Vol. 9 No. 4, July 2004

Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de

Joachim Griesbaum
University of Konstanz
Information Science
Fach D 87
D-78457 Konstanz
Germany

change font

Abstract

The goal of this study was to investigate the retrieval effectiveness of three popular German Web search services. For this purpose the engines Altavista.de, Google.de and Lycos.de were compared with each other in terms of the precision of their top twenty results. The test panelists were based on a collection of fifty randomly selected queries, and relevance assessments were made by independent jurors. Relevance assessments were acquired separately a) for the search results themselves and b) for the result descriptions on the search engine results pages. The basic findings were: 1.) Google reached the best result values. Statistical validation showed that Google performed significantly better than Altavista, but there was no significant difference between Google and Lycos. Lycos also attained better values than Altavista, but again the differences reached no significant value. In terms of top twenty precision, the experiment showed similar outcomes to the preceding retrieval test in 2002. Google, followed by Lycos and then Altavista, still performs best, but the gaps between the engines are closer now. 2.) There are big deviations between the relevance assignments based on the judgement of the results themselves and those based on the judgements of the result descriptions on the search engine results pages.

Introduction

On searching the Web, people usually rely on only a few search services such as Google, Yahoo, MSN and Lycos. At present, Google is the dominant service among them. For information retrieval purposes most people use either Google directly or one of the Google-powered portals such as AOL. Google's success is, for the most part, traced back to the assumed high quality of its search results (Sullivan, 2004). Though Google is classified as being superior to its competitors, the question is if and how far this is really true. If the quality of the search results is the most decisive factor for users when choosing a search engine, then there is an urgent need to find ways to measure it, at least approximately, in an objective way. As Sullivan recently pointed out:

> Why is getting a relevancy figure so important to consumers? First and foremost, it would bring a greater awareness that there are real choices in search.... If it turns out that relevancy testing finds that Google and its competitors are all in roughly the same degree of relevancy, then users might then be willing to experiment more with others. (Sullivan, 2002, December 5)

The aim of this investigation was to compare the retrieval effectiveness of Google and some of its prominent competitors. The intension was on the one hand, to measure the quality of the search results which give users hands-on hints as to which of the engines is best to choose when in need of information and, on the other hand to contribute to the theoretical question of how to measure the quality of Web retrieval systems.

Related Work

A growing body of inquiries is concerned with the usage and quality of search engines, e.g., (Bar-Ilan, 2002; Gordon and Pathak, 1999; Leighton and Srivastava, 1999; Hawking *et al.*, 2001; Ford *et al.*, 2001; Dennis *et al.*, 2002; Spink, 2002; Eguchi *et al.*, 2002; Gurrin and Smeaton, 2003; Mandl, 2003).

Evaluation can be seen as an important as well as difficult part of information retrieval. It is a complex topic lacking a safeguarded theoretical framework. Therefore, there is a wide range of different approaches, all facing the problem of developing or using approximate test designs (Griesbaum, 2000). The TREC3conferences offer a continuous, institutionalised infrastructure for large-scale evaluations that allows comparative studies with standard guidelines in controlled test environments. TREC Web Trac retrieval tasks are based on test collections containing documents, topics and relevance judgements. Eguchi *et al.*, (2002) give an overview of the Web Retrieval Task at the Third NTCIR Workshop. Its evaluation measures primarily rely on recall and precision. Although TREC applies as a standard test-environment, the methodologies it uses and the achieved results could be scrutinized. Gurrin and Smeaton, (2003), for example, pointed out that Web Trac collections heavily underestimate the off-site link density of the real Web. Furthermore, Web Trac collections are static, not reflecting the high volatility of Web pages. This means if the goal of an evaluation is to compare the real effectiveness of search engines, Web Trac methodologies could serve as an important model. But a one to one adaption of the applied test designs seems to be insufficient and might be misleading.

Spink, (2002) goes even further and criticises the limitations of TREC-like evaluation approaches as basically ignoring the important point of interaction in the retrieval process. The results of her user-centered approach showed no correlation between user-based evaluation measures such as change in information seeking stage and precision. Shang and LongZhuang, (2002), on the other hand used a largely automatic test design with thousands of queries and computed relevance scores for a statistical comparison of the precision of fifteen Web search engines. In 2003 Inktomi commissioned Veritest to perform a relevance test for Google, Wisenut, Fast, Teoma and Altavista. The aim was to compare the relative quality of the top ten results. The test collection was built of 100 randomly selected queries from the Inktomi engine logfiles. The URL of the top ten hits were judged for relevancy by three independent jurors. (Inktomi..., 2003)

These example investigations show there is wide diversity of evaluation approaches. The answer to the question as to which evaluation approach is appropriate and should be used is: it depends on the purpose of the inquiry. For example, if the goal is a comparison of different engines, i.e., concerning the number of relevant items retrieved within a certain range, then the neutrality of the test design is of capital importance. In such a case it could be justified or even necessary to abstract test parameters from real circumstances to avoid biases in favour of one or some of the compared engines. A test design with the aim to evaluate the effects of machine-user interaction regarding individual engines should be based on actual circumstances such as real users in interactive search sessions.

Methodology

This inquiry follows the advice given by Tague-Sutcliffe, (1992) for developing a retrieval test. She proposed constructing a retrieval test design with ten steps to guide the investigators in building a valid and reliable evaluation.[1]

1. Need for testing – motivation of the inquiry
2. Type of test – determination of the test procedure
3. Definition of variables
4. Database development – search engine selection
5. Finding queries
6. Processing queries
7. Experimental design

8. Data collection
9. Data analysis
10. Presenting results

Afterwards, the designed test setting was reviewed with a pre-test. Depending on its results, the experimental design would have been modified. Then the tests were carried out, data was collected and analyzed. Finally, a critical reflection of both the results and the test design itself showed on the one hand, how expressive the results really are and, on the other hand, illuminated the advantages and shortcomings of the employed test design.

Developing the test design

Need for testing

The aim of this investigation was to compare the retrieval effectiveness of Google and some of its competitors, wherein the quality of the search results serves as the benchmark for the performance appraisal. What is meant by the quality of search results? Is it the quality of the results themselves or the assumed quality of the result descriptions on the search engines' results pages? What is the difference? If there is a difference, is it relevant? Some evaluations do not differ between both *kinds* of results. Results and presentation of results are treated the same way, i.e., (Spink, 2002). Most evaluations either use the results themselves or the search engine result pages as the basis for the performance assessment (Griesbaum *et al.*, 2002). When using search engines the descriptions of results are the first things searchers see, and the results are what they get in the end. Figure 1 shows the search engines' result pages for the query *travel information st. lucia* of three evaluated search engines. Figure 2 gives a detailed picture of the result presentation of the first result, which is the same Web page on all engines. Finally Figure 3 shows the result itself that is described by the engines in Figure 2.

Why is it important to differ between the results themselves and the presentation of results on search engine result pages? Generally, in a first step, users decide with the help of the result presentations which hits deserve further investigation. Only those that seem to be relevant have a chance of being selected. Hence the presentations of the results predetermine the user choices of assumed relevant items, regardless of the real quality of the Web pages. If the quality estimations given by the result lists correspond with the true quality of the results, there is no problem; on the contrary, the overview enables users to differentiate rapidly between good and bad results. But what if relevance assessments of the result representations and the results differ in a perceivable way?

Users could waste time trying to detect whether would-be relevant pages serve their information need. In a worst case scenario users would miss relevant hits, if the result presentation is inadequate. But is this potential problem realistic or noticeable? One should expect that there are only minor differences between the promised and real page quality. But is that true? Differences may occur due to removed or changed Web pages by insufficient presentation algorithms or by manipulation techniques like cloaking.[2] In this evaluation the goal is to estimate both. The quality of the results themselves shows the real quality of the search engine output. The descriptions on the result pages are a strong indicator as to which of the relevant hits users would be most likely to pick. If the assumed quality of the result descriptions differs in a significant manner, new questions arise such as: are the engines spammed, or are the indexing frequencies too slow, and so on. In this inquiry the goal is only to get a rough picture: to determine if further examinations are actually worthwhile or even needed at all.

Type of test

This evaluation is based on a test collection. Information needs, queries, performance measures and relevance criteria were predetermined for users who served as independent jurors to give relevance assessments of the search results.

Definition of variables

The independent variables are evaluation criteria, retrieval performance measurement, queries, information needs and participants. The dependent variables are the relevance judgements which serve as indicators of the retrieval performance.

Evaluation criteria

In this inquiry the relevance of the search engine results serve as basic evaluation criteria. Although the word relevance has been used in different ways, it broadly corresponds to how well a document satisfies a user's information need (Robertson, 1981: 14). The use of relevance as an evaluation measure is a highly problematic decision. Relevance assessments are bound to the subjective views, circumstances, knowledge, etc., of the people who serve as jurors and can not be easily generalized. Moreover relevance judgements are not stable (Harter, 1996: 38). Eighty factors have been suggested as affecting relevance assessments (Schamber, 1994: 11). Furthermore, relevance judgements are document-centric and do not reflect the relevance of the search result set as a whole. In particular learning effects on the side of the searcher are widely ignored. These problems illustrate clearly the antagonism between the function of relevancy as an independent, and therefore objective, effectiveness measure on the one hand and the subjective expressiveness of real judgements on the other hand.

The contradiction cited above can be dealt with relevance assessments made by independent jurors. This ensured that the results were not biased by preferences or indispositions of the researchers. To minimize the influences of biases concerning the relevance assessments made by the jurors, the origin of the results had to be disguised. This is not possible for the presentation of results on the search engine result lists. To avoid learning effects the results and the presentation of results had to be judged by different people for the same query. Within these restrictions all hits per query were judged by the same person. This helped to secure the uniformity of the assessments.

The World Wide Web is based on a hypertext structure. Therefore, it is possible that documents that are not relevant by themselves allow access to relevant pages by hyperlinks. The dichotomy of *relevant* and *not relevant* was ignored at first and a third relevance judgement possibility *links to relevant page(s)* was supplemented. Assessed *links to relevant page(s)* were finally added to the *relevant* pages, for the simple reason that they are also helpful in satisfying information needs. The results representations on the search engines result lists are treated in a different way because they generally do not satisfy the users' information need by themselves but predetermine which hits have a chance at being selected. Therefore, they are either judged as *seems to be relevant, I would click this link* or '*seems not to be relevant, I would not click this link*.

Retrieval performance measurement

Recall and Precision are standard retrieval performance measures (Cleverdon *et al.*, 1966). Recall can hardly be measured in the Web. Furthermore, users usually view only the first two result pages of the engines' output (Jansen *et al.*, 2000). For this reason this evaluation restricted the retrieval performance measurement to the number of relevant hits retrieved within the first twenty results, the so called top-twenty precision. Although recall was principally disregarded, the *number of retrieved items* and *number of answered questions* were counted. *Number of retrieved items* shows the Web coverage deviations among the different engines. *Number of answered questions* points out in how many cases the engines deliver at least one relevant hit within the top-twenty results. This shows how often the engines are able to be at least marginally useful in that they deliver something relevant. The main results of this inquiry are based on the analysis of top-twenty micro- and macro-precision. Micro-precision relies on the documents as base units, it is calculated as the ratio of relevant documents to all documents for all queries. This means all documents are of equal importance. Macro-precision focuses on the precision of the single queries. This means all queries are of equal importance.

