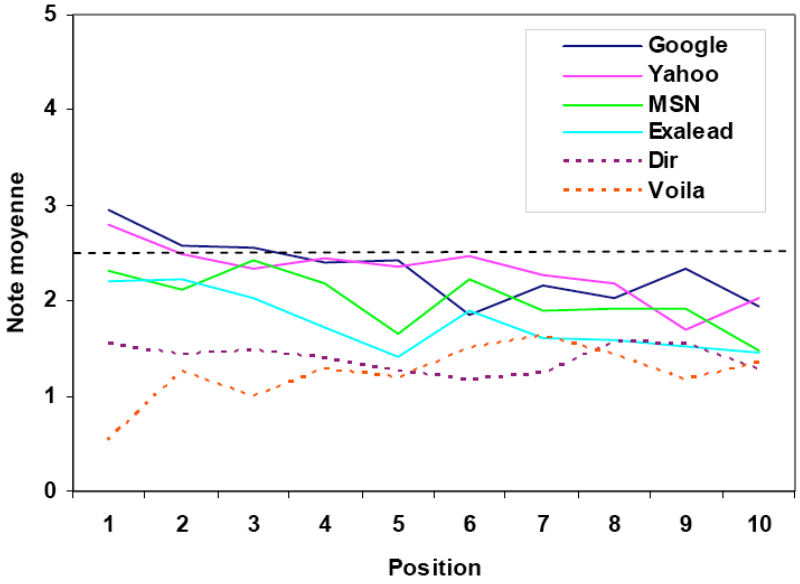


Fig. 2. Top10 recall-precision graph for all results (taken from Véronis, 2006)

Uniqueness of search results

Regarding the uniqueness of search engines, we have to distinguish between the uniqueness of the databases (defined by their overlap, see above) and the uniqueness of the search results (up to a certain cut-off rate). Two search engines based on the completely same index could deliver a completely different order of the results based on their ranking algorithms. This is an important point in Web-based research. The result sets tend to be overwhelmingly large, so that it is impossible for the user to look through all pages of the results list. Therefore, it could be useful to compare the top 10 or so results from different search engines to get different views on the same topic.

An important factor for the user is the uniqueness of the results of the different search engines (Spink, Jansen, Blakely, & Koshman, 2006; Véro-

nis, 2006). If switching the search engines brings different results, this is a good option if one does not find what was intended. In addition, the difference of the results is highly important for the discussion about the problems of a monopoly (or oligopoly) on the search engine market.

Studies discussing the overlap of search results from different engines were conducted to a large extent. We will not discuss in detail earlier studies (such as Chignell, Gwizdka, & Bodner, 1999; Gordon & Pathak, 1999; Nicholson, 2000; Schwartz, 1998). These all find little overlap between the search engines' results, but these findings are now of limited use because of the constantly changing search engine landscape.

A newer study focussing on the topic is the one by Spink et al. (2006). Search engines covered are Ask Jeeves, Google, Yahoo and MSN. For each search engine, the first 10 results are considered. The authors say that this limitation on the first page of results corresponds well to the user behaviour because users seldom go beyond the first page. The study also takes into account organic results and sponsored listings, but we will only report results for the organic listings.

The study is based on two sets of queries from April, 2005 (10,316 queries) and July, 2006 (12,570 queries). For every query, the top 10 results from each engine are downloaded. The comparison is done automatically using a direct comparison of the URLs. This approach is somehow problematic because of identical content under different URLs, where the search engines omit all but one URL for duplication (Bharat & Broder, 1998). This affects the results, and so we think that the actual overlap between search engines is higher than the numbers given in the results of studies just comparing URLs.

Spink et al. (2006) find that 84.9 percent of all hits are just listed by one search engine, while 11.4 percent by two, 2.6 percent by three and only 1.1 percent by all engines considered. The authors conclude that "using a single Web search engine only for a query means that a user misses exposure to a range of highly ranked Websites that are provided on the first page of results retrieved to any query" (p. 1,385). This may be true, but for a user not only the changing of the search engine, but also clicking the next button on the first results page to retrieve more results could be useful. Further research is needed that takes into account more results from each engine and applies a comparison between results that goes beyond the mere comparison of URLs.

