On the Precision of Search Engines: Results from a Controlled Experiment*

Hasan Girit, Robert Eberhard, Bernd Michelberger, and Bela Mutschler

University of Applied Sciences Ravensburg-Weingarten, Germany
{giritha,eberharr,michelbe,mutschlb}@hs-weingarten.de

Abstract. Handling the growing amount of digital information is one of the major challenges when dealing with the World Wide Web (WWW). In particular, users crave for an effective and efficient retrieval of needed information. In this context, search engines adopt a key role. Besides conventional search engines such as Google, semantic search engines have emerged as an alternative approach in recent years. The quality of search results delivered by search engines is influenced by many criteria. This paper picks up one specific issue, the precision, and investigates and compares the precision of current both conventional (i.e., non-semantic) and semantic search engines based on a controlled experiment with 77 participants. Specifically, Google, AltaVista, MetaGer, Hakia, Kngine, and WolframAlpha are investigated and compared.

Key words: conventional vs. semantic search engines, experiment

1 Introduction

When handling the growing amount of information in the WWW, search engines adopt a key role [1]. The simple use case from a user's perspective: to get an answer (i.e., information) for a specific question (i.e., a search query). However, asking questions (by means of a collection of keywords) and getting suitable answers (by means of relevant search results) remains a big challenge. The reason is that relevant information is indeed typically available, but it remains a complex task to accomplish to identify those information out of the huge amount of available information which are really helpful [2]. Thus, a lot of research is still performed to enable search engines to better answer the questions of their users.

Regarding this simple goal, two major approaches can be distinguished: First, conventional, non-semantic search engines index and rank web pages [3]. When a user enters a search query into a conventional search engine, the engine examines its index (cf. Section 2.1) and provides a list of best-matching web pages according to its internal ranking criteria (which are interpreted by a ranking algorithm). While some conventional search engines, such as Google, index only

* This paper was done in the niPRO research project. The project is funded by the German Federal Ministry of Education and Research (BMBF) under grant number 17102X10. More information can be found at http://www.nipro-project.org.

selected parts of web pages, others, such as AltaVista, index every single word of every web page [4]. Besides, additional metadata (e.g., author, title, keywords, description, date, language, format) about indexed web pages is used by many conventional search engines as well.

Second, semantic search engines seek to improve search accuracy by understanding user intent and the contextual meaning of terms appearing in the searchable data spaces, whether in the WWW or within closed systems, to generate relevant search results. Rather than using ranking algorithms (such as Google's PageRank) to predict relevancy, semantic search engines use semantics and the science of meaning in language to produce relevant search results. The goal is to deliver the information queried by a user rather than have a user navigate through a list of loosely related keyword results.

Now, which approach is better? This question is difficult to answer. Many issues determine the quality of search results delivered by search engines. This paper picks up one specific issue, the precision, and investigates and compares the precision of both conventional (i.e., non-semantic) and semantic search engines based on a controlled experiment with 77 participants. The investigated search engines include Google, AltaVista, MetaGer, Hakia, Kngine, and WolframAlpha (the reasons for having selected these engines are discussed in Section 3).

This paper is organized as follows. Section 2 provides important background information. Section 3 describes the research design underlying our empirical study. Section 4 presents the experiment results. Section 5 discusses related work. Section 6 concludes with a summary and an outlook.

2 Background Information

This section provides background information needed for the further understanding of the paper. Section 2.1 deals with the underlying concepts of both conventional and semantic search engines. Section 2.2 discusses the issue of precision, the key performance indicator we are investigating in our experiment.

2.1 Conventional vs. Semantic Search Engines

Conventional search engines gather, index, and rank information [5]. Specifically, these tasks are performed by a *crawler*, an *indexer*, and a *query engine*. Figure 1 illustrates how these three components are applied to process a query [6].

First, crawlers (also named *web spiders* or *robots*) autonomously collect available content, e.g., web pages. Think of a web browser which automatically follows every link on a web page. Doing so, the crawler captures as many web pages as possible. Each gathered web page is stored in a database and then indexed by the indexer [7]. When a user now enters a search query (i.e., a set of keywords) in the search engine's user interface (i.e., the search field), the (inverted) index is combined with a ranking algorithm to generate a list of potential matches, i.e., search results which probably provide relevant information (answers) with respect to the specified search query (question).