Queries and information needs

The information needs and queries are the core of each retrieval test. Subject area, complexity, specificity of the information needs and the method of query formulation specify the details of the test and determine quantity and quality of the retrieval output of the search engines. That means the kind of queries themselves, i.e., specificity, predetermines precision and recall values of the results. Hence the test results can only be seen as *objective* general retrieval performance values to a smaller extent. They rather show effectiveness differences between the engines for certain kinds of queries. The goal of this inquiry was to reflect real Web search contexts. For this reason, typical Web search queries were used. That is, a random sample set of real queries was selected. There was one exception: queries explicitly aimed at pornography or violence, for example *tierquälerei* (cruelty to animals) or *porno* (porn), were excluded.

This is a normative decision. Typical Web queries are short—in most cases one, two or three terms—and therefore very often rather unspecific. After the queries were selected the information needs were reconstructed. This was a very difficult task because the less specified the query is, the more diverse the underlying information needs can be. For example, the query 'travel' can mean that someone is searching for a) *general information about travel* b) *travel agencies* c) *commercial vacation offers* d) *online booking possibilities* or e) *meta-information services about travel* or all of them. Although there are different probabilities for the different information needs it is dangerous to restrict the relevance judgements of the jurors to a *most cases right* subset. Therefore, all of the reconstructed information had to be considered by the jurors when judging the relevance of the search results. That is, the less specific the query and the more diverse the underlying information need, the higher the risk that the relevance judgements of the jurors did not reflect the real context of the search which was initially done by a specific user. In such cases there is a tendency that the relevance judgements rather reflect semantic similarity between queries and results and not the intentions of real searchers. Even this is not the case for polysemantic terms like *jaguar*, which could mean that people want to find information a) of *the animal* or b) of *the car manufacturer*, such queries were avoided too. Queries were employed without operators, since most users do not use them (Jansen *et al.*, 2000). Fifty queries were applied, this number corresponds to the TREC standard values, and seems high enough to get reliable results that can be generalized within the given qualitative context in respect to topicality, specificity, etc., of the chosen queries (Buckley and Voorhees, 2000).

Participants

To ensure the reliability of the relevance judgements the number of independent jurors should be representative for typical Web users. This is a very difficult task. Given the constraint resources of this investigation a pragmatic approach was chosen. The aim was not to rebuild typical Web users as jurors but to secure that the relevance judgements can be seen as representative for at least a certain, homogenous group of competent users. It was decided that jurors should consist of people well experienced with the Web and search engines. At TREC conferences a relevance assessment consistency of 70%-80% between different jurors was observed (Kowalski, 1997: 225). To balance these effects of relevance judgement deviations between different jurors, a number as high as possible, ideally fifty jurors, one for each query should be chosen.

Students selected were participants of the retrieval lecture in Konstanz. The members of the staff of information science and the employees of QualiGO are well experienced with information retrieval, in particular Web searching and search engines, through their daily work. Therefore, it was established that all jurors were familiar with terms and concepts such as relevance, recall and precision.

Twenty-nine jurors could be recruited, which was not enough to assign a different juror to each query. Similiar to the preceding evaluations (Griesbaum *et al.*, 2002: 211; Griesbaum, 2000: 70) it was decided that each juror had to deal with two tests. It was assumed that twenty-five jurors is a high enough number to compensate for biases implicit in the relevance assessment consistency of 70%-80% (Kowalski, 1997: 225). Finally, twenty-seven jurors were employed because two tests had to be repeated due to technical problems with the display of four results, so that two persons from the remaining four persons had to be used.

There were seventeen male jurors and ten female jurors; eighteen were students of the university, five were members of academic staff, three were employees of QualiGO and one was a former employee of QualiGO.

The profile self-assessment of the jurors concerning the judgement of the fifty queries is shown in the following tables.

Age	0-20	21-30	31-40	41-50	51-60
Self assessments on the fifty queries	0	43	3	2	2
Number of jurors		23	2	1	1

Table 1: Age of jurors.

Computer and software knowhow	Beginner	Advanced	Expert

Self assessments on the fifty queries	1	20	29
Number of jurors	1	11	15

Table 2: Computer and software know how.

Web usage	Seldom / never	Several times per week	Daily
Self assessments on the fifty queries	0	9	41
Number of jurors	0	5	21

Table 3: Web usage.

Search engine usage	Seldom / never	Several times per week	Daily
Self assessments on the fifty queries	0	11	39
Number of jurors	0	6	21

Table 4: Search engine usage.

These tables show that the jurors fulfilled the qualitive criteria. Their self-assessment confirmed their Web and search engine experience. Hence they can be rated as experts.

Database development—search engine selection

Three engines were compared with each other. The base requirement selection criterion was that the engines were also tested in a preceding test in 2002 (Griesbaum *et al.*, 2002). Furthermore, selection criteria were audience reach, as '...the percentage of home Web surfers estimated to have visited each site during the month' (Sullivan, 2002, April 29) and international name recognition. Taking this into consideration, Altavista.de, Google.de and Lycos.de were selected.

Finding queries

The queries should represent real searches in form and content as closely as possible. To avoid a topical or other kind of narrowing, a two-way approach was chosen.

The first half of the queries, twenty-five for the actual test and two for the pre-test, were composed of a randomly selected subset of the most frequently asked queries (Top 500) that are recorded in the logfiles by the pay-per-click engine QualiGO. QualiGO logfiles contain the queries of a large number of its partner search sites like Bluewin.ch, Blitzsuche.de, Sharelook.de, etc. The idea is that this logfile, which is composed of queries that were employed by many different engines, is more representative to the typical Web users' queries than logfiles which contain records from only one engine.

Table 5 shows the selected queries from QualiGO Logfiles.

Selected queries from QualiGO	English translation
flirt	flirt
wetter	weather
gartenarbeit	gardening

vergnuegungsparks	amusement parks
auktionen	auctions
heilbaeder	therapeutic baths
private hochschulen	private universities
deutsche bahn	german railways
reisen	travel
babykleidung	baby clothes
pollenallergie	pollen allergy
altersteilzeit	partial retirement
bewerbungsschreiben	job application
sonja kraus	sonja kraus
parfum sonderangebote schnaeppchen	perfume specials bargain
online weiterbildung	online further education
schuldentilgung	depth redemption
motorradbekleidung	motorcycle clothes
return to castle wolfenstein komplettloesung	return to castle wolfenstein walkthrough
harry potter und der stein der weisen	harry potter and the sourcerer's stone
online privathaftpflichtversicherung	online private liability insurance
die fabelhafte welt der amelie	the wonderful world of amelie
sms kostenlos ohne anmeldung	free sms without registration
window color malvorlagen	window color templates
weihnachtslieder download	christmas songs download

Table 5: 25 randomly selected queries from a subset of the most frequently asked queries (Top 500) that are recorded in the logfiles by the pay-per-click engine QualiGO.

The challenge lies in rebuilding the underlying information needs and relevance criteria. In this study they were reconstructed in different steps. First, the researchers wrote down in one sentence what they thought was the aim of the search. Then they formulated in three or four sentences the relevance criteria that specify which kind of pages could or could not be relevant. Once this was done, three assistants and former members of the working group Information Science at Konstanz subsequently checked and modified these drafts. In a final discussion the final versions were formulated.

Table 6 shows an example for the query *travel*.

Query	Information need	Relevance criteria
travel	User seeks information about travel in the Web, he wants to research travel agencies, vacation offers, online booking possibilities or information about that topic, eventually he wants to book a trip.	Relevant pages contain information about travel or contain travel offers also relevant are travel agencies or travel (meta) information services or online booking possibilities.

Table 6: Example for the reconstruction of information needs and relevance criteria for the query "travel".

To avoid the danger of equating representativity of the queries with their occurrence frequency solely in QualiGO, the second half of the queries was selected with the help of a second service. Askjeeves.com *Top Searches* service was selected because queries on this search service often deliver a set of natural language phrases, which are very helpful in determining the possible underlying information needs.[footnote3] A random sample of queries from the *Top Searches* was chosen. This sample was used in searches on Askjeeves.com and the given set of natural languages phrases under the heading *Did you ask for* on the corresponding result page was used as the base of the information need. The reconstruction of the information needs was much easier this way. After that, the queries were translated into German if necessary. Finally the queries and relevance criteria were formulated and checked the same way as the queries taken from the QualiGO logfiles.

Table 7 shows the selected queries from from Askjeeves *Top Searches* service.

Selected queries from Askjeeves "Top Searches" service	English translation
reiseinformation st. lucia	travel information st. lucia
elvis presley bilder	elvis presley pictures
kinofilm volcano video dvd	movie volcano video dvd
scheidung informationen	divorce information
robert swindells	robert swindells
fehlgeburt	misscarriage
hotmail	hotmail
christina aguilera songtexte	Christina aguliera song lyrics
harley davidson bekleidung	harley davidson clothing
brad pitt videos	prad pitt videos
yahoo	yahoo
offspring musikvideos	offspring music videos
luftverschmutzung	air pollution
amnesty international	amnesty international
easy jet	easy jet
planet merkur	planet of mercury
charles dickens biographie	charles dickens biography
disneyworld tickets	disneyworld tickets
diät	diet
geschichte griechenland	history of greece
kostenlos schnittmuster	free sewing patterns
stammbaumforschung	genealogy
toyota	toyota
clipart	clip art

vulkane japan	volcano japan

Table 7: 25 randomly selected "Top Searches" queries on Askjeeves.com.

The whole query set containing information needs and relevance criteria can be found in the Appendix.

To get a picture of the characteristics of the query corpus some attributes were classified. Queries were categorized as rather specific if the information need was relatively clear. An example is the query *reiseinformation st. lucia (travel information st. lucia),* the underlying information need being that the user is seeking travel information on the island of St. Lucia for vacation planning.

Queries were categorized as rather unspecific if the aim of the searcher was not clearly laid out. An example is the query *reisen (travel)* which could mean the user is searching either information about travel, or travel opportunities, if he even wants to book a trip online.

Grading of topical specificity		Number of query terms			
Rather specific	Rather unspecific	1	2	3	4 or higher
18	32	21	16	8	5

Table 8: Characteristics of the queries.

The attributes of the queries of this evaluation can be summarized as follows. There were no topical constraints, but the restriction that queries explicitly aimed at pornography or violence were excluded. The topical specificity was predominantly low because two thirds of the queries are categorized as rather unspecific. That means there was often more than one kind of possible *right answer* in the result set. Nearly half of the queries were made of only one keyword and scarcely one third of the queries of two terms: only fourteen queries contained three or more keywords.

Processing queries

The queries were processed in January 2003 with the help of WinHTTrack Website Copier 3.22-3, an open source easy-to-use offline browser utility on a Windows XP system. It was fed with an HTML page that contains the query URLs of the two queries for the pre-test. The aim was to save the result lists of the engines and also the results themselves within a local folder with minimal time delay and true to the original as possible. Errors in the test results that could be traced back to deviations between query processing date and relevance judgement date should be avoided as much as possible. Furthermore the test collection, that is query URLs, search engine result pages and result pages should be preserved in a persistent way to allow further analysis after the tests were executed. In processing the two pre-test queries the following problems arose:

1. redirects
2. absolute URLs within the domain of the result pages
3. dynamic elements on the result pages like javascript code.

Because of the highly volatile nature of the Web all results had to be checked immediately after query processing, and if there were problems they had to be manually corrected. This was a time-consuming task. Every corrected page that could not be originally rebuilt was reviewed until the researchers were sure that there would be no deviance of the relevance judgements caused by insufficient replication of the result page. The fifty test queries were processed on 15th January, 2003 and checked for errors and corrected on the 16th and 17th. Therefore there were some differences possibly due to the fact that result pages could have been modified by its vendors unknown to the researchers.