In Véronis' study (Véronis, 2006; see above), the overlap of the top 10 results is also measured on the URL basis. He finds that the overlap between every two engines is very low, ranging from 2.9 percent to 25.1 per-

cent. Interestingly enough, the pair Google/Yahoo produces the highest degree of overlap.

Quality of the search features

This section discusses results from studies concerning the comparison of the power of the command languages and advanced search features, but also on the operational reliability of these.

There is no shortage of comparisons of search engine features and commands (e.g., Hock, 2004; Notess, 2006; Ojala, 2002). Early search engines such as AltaVista adapted their search functionality from classic online databases, which usually offer a wide range of operators and search functions. Later instalments are more oriented towards the average user who is not interested in advanced search. Nevertheless, search features and operators are necessary for conducting serious Web-based research. A discussion of search features that should be offered by search engines and the degree to which they are applied in the major search engines is given in Lewandowski (2004a). Unfortunately, the comparison of the search engines itself is hopelessly outdated. The reader here is referred to Notess' (2006) compilation in table form.

From a comparison of search engines and online databases, Othman & Halim (2004) can show how limited the search functionalities in search engines are in general. Even the functions regarded as common (i.e., five of the databases/search engines investigated offer this function) are only in part applied in the search engines.

A problem with search features that is often overlooked is their operational reliability. While there are functions clearly without any problematic potential (such as restriction to the top level domain), other functions that are relatively easy to apply do not work properly in some major search engines (e.g., Boolean OR in Google; see Notess, 2000). With trickier functions it is, to a large degree, unclear how well they work in different search engines. Such features are the language restriction, searching for related pages, content filters, and the date restriction.

This last feature is, to our knowledge, the only one of them systematically studied, as of yet. In a study testing the ability of search engines to determine the correct date of web documents, Lewandowski (2004c) finds that the major search engines all have problems with this task. He uses 50 randomly selected queries from the live ticker of the German search engine Fireball, which are sent to the major search engines Google, Yahoo and Teoma. These engines were selected because of their index sizes and their popularity at the time of the investigation. All searches were done twice:

once without any restrictions, and once with a date-restriction for the last six months. For each query, 20 results were examined for date information. The study reveals that about 30 to 33 percent of the pages have explicit update information in their content. This information was used to compare the non-restricted with the date-restricted queries.

The number of documents from the top 20 list that were updated within the last six months was counted and was defined as the up-to-dateness rate. The proportion of these documents, out of all the documents, was defined as the up-to-dateness rate. The corresponding sets of documents retrieved by the simple search, as well as by the date-restricted search, were calculated. The up-to-dateness rates for the simple search are 37 percent for Teoma, 49 percent for Google, and 41 percent for Yahoo. For the date-restricted search, the rates are 37 percent for Teoma (which means no improvement), 60 percent for Google, and 54 percent for Yahoo. Taking this into consideration, even Google, proved to be the best search engine, in this test fails in 40 percent of all documents. All in all, the study shows that the tested search engines have massive problems in determining the actual update of the documents found. But this data could be very useful for the indexing and even the ranking process (Lewandowski, 2006a).

The study recommends using information from several sources to identify the actual date of a document. The following factors should be combined: server date, date of the first time the document was indexed, metadata (if available), and update information provided in the content of the page.

Search engine usability

With respect to the users' searching behaviour, we use findings from our online survey conducted in 2003 (Schmidt-Maenz & Bomhardt, 2005), and other studies concerning search engine users. Additionally, we have observed the live tickers of three different search engines (Fireball (FB), Lycos (L), and Metaspinner (MS)), since Summer 2004 (Schmidt-Maenz & Koch, 2005, 2006). In our live tickers observed, the list could be updated automatically by refreshing those pages by use of a program. With that, we collected a nearly complete list of search queries performed on these engines during our observation period. Table 4 shows the most important results concerning interaction points between search engines and users.

We have analyzed these longitudinally and simultaneously collected observation data based on different search engines. As a consequence, we have a representative view of what searching persons do, since we have comparable data sets regarding observation length, time, and method. The

results of all three observed search engines are similar, for that reason it is assumed that these patterns will be the same for other engines, too.