Experimental design

The experimental design in this evaluation is equivalent to a *repeated measures design* (Tague-Sutcliffe, 1992). The jurors had two tasks. In the first part each juror had to judge the results one to twenty, one after the other, in the order given by the search engine for each engine successively for a certain query, not knowing which hits were from which engines. This is shown in Figure 4.

In the second part each juror had to judge the result presentations one to twenty for each engine for a different query, this time knowing which result presentation belonged to which engine. This is shown in Figure 5.

This means one could conclude that preferences and indispositions may have influenced the relevance judgements of the result presentations. This problem could not be solved. One may further ask, if the jurors judged the same query, i.e., first the real result of query X and then the result presentations of query X, would not the relevance assessments be more consistent? The answer is yes, and that is a big problem in this experimental design. It should be mentioned that there is a trade-off between consistency and neutrality in this experimental design. Learning effects could also appear on identical results or very similar result presentations between different engines. They should be compensated for by varying the engine order for each query.

Data collection, analysis and presenting results

Two kinds of data were collected. First, the personal juror data concerning age, sex, occupation, computer and search engine knowledge as well as usage and user affection before and after the test was documented. This allowed analysis of the competence profiles of the jurors and checked if they were really experienced with computers and search engines and were able to be rated as experts. Analyzing the participants' state of mind before and after the test gave a rough picture as to if and if so, to what extreme, the jurors found the test to be arduous.

Secondly, the relevance judgements were collected. The judgements of the results were considered to be *real* results because the jurors did not know their origin. For that reason it is assumed that the neutrality of theses relevance assessments was higher than that for the judgements on the result presentations because the source delivering the result presentations was apparent. Next, relevance judgements of the results and results representation were compared with each other. The aim was to find out if relevance assessments of the results representations and the results differ in a perceivable way. Result pages that were judged *links to relevant page(s)* were added to the *relevant* pages to achieve a simple binary benchmark scale. Finally, the retrieval performance was measured on the basis of the top-twenty precision. Micro- and also macro-precision were analyzed. This main result was statistically validated with the help of the sign test. Additionally, the *number of answered queries* and the *number of retrieved items* were observed.

Pre-test and test

Pre-test

A pre-test was done on 19th January, 2003 with the help of a member of the staff of the Working Group of Information Science who served as a juror. Queries were *gesundheit* (health) and *greenpeace*. The duration of the pre-test was less then one hour. The pre-test juror found the test to be easy and to be neither arduous nor tiring.

Test

The real tests were conducted within four days, between 21st and 24th January. Two tests had to be repeated due to technical problems with the display of four results. Otherwise the tests proceeded without problems. Test length varied from one-half to one-and-a-half hours per juror.

The tables below show the data about user state of mind before and after the test.

It deteriorated in an acceptable manner; most jurors felt one scale grade worse after the test than before. Some even felt better. On a five-point scale from -2 'I'm feeling very bad' to +2 'I'm feeling very well', the average fell between +0,66 and +0,08.

Answers to the question "how do you feel" before the test				
-2	-1	0	1	2
0	5	13	26	6

Table 9: Juror feeling before the test.

Answers to the question "how do you feel" after the test				
-2	-1	0	1	2
2	9	15	17	7

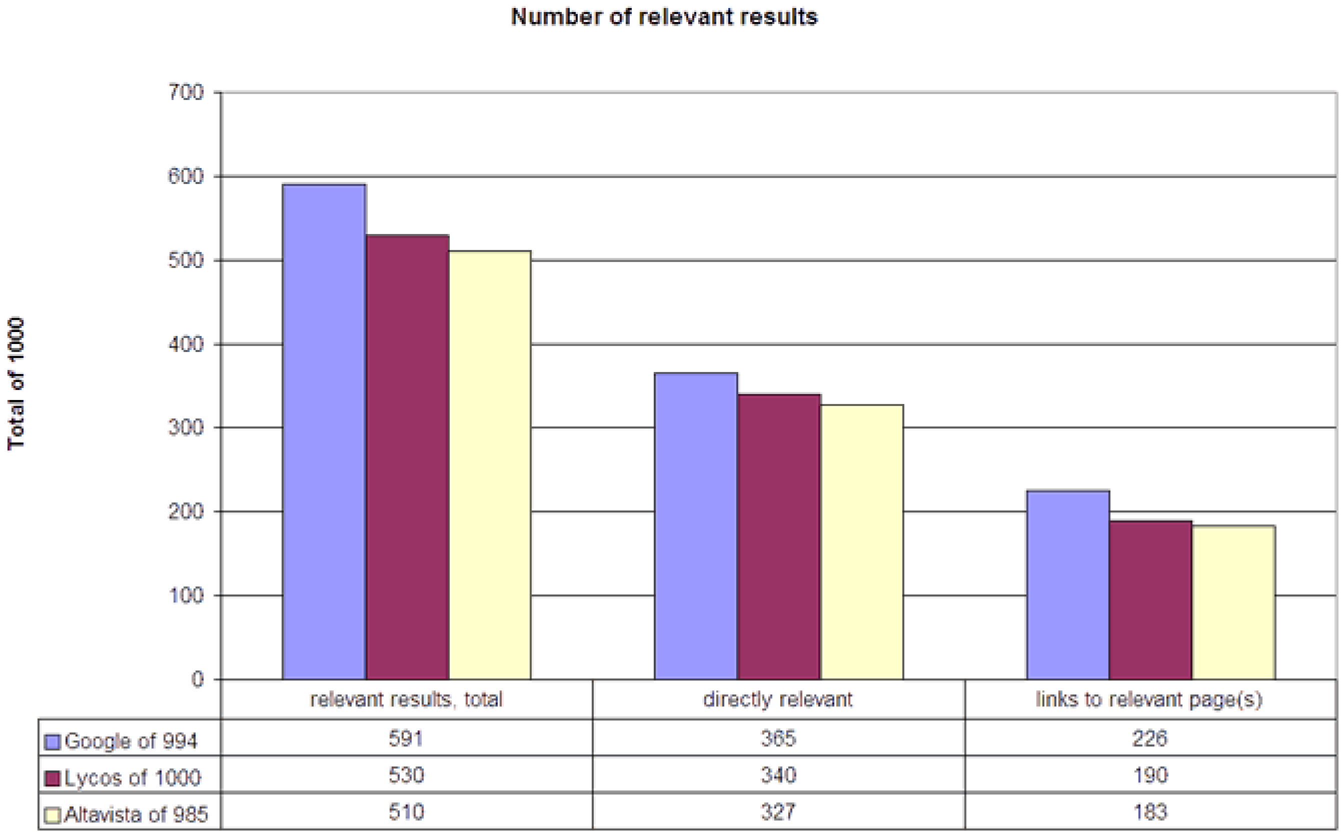
Table 10: Juror feeling after the test.

Results

The test results were analyzed as follows: first, the number of relevant hits was counted separately for the results and for the results representations. Possible deviations are discussed closer in order to examine possible reasons for the differences. Then the micro- and macro-precision for the results themselves were calculated, delivering the core findings of this evaluation whose purpose was to point out which engine delivers the highest number of relevant results for all queries and which engine answers the queries best. These main results are statistically verified with the sign test, giving the overall results of this evaluation. After that the *number of retrieved items* and the *number of answered questions* were counted to show coverage deviations between the engines as well as the number of queries that were answered in an at least marginally useful way.

Number of relevant hits and number of relevant result presentations

The overview of the number of relevant results shows how many results and result presentations of the maximum result number of 1000 hits were judged as *relevant* or as *links to relevant page(s)*.



	relevant results, total	directly relevant	links to relevant page(s)
🟦 Google of 994	591	365	226
🟥 Lycos of 1000	530	340	190
🟨 Altavista of 985	510	327	183

Figure 6: Number of relevant results in proportion to the possible number of 1000 results

Google is the best engine: it delivers the highest number of results that are 'directly relevant' as well as the highest number of 'links to relevant page(s)', thus leading the field with a total of 591 relevant results followed by Lycos with 530 relevant results and then Altavista in last place with 510 relevant results. The first finding of the test is that

Google delivers more relevant results than the other engines.

The overview of the number of relevant result presentations shows how many results were assumed to be relevant by judging the presentations of the results on the search engines result pages.

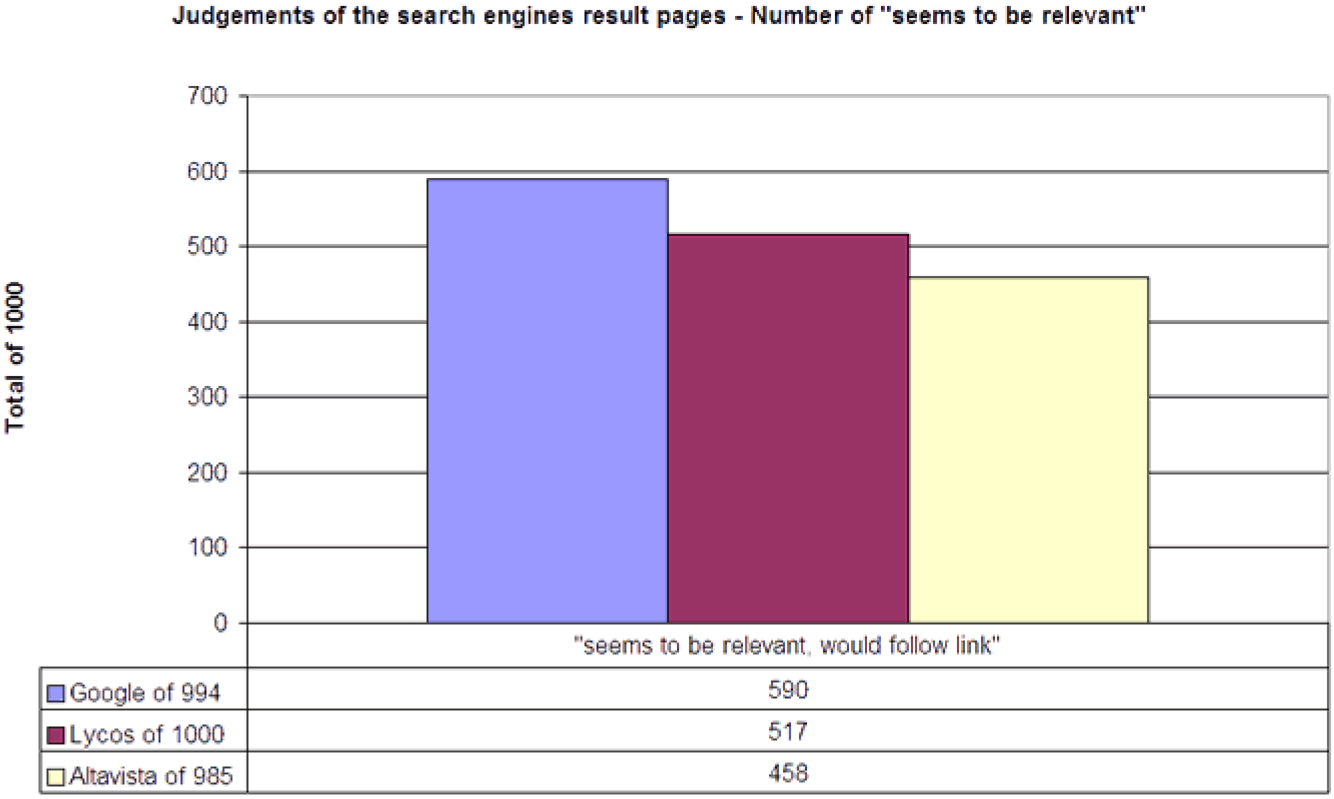
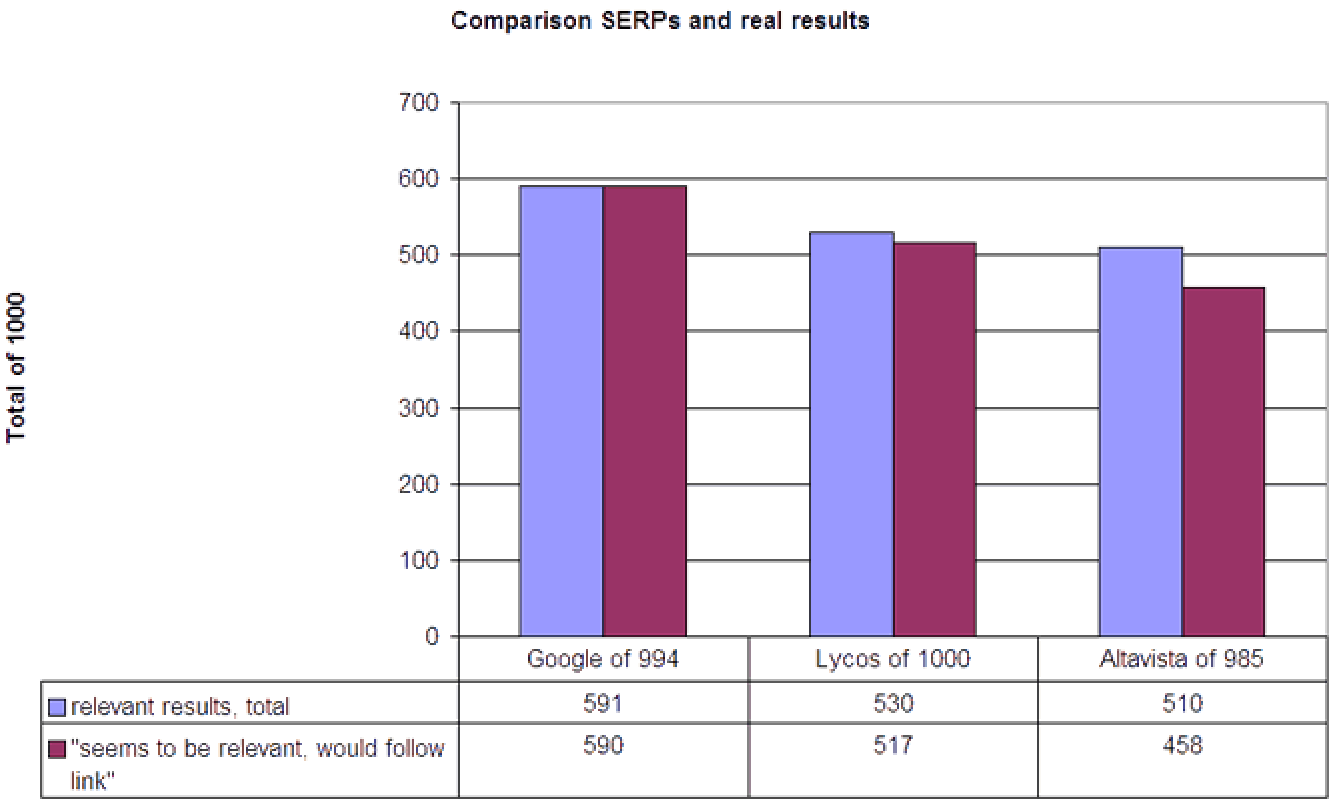


Figure 7: Number of 'seems to be relevant' result presentations in proportion to the possible number of 1000 results

The judgement of the result pages gives a picture very similar to the judgements of the results. The order of the engines is the same. Google reached the highest values with 590 'seems to be relevant' result presentations. Lycos reached a score of 517 whereas Altavista got only 457 positive assessments.

Whereas the sum of relevant hits and relevant result presentations was nearly the same for Google and relatively equal for Lycos, the difference for Altavista was more than 5%.

The comparison of both kinds of search engine effectiveness can be found in the following summary in Figure 8.



	Google of 994	Lycos of 1000	Altavista of 985
🟦 relevant results, total	591	530	510
🟥 "seems to be relevant, would follow link"	590	517	458

Figure 8: Number of *seems to be relevant* result presentations in proportion to the possible number of 1000 results

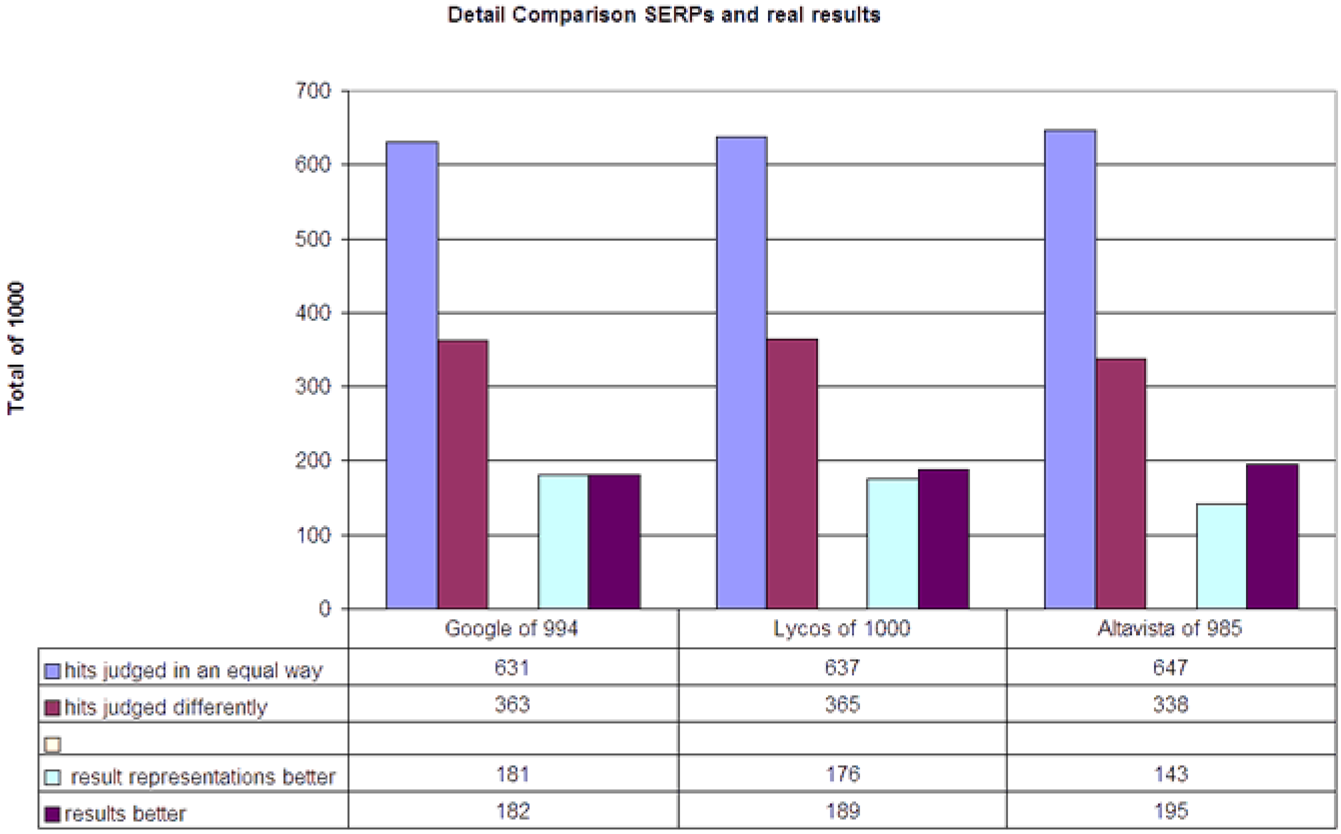
What do these results mean? At first sight it seems that Google's and Lycos's representations of results and results retrieval effectiveness are largely congruent and that Altavista has a) trouble with its result presentations or b) the jurors' judgements of Altavista result interpretations are at least partially influenced by preferences and indispositions in a negative way or a mix of both possibilities. Such an interpretation is appropriate to a certain degree, but things are a little bit more complicated. First of all, as mentioned above, the jurors for the results and the result presentations for the same query were different people. The advantage of this adjustment is that learning effects are minimized. The big disadvantage is that there is a much higher probability that relevance judgements are more deviant than they would be if the results and the presentations of results were judged by the same person because of different preferences (Schamber, 1994: 11). It can be assumed that preferences and indispositions of the single jurors towards the search engines influenced the assessments of the result presentations. One the one hand, this could mean that formal and contextual differences in the presentation of results are able to bias the judgements. For example, two jurors remarked that the result presentations of Lycos are better because the text description of these results enabled them to judge more accurately than the text descriptions of Google and Altavista whether the corresponding results were really relevant. One can imagine that identical result presentations could be judged diversely on different engines because the juror thinks that one engine generally delivers very good results while the other engine is generally spammed.

Before reading and interpreting the following numbers it is important to keep in mind that these figures are not proven to be valid but are at least partially influenced by the following factors

- The judgement of the results and the presentations of results were made by different jurors.
- There are diversities in the formal and contextual result representations of the different search engines.
- There are diverse preferences of jurors regarding the different search engines.

Nevertheless the figures give some interesting insights. As said before the sum of relevant hits and relevant result presentations were nearly the same for Google and relatively equal for Lycos, whereas the sum of the difference between the judgements of relevant hits and relevant result presentations was more than 5% for Altavista. However, the judgement of the single hits was different in roughly one third of the cases for all three engines, even for Google.

Of all possible 1000 results, Google delivered 994 hits of which only 631 were judged in a consistent manner and 363 were judged differently. These 363 assessment deviations were composed of 181 items in which the results were judged as *not relevant* and the result presentations were judged as *seems to be relevant, would follow link,* and 182 items in which the results were judged either as *relevant* or *links to relevant page(s)* whereas the result presentations were judged as *seems not to be relevant, wouldn't follow link*. For Lycos, 365 hits were also judged differently and 637 equally. In 176 of the cases the result presentations were judged to be better than the results and 189 cases were converse. For Altavista, 338 hits were judged differently and 647 hits were given similar ratings. In 143 cases the result representations were judged as being better than the results and in 195 cases the results were judged better as the result representations.



	Google of 994	Lycos of 1000	Altavista of 985
■ hits judged in an equal way	631	637	647
■ hits judged differently	363	365	338
□			
□ result representations better	181	176	143
■ results better	182	189	195

Figure 9: Comparison of the deviations of positive relevance judgements for the results and the search engine presentations of results

The tentative results of these comparisons are:

1. The relevance assessment of results and result representations deviated in roughly one third of the cases for all engines.
2. For Google and Lycos the deviations were mutually balanced.
3. Altavista delivered more relevant results than it was assumed to deliver.

Though tentative, these results indicate that further investigation and comparison of the effectiveness of the results and the result representations on the search engine result pages could be actually worthwhile. A closer look at the deviations should give hints as to why one third of the results and result presentations were judged differently.

An analysis of the '404 errors' gives following picture.