The following results show how users interact with search engines regarding different parameters that reflect search engine usability.

Interface design

Interfaces of search engines have only one dimension, but there are different groups of search engine users which have different needs (Hotchkiss, Garrison, & Jensen, 2004). Most searching persons only evaluate the result listings very quickly before clicking on one or two recommended web pages (Hotchkiss et al., 2004; Spink & Jansen, 2004). Google has a very clear input window, while Yahoo! is overloaded by adverts and news (Geoghegan, 2004). Paid placements are often not clearly separated from the organic lists. They highlight those links with very light backround colors (e.g., Google) or give only hints written in very small and slightly coloured letters (e.g., Altavista). That's why users often cannot differentiate between those two or have the feeling that the link they clicked on could be a paid listed link. Additionally, it is important to present only a few results (10 to 15) since search engine users are not willingly to scroll (Hotchkiss et al., 2004).

Additionally, search engines have to provide features to help users to specialize their search queries. Especially advanced users apply operators and features. Every major search engine provides advanced search features except Excite (Fauldrath & Kunisch, 2005).

Acceptance of search features and operators

Search queries are very short and do not show any variations over a longitudinal period. German search queries are, on average, a little bit shorter than English queries since in German word compositions are used instead of words stringed together. Nearly half of the search queries contain only one term. Regarding search terms which occur nearly every day (Schmidt-Maenz & Koch, 2006) one finds many operators used inappropriately and fillers such as "in" or "for". This shows how intuitively online searching persons formulate their queries.

The results from studies mentioned above could not be confirmed, here, since only operators presented at the beginning such as ' + ', ' - ' or the phrase search were used relatively frequently. But altogether, the usage of operators accounted for less than 3 percent of all search queries observed. The phrase search was the most frequent form to arrange search queries in a complex way. Here, search queries with phrases were 2.1 percent for

Fireball, 2.4 percent for Lycos, and 2.5 percent for Metaspinner (Schmidt-Maenz & Koch, 2005).

In the Live Ticker, the German search engines Fireball and Metaspinner also show the selected search area in addition to the current search queries. The search for German pages, only, is selected most frequently. This results from the fact that this area is a pre-adjusted standard in both search engines. In more than two thirds of all search queries, users do not personalize their search by using such features. People in the context of our Internet survey also told that they do not personalize their favourite search engines according to their needs. That means that, all in all, search features such as operators are not accepted. To tell the truth, John Q. Public does not even know how to use operators or what to do with search features.

Performance of search engines

The most disliked factor in search engine result lists are web pages that are optimized to high rankings in result lists, only, and are therefore of little value to the user, and other pages that do not fit the search queries performed (24.4 percent of 2014) and advertisement pages (21.4 percent of 2014). We think that Internet users often do not know whether they click on paid or organic results. In Machill et al. (2003) respondents said that they are unsatisfied with results of which nobody knows whether they are paid. A high percentage of respondents (76.6 percent of 6133) do not think about personalization possibilities of their preferred search engines. These results show that it is possible to evaluate the search engine usability by user surveys. Responders also find what they were supposed to. But the quality of results found is unclear. 70.8 percent of 6722 responders very often return to the search engine, instantly, when they do not find what they want on a recommended web page.

User guidance

Internet users commonly do not know how search engines work. We asked five general questions about search engines, such as “Is the following statement correct? Meta search engines have their own index”. But only 44.2 percent of 5944 interviewees were able to answer four or five of these questions correctly. We also find that users with more correct answers use significantly more operators (Schmidt-Maenz & Bomhardt, 2005). We show by our results that users, generally, don’t understand search engines. Considering this, it is important to have a clear and simple search engine interface to improve the usability of search engines.

Help functions are provided by most search engines, but it is always a very small button (Google, Yahoo!). Fauldrath & Kunisch (2005) stated that only 57 percent of examined search engine have a help page, which is easy to find. In most cases this is titled with "all about…" instead of a precise anchor text, such as "help". It is also hard for beginners to know what they are looking for. A general description of what search engines definitely do is missing. Only 71 percent of major search engines give some help on how to process a search session.