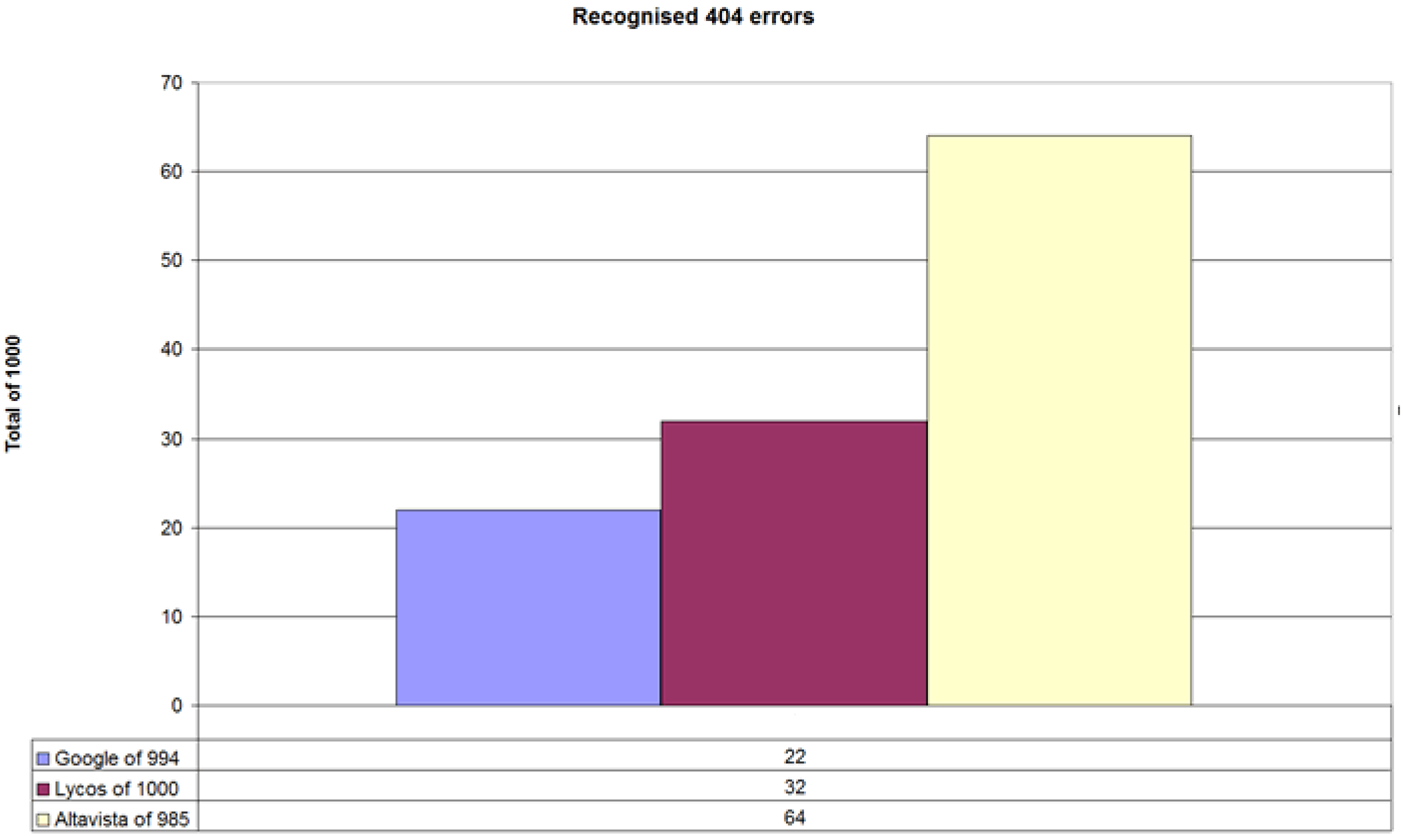


Figure 10: '404 page not found' errors

Google is the engine with the smallest number, namely twenty-two, of the recognised '404 errors', two of which were even judged as *relevant* or *links to relevant page(s)*. Lycos followed with thirty-two and Altavista was again in last place with a clearly worse score of sixty-four '404 errors', one of these was assumed to be relevant. Only a small number of these '404 error' results were judged as *seems to be relevant*, *would follow links* on the search engines' result pages. This means the so called '404 errors' are in no way a sufficient explanation for deviations between the results and the result presentations assessments. They made up only for a small fraction of the cases in which the result representations were given more positive judgements. These kinds of the deviations can not be easily explained and deserve further and more elaborate studies which can not be conducted here.

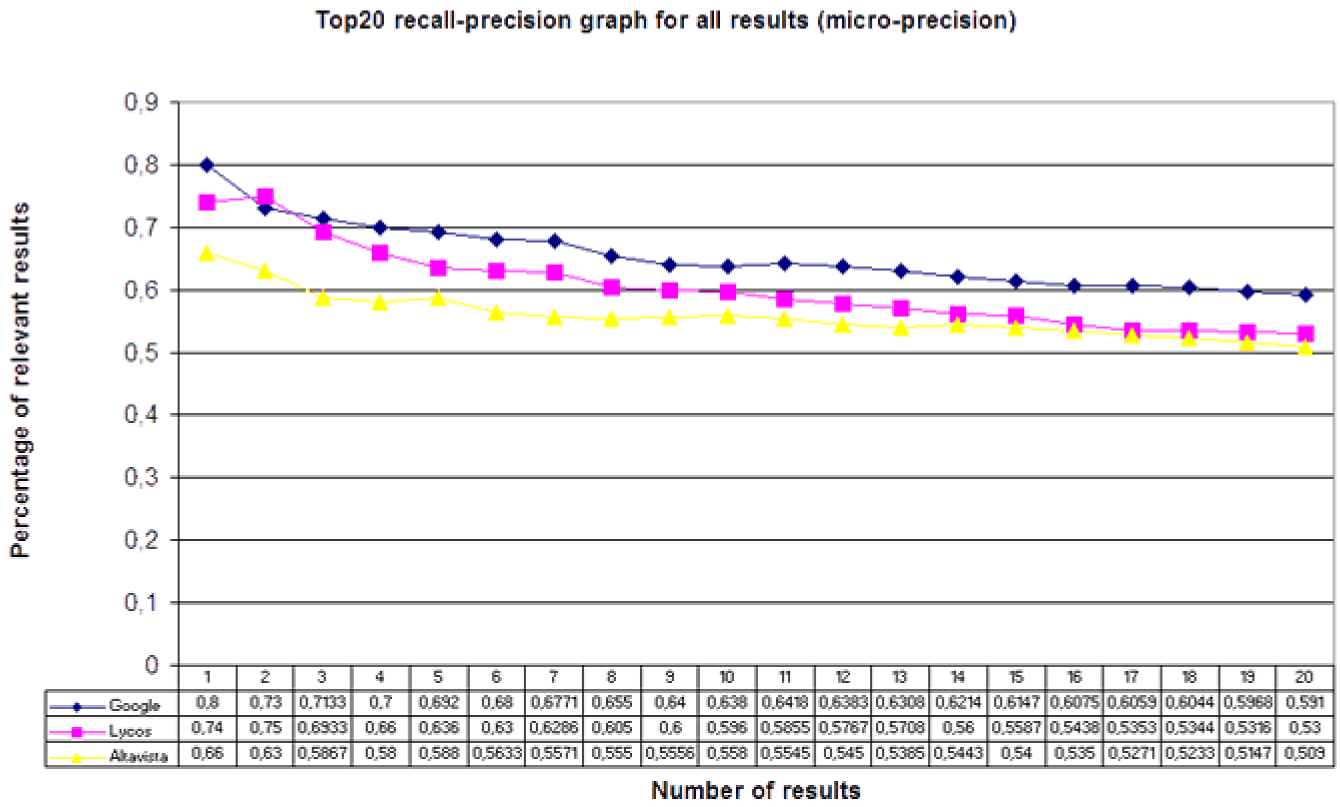
There is no room here to examine the possible reasons for why results are judged as relevant while their representations are not. Detailed views seem to suggest that pages which get a rather *broad* topical page description focusing not only on the searched topic are rather judged as not relevant even if the content of the results is judged as relevant.

The following analysis will only consider the real results, because it is assumed that they give a more trustworthy picture of the retrieval effectiveness, in that the relevant judgements are not biased by preferences and indispositions concerning their delivering engine.

Mean average precision of the results (micro-precision view)

As mentioned previously Google delivers for the first twenty positions of all queries the highest number of relevant results. The micro precision shows how these relevant results spread among the single top twenty positions. It indicates differences in comparison of the effectiveness of the top twenty overall results and, for example, for the top ten results, the number of results most searchers take care of (Jansen *et al*, 2000).

The following recall-precision graph of the top twenty positions displays the number of relevant results for the corresponding top x views. For example, the top ten precision shows the percentage of results that are either judged as *relevant* or judged as *links to relevant page(s)* for the first ten positions as a whole.



	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
▲ Altavista	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

Figure 11: Top-twenty recall-precision graph for all results. Percentage of the number of relevant results in proportion to all results at the corresponding positions.

Figure 11 shows that Google reached the highest micro precision. The percentage of relevant results in the first position is 80%. The percentage of the sum of the top ten results rated as relevant is 63.8%. The percentage of the sum of the top-twenty results rated as relevant is still 59.1%. Google reached a higher effectiveness for all top-one to twenty values, with the exception of the top-two precision in which Lycos performed slightly better with 75% versus 73% of relevant results. Lycos scored the second highest values with 74% of relevant results for the top one, 59.6% for the top ten and 53% for the top twenty precision. Altavista was rated last with 66% relevant results for the top one, 55.8% for the top ten and 50.9% for the top twenty precision scores. The conclusion is, at every cut-off rate for the first twenty results Google retrieves for the most part a greater number of relevant hits than the other engines, followed by Lycos which always reaches higher values than Altavista. At first glance the reached values seem to be very high, 80% relevant results with Google for the first position and still far over 50% for all results within the first twenty positions. Could this mean that the engines are very good, or do these extraordinary high values indicate that there is something wrong with the experimental design? The answer is neither yes nor no. The values are rather an indicator of the differences in effectiveness between the engines and could be regarded as a measure of global retrieval quality only to a minor extent. Recall and precision values are, among other things, predetermined by the kind of queries, i.e., specificity. The used queries are rather short and, in most cases, rather unspecific. Furthermore the reconstructed information needs are rather broad and not restricted to an *in most cases right* subset, so there is a possibility that they reflect a wider context than the basic information need of a real searcher for the same query, cp. the example *travel*. This means the absolute quantitative values are probably biased in favour of the engines in general. For this reason, any interpretations of the results should focus on the differences between the engines. These are assumed to be valid because the effects of these discrepancies from reality affect all engines widely in the same way.

The first core finding of this investigation is: Google delivers a greater number of relevant hits than the other two engines with only one exception if users consider selecting only the first two results, then Lycos seems to be slightly better.

Lycos is awarded the second place because it always delivers a greater number of relevant items than Altavista. If we take the *mean average precision* as the average of the precision values at the top one to twenty positions then we get the following values in Table 11.

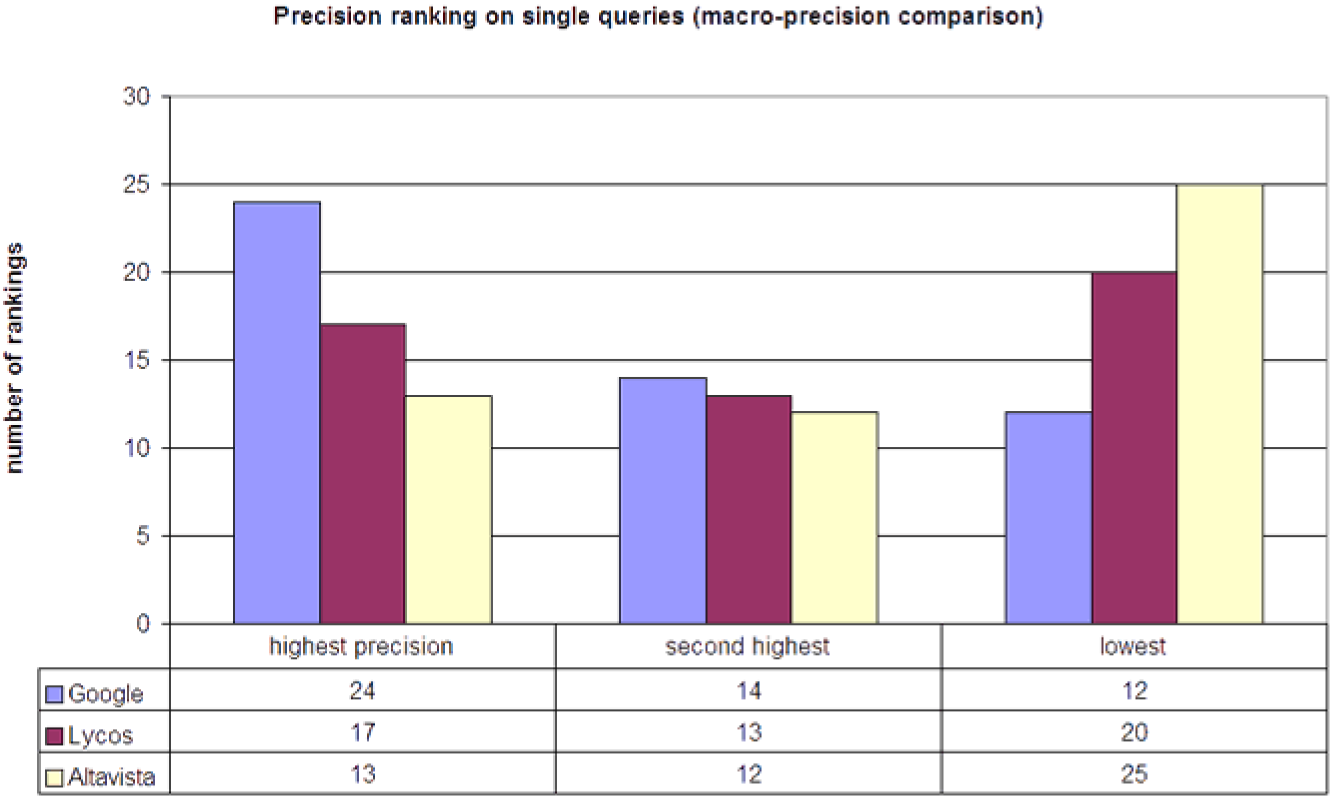
Search engine	Mean average precision
Google	0.65
Lycos	0.60
Altavista	0.56

Table 11: Mean average precision for top one to twenty precision.

These measurements illustrate this first core finding in one single date. Google reaches the highest values, followed by Lycos and then Altavista.

Answering queries (macro-precision view)

If Google delivers the highest number of relevant results, does this mean that it is the engine with the highest effectiveness? Not necessarily, because it is possible that one engine delivers more relevant results in the total over all queries but answers the single queries with minor effectiveness (Griesbaum, 2000: 67). To answer the question as to which engine answers the single queries best, the precision values concerning the single queries for each engine were compared with each other.



	highest precision	second highest	lowest
■ Google	24	14	12
■ Lycos	17	13	20
□ Altavista	13	12	25

Figure 12: Number of rankings in reference to the precision values reached on the fifty queries.

Highest precision means the engine reached the highest precision value concerning a certain query in comparison with the precision values of the other engines. *Second highest* means the engine reached the second highest precision values comparatively for a certain query. *Lowest* means the engine reached the lowest precision values comparatively for a certain query. For example: for the query *reisen* (travel), Google reached a precision of 0.8, Lycos a precision of 0.95 and Altavista a precision of 0.85. Therefore Lycos got *highest precision*, Altavista *second highest* and Google *lowest* ranking for this query.

Figure 12 indicates that Google reached the highest values for macro-precision view, too. It answered the queries best twenty-four times, second best fourteen times and had the lowest precision values on only twelve occasions. Google was followed by Lycos, which reached the highest ranking seventeen times, the second highest thirteen times and the lowest twenty times. Again, Altavista seems to be the engine with the lowest effectiveness. It got the highest ranking thirteen times, the second highest seventeen times, and the lowest ranking half of the time for twenty-five queries.[4]

The second core finding of this investigation is: Google seems to satisfy information needs best, because it delivers the highest numbers of relevant results for the single queries when compared with the other two engines. Lycos reached second place and, again, Altavista was last.

In terms of micro- and macro-precision, the core finding of the test seems to be:

Google is the engine which performs best, followed by Lycos and then Altavista. This is the same ranking as in the preceding retrieval test in 2002 (Griesbaum *et al.*, 2002).

The question arises if the differences of the retrieval effectiveness are high enough to support this result as a valid statement. For this purpose the results were statistically validated with the sign test (Siegel, 1987). According to the sign test, only the differences between Google and Altavista are significant, whereas the differences between Google and Lycos, and Lycos and Altavista are too little to be statistically validated. This is different from the preceding evaluation, in which Google's higher retrieval effectiveness was statistically validated in comparison to all other engines. What can we conclude? According to the numerical result values Google again seems to be superior to the other engines but the gap between Google and Lycos is no longer statistically sound. This means there is a high probability that the gaps between Google and its competitors are decreasing.

Coverage – number of answered queries and number of items retrieved

The question of how many queries can be answered by the engines is an important one. It shows how often the engines are actually helpful for searchers in so far as they return something at least minimally relevant within the first twenty hits. In this evaluation the number of answered queries was taken as the number of queries to which the engines delivered at least one relevant result. The values are displayed in Table 12.

Search engine	Number of answered queries	Number of not answered queries
Google	50	0
Lycos	50	0
Altavista	48	2

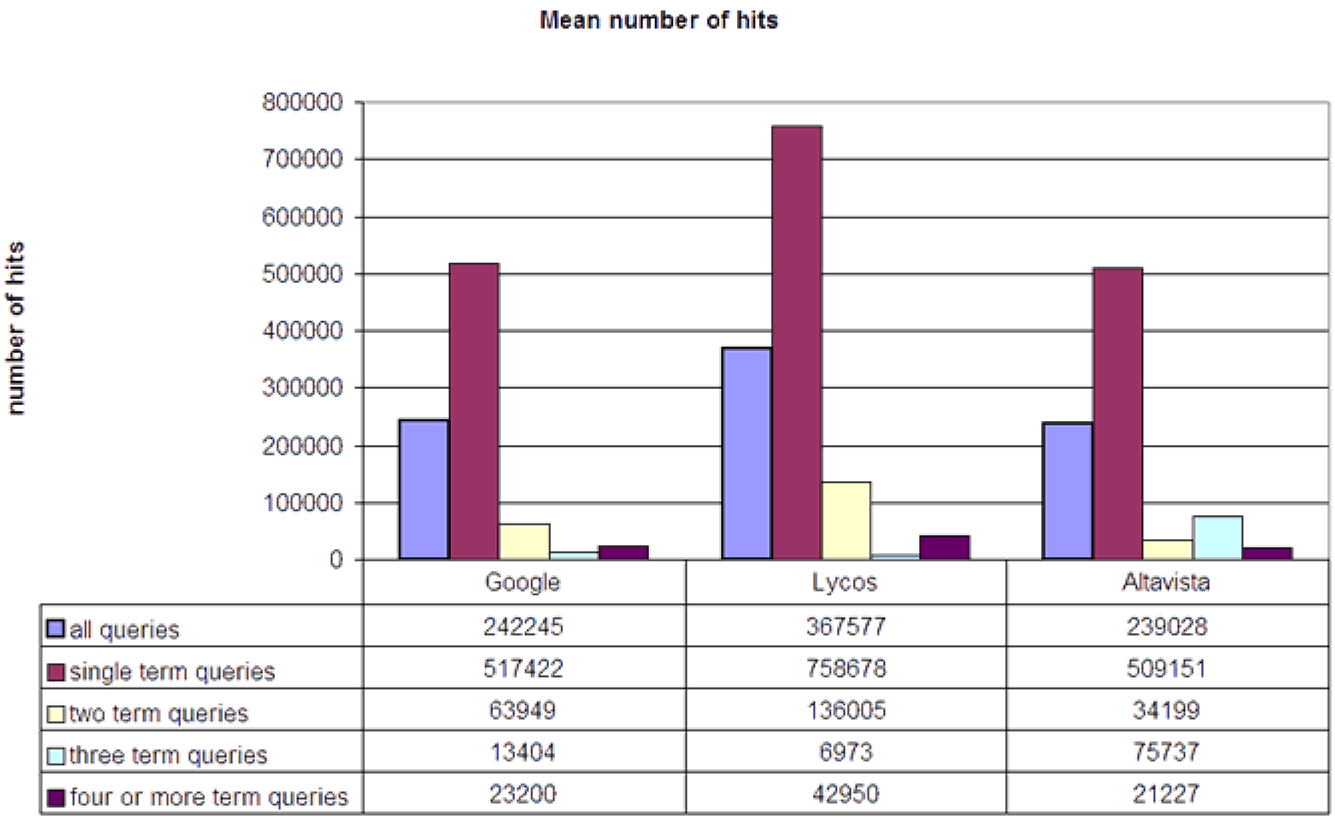
Table 12: Number of answered queries and number of items retrieved.

For all of the queries, Google and Lycos retrieved at least one relevant result. Altavista failed to answer queries two times. This suggests that the engines are capable of answering typical Web search queries in a somewhat helpful way in nearly all cases. The two queries that were not answered by Altavista were the queries *hotmail* and *sonja krauss*. Although Altavista returned more than twenty hits for both queries, none of the first twenty results was assessed as being *relevant* or as being l*inks to relevant page(s)*.

Only Lycos always delivered the full number of twenty results per query, whereas Google delivered only fourteen results for the query *kinofilm volcano video dvd*, thus retrieving 994 of the 1000 possible results within the first twenty positions. Altavista delivered the smallest number of results within the test collection. It retrieved only fourteen results for the query *reiseinformation st. lucia* and only eleven results for the query *kinofilm volcano video dvd*. So within the first twenty evaluated positions, Lycos delivered the greatest number followed by Google and then Altavista.

Nevertheless the question as to which engine enables access to the greatest part of the Web is difficult to answer. Each engine releases figures about its index size. Furthermore the results which searches actually retrieve are dependent on additional factors, document representation for instance.

The coverage compared as the sum of the number of items retrieved indicates the following figure.



	Google	Lycos	Altavista
□ all queries	242245	367577	239028
■ single term queries	517422	758678	509151
□ two term queries	63949	136005	34199
□ three term queries	13404	6973	75737
■ four or more term queries	23200	42950	21227

Figure 13: Number of retrieved items. Mean for all queries and mean for different query term frequencies.

Figure 13 shows that Lycos retrieved the greatest number of hits, with an average of about 368,000 hits per query, followed by Google with an average of 242,000 hits and Altavista with an average of 239,000. The numbers roughly doubled for single term queries and grew smaller the more terms were used [5]. In numbers of declared hits it seems that Lycos disposes the biggest index. This finding corresponds to the fact that only Lycos constantly delivered the full number of twenty results per query.