Another point to improve user guidance is to give additional information to the ranked pages. Here the title of the documents, a short description, and the URL are helpful. Every search engine provides this information. But it is also interesting for users to see when last changes were made on the recommended page, or similar search terms are given. 71 percent of major search engines provide temporal information and only 29 percent suggest similar terms (Fauldrath & Kunisch, 2005).

Table 5. Empirical results of the observation of three different search tickers (Schmidt-Mänz; 2007)

ID	Year	Days	# Search Queries	Avg. Length	1-Term Queries	Complex Queries	Phrase Search	Search Feature
FB	2004	399	132,833,007	1.8	50.1%	<3.0%	2.1%	65.8%
L	2004	403	189,930,859	1.7	51.9%	<3.0%	2.4%	-
MS	2004	314	4,089,731	1.8	48.4%	<3.0%	2.5%	87.9%

Conclusions

Today, nobody knows the real performance or accuracy of search engines. There are several studies dealing with a single aspect of quality measurement, but none that tries to evaluate search engine quality as a whole. There was a lack of an overview of empirical results and of quality measures to be used. Our measurement perspectives initiate the discussion about the important matter of search engine quality. With this, it is possible to enhance transparency and diversity on the search engine market.

We showed that there definitely is a gap between the performance of search engines and user needs, respectively capabilities. Regarding user searching behaviour, there are several possibilities, which could be improved. Our assumption is that users do not know how to best interact with search engines. For that reason help functions have to be offered so that more intuitive users also can learn to handle Internet search engines. The next point is the presentation of search results. Search engine should

clearly separate paid listings from organic results. User should also get the possibility to learn about the functionality of search engines. Users search often in an intuitive way, for that reason search engines should give accurate results based on very short or very specialized search queries.

Some questions are still open. What does the European country bias of search Engines look like? How large is the intersection of search engines regarding more than the first results page? Which design of search engine user interfaces will be best suitable for the users' needs? Our next steps will be to give answers to these questions.

Our search engine quality parameters will help to conduct quality studies to compare different search engines with the same measures. This will again help users to decide which search engines they will prefer to use.

Another important point in the future will be to enlighten users about how search engines work, what they really do and how to use them.

Most research deals with very special parameters to measure search engine quality and the user behaviour is often completely omitted. In this chapter, we introduced a comprehensive approach to measure both, search engine quality with all technical aspects and with aspects from the users' perspective.

Our further research will be to conduct such a comprehensive study by comparing search engine quality of the major search engines. Here, we will include user surveys and laboratory studies.

References

Acharya, A., Cutts, M., Dean, J., Haahr, P., Henzinger, M., Hoelzle, U., et al. (2005). Information Retrieval Based on Historical Data, USA.

Beitzel, S., Jensen, C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). In Hourly Analysis of a Very Large Topically Categorized Web Query Log (pp. 321-328). Paper presented at the ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK. ACM Press.

Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing, 7(1).

Bharat, K., & Broder, A. (1998). A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. Computer Networks and ISDN Systems, 30(1-7), 379-388.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7), 107-117.

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

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph Structure in the Web. Retrieved 15.4.2006, from http://www.almaden.ibm.com/webfountain/resources/GraphStructureintheWeb.pdf

Cacheda, F., & Viña, Á. (2001). Understanding How People Use Search Engines: A Statistical Analysis for E-Business (Vol. 1, pp. 319-325). Paper presented at the e-2001 E-Business and E-Work Conference and Exhibition.

Ding, W., & Marchionini, G. (1996). A Comparative Study of Web Search Service Performance, Proceedings of the 59th American Society for Information Science Annual Meeting (pp. 136-142): Learned Information.

Ford, N., Miller, D., & Moss, N. (2002). Web Search Strategies and Retrieval Effectiveness: an Empirical Study. Journal of Documentation, 58(1), 30-48.

Geoghegan, T. (2004). Search Wars - Which is Best?, from news.bbc.co.uk/2/hi/uk_news/magazine/4003193.stm

Greisdorf, H., & Spink, A. (2001). Median Measure: An Approach to IR Systems Evaluation. Information Processing & Management, 37(6), 843-857.

Griesbaum, J. (2004). Evaluation of three German Search Engines: Altavista.de, Google.de and Lycos.de. Information Research, 9(4).