Conclusions

The first and main result of this inquiry is, Google reached the highest values, followed by Lycos and then Altavista. Google scored the highest number of relevant items across all fifty queries and achieved the highest top twenty precision values for nearly half of the queries. But the differences are rather low. In fact they are so low that a statistical validation with the help of the sign test indicates that Google performs significantly better than Altavista, but there is no significant difference between Google and Lycos. Although Lycos also attains better values than Altavista, the differences between these two engines reach no significant value. The conclusion is that the effectiveness of the engines in answering queries with relevant results from position one to twenty is very close to each other, with the one exception that Google seems to be clearly better than Altavista. To quote Sullivan again. 'If it turns out that relevancy testing finds that Google and its competitors are all in roughly the same degree of relevancy, then users might then be willing to experiment more with others' (Sullivan, 2002, December 5)

Hence, one conclusion of this evaluation is that ,users could, and should, contemplate Lycos as a real alternative to Google, especially because both engines were able to answer all of the 50 queries with at least one relevant result. Beyond that, Lycos was found to retrieve the greatest number of hits. This indicates at least Lycos' coverage is not worse than Google's coverage. This is the opposite case if one regards only the official number of indexed Web pages on both engines. Lycos results relied heavily on the AlltheWeb search engine. If we keep in mind that AlltheWeb, Altavista and Inktomi too are now owned by Yahoo, it will be very interesting to see what kind of new search engine will come into being if the capabilities of these three engines are combined.

The second and supplementary result of this evaluation is: there are big deviations between the relevance judgements of the results themselves and the judgements of the result representations on the search engine results pages. The overall results in regards to the sum of all relevant items are about the same size, but the assessments of the single results are dissenting in about one third of the cases. For Google and Lycos the deviations widely cancel each other out, but Altavista result presentations are more often judged worse than Altavista results rather than the other way round. A closer look at the deviations between relevance assessments of results and representation of results on the search engine result pages (SERPs) makes it apparent how tentative these supplementary test results are. Results and result presentations are judged by different people. Hence the consistency of the judgements is debatable. It seems probable that the experimental design distorts the results. Nevertheless, there is an astonishing number of deviations, and it could be dangerous to allocate them only to an insufficient experimental design. The great variation of relevance assessments among the results themselves and their representation suggests that further research is necessary to determine the reasons and consequences of this rather unexpected observation.

In contrast to the supplementary results, the values concerning the comparison of the effectiveness of only the results between the different engines are seen as valid and reliable. It is important to note that there are also problematic aspects. Hence there is also room for improvement.

1. Retrieval effectiveness results were based only on the judgement of the real results of the engines. Hence, surplus values of the result presentations that influence a real user's decision as to which hits to select are as equally ignored as features that influence the search process as a whole. Examples of this are: direct links to Web catalogue result categories or the Google archive.
2. Although randomly selected, real Web search queries were employed, the question remains: Are real Web search contexts actually reflected by the selected queries? Is the query collection as a whole a representative sample? Furthermore, the information needs were rebuilt. It was done carefully in different steps with several people to avoid biases. But, for rather unspecific queries, the reconstructed information needs were defined rather broadly. Therefore, it is sensible to assume that the relevance judgements were distorted in a positive way because there was a tendency that they reflected rather a somewhat semantic similarity than intentions of real searchers.
3. The evaluation criterion was relevance and relevance is very difficult to measure. It is bound to the subjective views, circumstances and knowledge of those users who served as jurors, and it is difficult to generalize. The goal of this evaluation was not to measure retrieval effectiveness as an absolute value, but to determine mainly the differences between the engines.
4. The qualitative evaluation criterion is limited to the number of somewhat relevant results which crossed the threshold from being not-relevant. Differences in the degree of relevancy were not measured. It is possible to measure the degree of relevance additionally. This point should be noted for further evaluations. The researchers would be able to detect differences in various kinds of relevancy.

This evaluation tried to achieve results as valid and reliable as possible within the given resources. The available knowledge and evaluation experience was used to consider Web-specific retrieval peculiarities and evaluation standards in, one hopes, a sufficient manner. The goal was to develop the experimental design in the best possible way, within the given resources. It is assumed that the results are valid and reliable enough to get some helpful insights into the effectiveness of the investigated search engines. The mentioned problem fields show, on the one hand, the constraints of this evaluation and, on the other hand, give some hints on how to improve further investigations.

Notes

1) Validity is the extent to which the experiment results are correct, whereas reliability is the extent to which the results can be replicated. (Tague-Sutcliffe, 1992: 467).

2) 'The act of getting a search engine to record content for a URL that is different than what a searcher will ultimately see. It can be done in many technical ways.' (Sullivan, n.d.)

3) After modifications by Ask.com on www.askjeeves.com this service has changed and is no longer available as described above.

4) The problem of zero relevant item answers/result sets is ignored at this point. (For further information cf. Womser-Hacker, 1989: 151-156).

5) One should note that in the German language there is no strong connection between number of query terms and

specificity, unlike e.g., in English, because a large part of single terms are pre-coordinated. Example: query *uftverschmutzung* (air pollution). That means even single term queries could be very specific.

Acknowledgements

The author wishes to thank all the people who supported this evaluation and the suggestions of the anonymous referees. Special thanks to Anna Weber, without whose help this study would not have been possible.

References

- Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology*, **53**(4), 308-319.
- Buckley, C. & Voorhees, E.M. (2000). Evaluating Evaluation Measure Stability. In N. J. Belkin, P. Ingwersen, M.-K.Leong and E. Yannakoudakis (Eds.); *Proceedings of SIGIR'00*, (pp. 33-40). New York: ACM Press.
- Cleverdon, C. W., Mills, J. & Keen, M. (1966). *Factors determining the performance of indexing systems*, Vol. 1: Design, Vol. 2: Test results. Cranfield, UK: College of Aeronautics.
- Dennis, S., Bruza, P. & McArthur, R. (2002). Web searching: a process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, **53**(2), 120-133.
- Eguchi, K., Oyama, K., Ishida, E., Kando, N. & Kuriyama, K. (2002). Overview of the Web retrieval task at the Third NTCIR Workshop. *Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering (September 2001-October, 2002)*, Tokyo: National Institute of Informatics (NII). Retrieved 27 April, 2003 from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-WEB-EguchiK.pdf.
- Ford, N., Miller, D. & Moss, N. (2001). The role of individual differences in Internet searching: an empirical study. *Journal of the American Society for Information Science and Technology*, **52**(12), 1049-1066.
- Gordon, M. and Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, **35**(2), 141-180.
- Griesbaum, J. (2000). *Evaluierung hybrider Suchsysteme im WWW*. Unpublished diploma thesis, Universität Konstanz, Konstanz, Germany. Retrieved 28 April, 2003 from http://www.inf.uni-konstanz.de/%7Egriesbau/files/evaluierung_hybrider_suchsysteme_im_www.pdf.
- Griesbaum, J., Rittberger, M. & Bekavac, B. (2002). Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de. In R. Hammwöhner, C. Wolff, and C. Womser-Hacker (Eds.); *Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information, Proceedings des 8. Internationalen Symposiums für Informationswissenschaft*, (pp.201-223). Konstanz: UVK, 201-223.
- Gurrin, C. & Smeaton, A. (2003). Improving the evaluation of Web search systems. In F. Sebastiani (Ed.), *Advances in information retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003. Proceedings*, (pp.25-40) Berlin; New York: Springer. (Lecture Notes in Computer Science **2633**)
- Harter, S.P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, **47**(1), 37-49.
- Hawking, D., Craswell, N., Bailey, P. & Griffiths, K. (2001). Measuring search engine quality. *Journal of Information Retrieval*, **4**(1), 33-59.
- Inktomi Corp. Web search relevance test. (2003). Retrieved 6 May, 2003 from the Veritest Website at http://www.veritest.com/clients/reports/inktomi/inktomi_Web_search_test.pdf
- Jansen, B., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing & Management*, **36**(2), 207-227.
- Kowalski, G. (1997). *Information retrieval systems, theory and implementation*. Boston, MA: Kluwer Academic Publishers
- 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.
- Lesk, M. (1995). The seven ages of information retrieval. In *Proceedings of the Conference for the 50th anniversary of As We May Think*. (pp. 12-14). Cambridge, MA: MIT Press.
- Mandl, T. (2003). Web- und Multimedia-Dokumente. Neuere Entwicklungen bei der Evaluierung von Information Retrieval Systemen. *Information: Wissenschaft & Praxis*, **54**(4), 203-210.
- Robertson, S.E. (1981). The methodology of information retrieval experiments. In K.S. Jones, (Ed.);

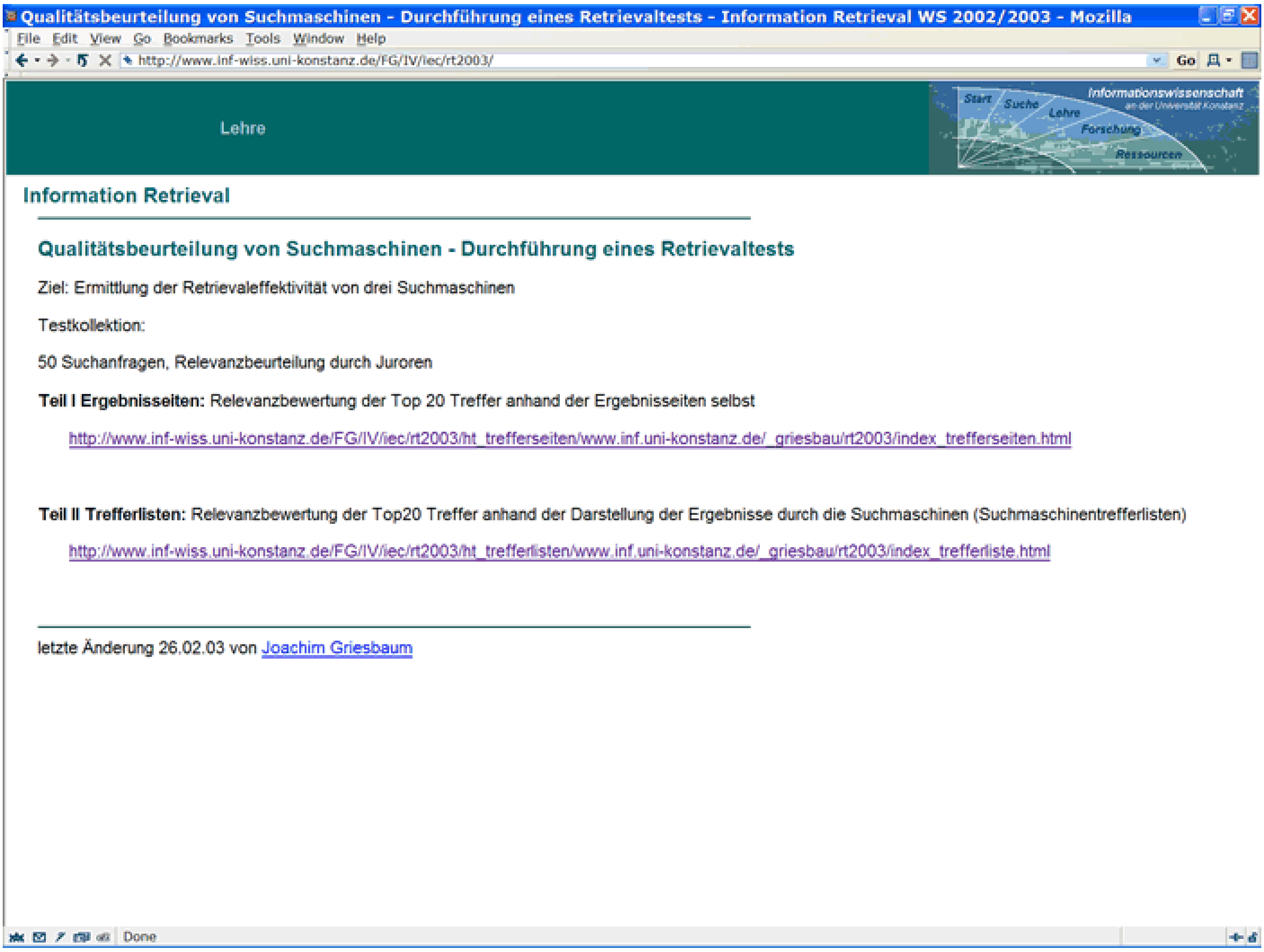
Information retrieval experiment. (pp.9-31) London: Butterworth. 9-31.

- Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*. **29,**, 3-48.
- Shang, Yi. & Longzhuang, Li. (2002). Precision evaluation of search engines. *World Wide Web*, **5**(2), 159-179.
- Siegel, S. (1987). *Nichtparametrische statistische Methoden* (3rd ed.). Eschborn bei Frankfurt am Main: Klotz
- Spink, A. (2002). A user centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing & Management*, **38**(3), 410-426.
- Sullivan, D. (2002, December 5). In search of the relevancy figure. *SearchEngineWatch.com* Retrieved 24 April, 2003 from http://www.searchenginewatch.com/sereport/article.php/2165151
- Sullivan, D. (2002, April 29). Jupiter MMXI European search engine ratings. *SearchEngineWatch.com* Retrieved 19 April 2004 from http://Web.archive.org/Web/20030618111440/ http://www.searchenginewatch.com/reports/article.php/2156441)
- Sullivan, D. (2004, April 28). Major search engines and directories. *SearchEngineWatch.com* Retrieved 24 April, 2003 from http://www.searchenginewatch.com/links/article.php/2156221
- Sullivan, D. (n.d.). *Search engine optimization & marketing glossary.* Palo Alto, CA: SEMPO: Search Engine Marketing Professional Organization. Retrieved 19 April, 2004 from the SEMPO Website at http://www.sempo.org/search-engine-marketing-glossary.php
- Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, **28**(4), 467-490.
- Womser-Hacker, C. (1989).*Der PADOK Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information- Retrieval-Systemen*. Hildesheim Georg Olms.

Appendix

Test Web Site

The test Website can be found at http://www.inf-wiss.uni-konstanz.de/FG/IV/iec/rt2003/.


Qualitätsbeurteilung von Suchmaschinen - Durchführung eines Retrievaltests - Information Retrieval WS 2002/2003 - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://www.inf-wiss.uni-konstanz.de/FG/IV/iec/rt2003/
Go
Lehre
Start
Suche
Lehre
Forschung
Ressourcen
Informationswissenschaft
an der Universität Konstanz
Done

Information Retrieval

Qualitätsbeurteilung von Suchmaschinen - Durchführung eines Retrievaltests

Ziel: Ermittlung der Retrievaleffektivität von drei Suchmaschinen

Testkollektion:

50 Suchanfragen, Relevanzbeurteilung durch Juroren

Teil I Ergebnisseiten: Relevanzbewertung der Top 20 Treffer anhand der Ergebnisseiten selbst

http://www.inf-wiss.uni-konstanz.de/FG/IV/iec/rt2003/ht_trefferseiten/www.inf.uni-konstanz.de/_griesbau/rt2003/index_trefferseiten.html

Teil II Trefferlisten: Relevanzbewertung der Top20 Treffer anhand der Darstellung der Ergebnisse durch die Suchmaschinen (Suchmaschinentrefferlisten)

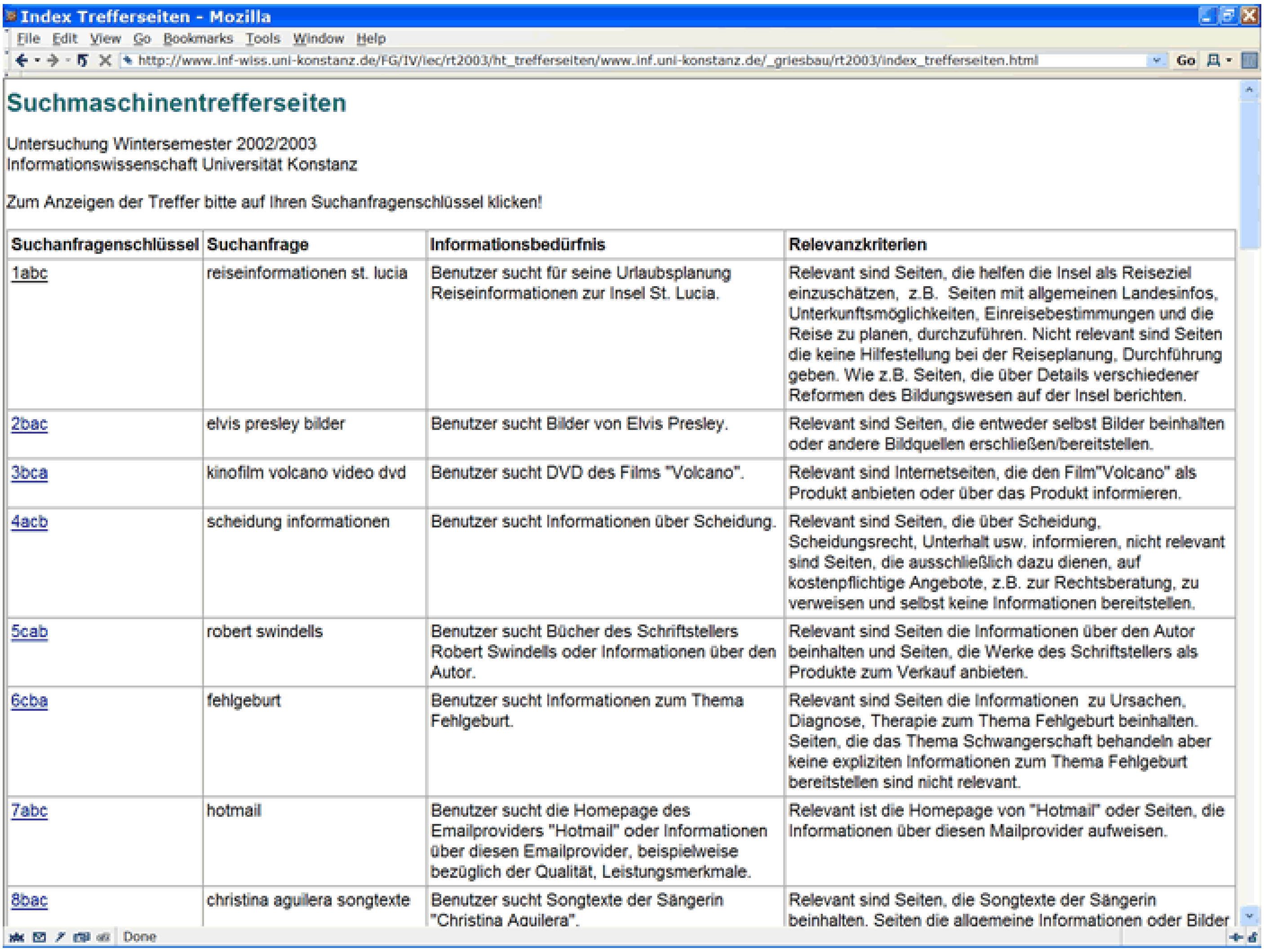
http://www.inf-wiss.uni-konstanz.de/FG/IV/iec/rt2003/ht_trefferlisten/www.inf.uni-konstanz.de/_griesbau/rt2003/index_trefferliste.html

letzte Änderung 26.02.03 von Joachim Griesbaum

Figure 14: Test Website.

Queries Information needs, relevance criteria

The whole query set containing information needs and relevance criteria can be found at http://www.inf-wiss.uni-konstanz.de/FG/IV/iec/rt2003/ht_trefferseiten/www.inf.uni-konstanz.de/_griesbau/rt2003/index_trefferseiten.html.


Index Trefferseiten - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://www.inf-wiss.uni-konstanz.de/FG/IV/iec/rt2003/ht_trefferseiten/www.inf.uni-konstanz.de/_griesbau/rt2003/index_trefferseiten.html
Go
Done

Suchmaschinentrefferseiten

Untersuchung Wintersemester 2002/2003
Informationswissenschaft Universität Konstanz

Zum Anzeigen der Treffer bitte auf Ihren Suchanfragenschlüssel klicken!

Suchanfragenschlüssel	Suchanfrage	Informationsbedürfnis	Relevanzkriterien
1abc	reiseinformationen st. lucia	Benutzer sucht für seine Urlaubsplanung Reiseinformationen zur Insel St. Lucia.	Relevant sind Seiten, die helfen die Insel als Reiseziel einzuschätzen, z.B. Seiten mit allgemeinen Landesinfos, Unterkunftsmöglichkeiten, Einreisebestimmungen und die Reise zu planen, durchzuführen. Nicht relevant sind Seiten die keine Hilfestellung bei der Reiseplanung, Durchführung geben. Wie z.B. Seiten, die über Details verschiedener Reformen des Bildungswesen auf der Insel berichten.
2bac	elvis presley bilder	Benutzer sucht Bilder von Elvis Presley.	Relevant sind Seiten, die entweder selbst Bilder beinhalten oder andere Bildquellen erschließen/bereitstellen.
3bca	kinofilm volcano video dvd	Benutzer sucht DVD des Films "Volcano".	Relevant sind Internetseiten, die den Film"Volcano" als Produkt anbieten oder über das Produkt informieren.
4acb	scheidung informationen	Benutzer sucht Informationen über Scheidung.	Relevant sind Seiten, die über Scheidung, Scheidungsrecht, Unterhalt usw. informieren, nicht relevant sind Seiten, die ausschließlich dazu dienen, auf kostenpflichtige Angebote, z.B. zur Rechtsberatung, zu verweisen und selbst keine Informationen bereitstellen.
5cab	robert swindells	Benutzer sucht Bücher des Schriftstellers Robert Swindells oder Informationen über den Autor.	Relevant sind Seiten die Informationen über den Autor beinhalten und Seiten, die Werke des Schriftstellers als Produkte zum Verkauf anbieten.
6cba	fehlgeburt	Benutzer sucht Informationen zum Thema Fehlgeburt.	Relevant sind Seiten die Informationen zu Ursachen, Diagnose, Therapie zum Thema Fehlgeburt beinhalten. Seiten, die das Thema Schwangerschaft behandeln aber keine expliziten Informationen zum Thema Fehlgeburt bereitstellen sind nicht relevant.
7abc	hotmail	Benutzer sucht die Homepage des Emailproviders "Hotmail" oder Informationen über diesen Emailprovider, beispielweise bezüglich der Qualität, Leistungsmerkmale.	Relevant ist die Homepage von "Hotmail" oder Seiten, die Informationen über diesen Mailprovider aufweisen.
8bac	christina aguilera songtexte	Benutzer sucht Songtexte der Sängerin "Christina Aguilera".	Relevant sind Seiten, die Songtexte der Sängerin beinhalten. Seiten die allgemeine Informationen oder Bilder

Figure 15: Queries, information needs and relevance criteria.

Scholar Search

Find other papers on this subject.

Google Search

How to cite this paper:

Griesbaum, J. (2004) "Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de" *Information Research*, **9**(4) paper 189 [Available at http://InformationR.net/ir/9-4/paper189.html]

Check for citations, using Google Scholar

13019
Web Counter

Last updated: 25 May, 2004

- Subject index |
- Search |
- Home