Griesbaum, J., Rittberger, M., & Bekavac, B. (2002). In: R. Hammwöhner, C. Wolff & C. Womser-Hacker (Eds.), Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de (pp. 201-223). Paper presented at the Information und Mobilität. Optimierung und Vermeidung von Mobilität durch Information. 8. Internationales Symposium für Informationswissenschaft. UVK.

Gulli, A., & Signorini, A. (2005). The Indexable Web is More Than 11.5 billion Pages (pp. 902-903). Paper presented at the Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Chiba, Japan.

Hoelscher, C., & Strube, G. (2000). Web Search Behavior of Internet Experts and Newbies (pp. 337-346). Paper presented at the 9th International World Wide Web Conference.

Ingwersen, P., & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer.

Jansen, B. (2000). An Investigation Into the Use of Simple Queries on Web IR Systems. Information Research, An Electronic Journal, 6(1).

Jansen, B., & Spink, A. (2003). An Analysis of Web Documents Retrieved and Viewed (pp. 64-69). Paper presented at the 4th International Conference on Internet Computing.

Jansen, B., & Spink, A. (2006). How we are Searching the World Wide Web? A Comparision of Nine Searech Engine Transaction Logs. Information Processing and Management, 42(1), 248-263.

Ke, Y., Deng, L., Ng, W., & Lee, D.L. (2006). Web Dynamics and their Ramifications for the Development of Web Search Engines. Computer Networks, 50(10), 1430-1447.

Kleinberg, J.M. (1999). Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46, 604-632.

Korfhage, R.R. (1997). Information Storage and Retrieval. New York: Wiley.
Lawrence, S., & Giles, C.L. (1998). Searching the World Wide Web. Science, 280, 98-100.
Lawrence, S., & Giles, C.L. (1999). Accessibility of Information on the web. Nature, 400(8), 107-109.
Leighton, H.V., & Srivastava, J. (1999). First 20 Precision among World Wide Web Search Services (Search Engines). Journal of the American Society for Information Science, 50(10), 870-881.
Lewandowski, D. (2004a). Abfragesprachen und erweiterte Suchfunktionen von WWW-Suchmaschinen. Information Wissenschaft und Praxis, 55(2), 97-102.
Lewandowski, D. (2004b). Bewertung von linktopologischen Verfahren als bestimmender Ranking-Faktor bei WWW-Suchmaschinen, Wissensorganisation und gesellschaftliche Verantwortung. 9. Tagung der Deutschen ISKO (Wissensorganisation'2004). Duisburg, Germany.
Lewandowski, D. (2004c). Date-restricted Queries in Web Search Engines. Online Information Review, 28(6), 420-427.
Lewandowski, D. (2005a). Web Searching, Search Engines and Information Retrieval. Information Services and Use, 18(3), 137-147.
Lewandowski, D. (2005b). Yahoo - Zweifel an den Angaben zur Indexgröße, Suche in mehreren Sprachen. Password, 20(9), 21-22.
Lewandowski, D. (2006a). Aktualität als erfolgskritischer Faktor bei Suchmaschinen. Information Wissenschaft und Praxis, 57(3), 141-148.
Lewandowski, D. (2006b). Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können. Zeitschrift für Bibliothekswesen und Bibliographie, 53(2), 71-78.
Lewandowski, D. (2006c). Zur Bewertung der Qualität von Suchmaschinen. In: J. Eberspächer & S. Holtel (Eds.), Suchen und Finden im Internet (pp. 195-199). Heidelberg: Springer.
Lewandowski, D., & Mayr, P. (2006). Exploring the Academic Invisible Web. Library Hi Tech, 24(4), 529-539.
Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2006). The Freshness of Web search engine databases. Journal of Information Science, 32(2), 133-150.
MacCall, S.L., & Cleveland, A.D. (1999). A Relevance-based Quantitative Measure for Internet Information Retrieval Evaluation (pp. 763-768). Paper presented at the Proceedings of the American Society for Information Science Annual Meeting.
Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2003). Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In M. Machill & C. Welp (Eds.), Wegweiser im Netz. Gütersloh: Bertelsmann Stiftung.
Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2004). Navigating the Internet: A Study of German-Language Search Engines. European Journal of Communication, 19(3), 321-347.

Notess, G.R. (2003). Search Engine Statistics: Freshness Showdown. Retrieved 4.1.2005, from http://www.searchengineshowdown.com/stats/freshness.shtml

Ntoulas, A., Cho, J., & Olston, C. (2004). What's New on the Web? The Evolution of the Web from a Search Engine Perspective. Paper presented at the Thirteenth WWW Conference, New York, USA.

Ozmutlu, H., Spink, A., & Ozmutlu, S. (2003). A Study of Multitasking Web Search (pp. 145-148). Paper presented at the International Conference on Information Technology: Computers and Communications.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Retrieved 24.7.2006, from http://dbpubs.stanford.edu:8090/pub/1999-66

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1988). SERVQUAL: A Multiple-item Scale for Measuring Consumer Perceptions of Service Quality. Journal of Retailing, 64(1), 12-40.

Risvik, K.M., & Michelsen, R. (2002). Search engines and Web dynamics. Computer Networks, 39(3), 289-302.

Saracevic, T. (1995). In Evaluation of Evaluation in Information Retrieval (pp. 138-146). Paper presented at the SIGIR'95, Seattle, CA. ACM Press.

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

Schmidt-Maenz, N., & Bomhardt, C. (2005). Wie Suchen Onliner im Internet? Science Factory/Absatzwirtschaft(2), 5-8.

Schmidt-Maenz, N., & Gaul, W. (2005). Web Mining and Online Visibility. In C. Weihs & W. Gaul (Eds.), Classification - the Ubiquitous Challenge (pp. 418-425): Springer.

Schmidt-Maenz, N., & Koch, M. (2006). A General Classification of (Search) Queries and Terms (pp. 375-381). Paper presented at the 3rd International Conference on Information Technologies: Next Generations, Las Vegas, Nevada, USA.

Sherman, C., & Price, G. (2001). The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today.

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

Singhal, A., & Kaszkiel, M. (2001). A case study in web search using TREC algorithms (pp. 708-716). Paper presented at the 10th international conference on World Wide Web, Hong Kong.

Spink, A., & Jansen, B. (2004). Web Search: Public Searching of the Web (Vol. 6). Dordrecht, Boston, London: Kluwer Academic Publishers.

Spink, A., Jansen, B., & Ozmutlu, H. (2000). Use of Query Reformulation and Relevance Feedback by Excite Users. Internet Research: Electronic Networking Applications and Policy, 19(4), 317-328.

Spink, A., Ozmutlu, S., Ozmutlu, H., & Jansen, B. (2002). U.S. Versus European Web Searching Processes. Journal of the American Society for Information Science and Technology, 53(8), 639-652.

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

Su, L.T. (1998). Value of Search Results as a Whole as the Best Single Measure of Information Retrieval Performance. Information Processing & Management, 34(5), 557-579.

Sullivan, D. (2005). Search Engine Sizes. Retrieved 24.7.2006, from http://searchenginewatch.com/showPage.html?page=2156481

Vaughan, L. (2004). New Measurements for Search Engine Evaluation Proposed and Tested. Information Processing & Management, 40(4), 677-691.

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

Véronis, J. (2006). A Comparative Study of six Search Engines. Retrieved 15.3.2006, from http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf

Wang, H., Xie, M., & Goh, T.N. (1999). Service Quality of Internet Search Engines. Journal of Information Science, 25(6), 499-507.

Williams, M.E. (2005). The State of Databases Today: 2005. In Gale Directory of Databases (Vol. 2, pp. XV-XXV). Detroit, Mich.: Gale Group.

Wolff, C. (2000). Vergleichende Evaluierung von Such- und Metasuchmaschinen, 7. Internationales Symposium für Informationswissenschaft (ISI 2000) (pp. 31-38). Darmstadt, Germany: Universitätsverlag Konstanz.

Xie, M., Wang, H., & Goh, T.N. (1998). Quality Dimensions of Internet Search Engines. Journal of Information Science, 24(5), 365-372.

Zien, J., Meyer, J., Tomlin, J., & Liu, J. (2000). Web Query Characteristics and their Implications on Search Engines: Almaden Research Center.