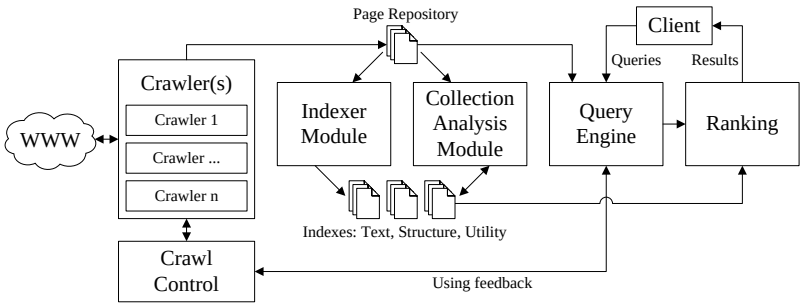

Fig. 1. Conventional search engines.

Semantic search engines, in turn, allow users to search not only based on a set of keywords. Natural language search phrases (e.g., when was Google founded?) are used. Moreover, semantic search engines typically allow to further refine the search space in order to increase the accuracy and relevance of search results [8, 9]. Generally, there exist three approaches of semantic search engines [10]: *context-based*, *evolutionary*, and *semantic association discovery search engines*. In our experiment we focus on context-based search engines as this approach is used by most existing semantic search engines. Figure 2 illustrates how a context-based semantic search engines generate its results [10].

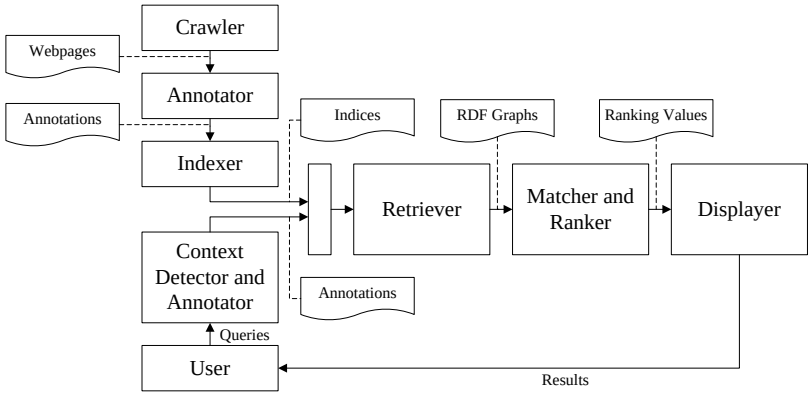

Fig. 2. Semantic search engines.

2.2 Evaluating Search Engines

As aforementioned, we investigate the precision of search engines in our experiment. This criterion is often used when evaluating a search engine's effectiveness [11]. The precision describes the ratio of relevant results with respect to the

total issued results of a search query, i.e., precision is a measure of the ability of a search engine to present only relevant results [12]. The most common way to analyze the precision of a search engine is to use a simple binary relevance judgment. A result for a search query is either relevant or not [11]. Let a be the number of *relevant results* and b be the number of *non-relevant results*. The precision of a search engine p can then be calculated as follows [13]:

$$p = \frac{a}{a+b} \tag{1}$$

Though, in order to calculate the precision of a search engine, it becomes obviously necessary to take a closer look at the notion of relevance [14, 15]. Crestani and Lalmas define relevance as logical relevance [16]: "A stored sentence is logical relevant to (a representation of) an information need if and only if it is a member of some minimal premiss set of stored sentences for some component statement of that need". We pick up this definition and additionally include two more variables into our definition of precision: Let c be the number of results containing *links to relevant content* (e.g., a search result is a web directory that is linking to other relevant web pages) and d be the number of *no results* (e.g., a search result links to a web page which is not reachable). This results in the following adapted equation (2), which we use in our experiment:

$$p = \frac{a + \frac{c}{2}}{a+b+c+d} \tag{2}$$

Note that a search result which contains links to relevant content is more valuable than a non-relevant result, i.e., it must be rated higher. For this purpose, links to relevant content are considered in the numerator of our precision equation. Relevant results (a) can be reached within one click (e.g., from the search result to the relevant content), whereas results containing links to relevant content (c) need at least two clicks (e.g., from the search result to the link to the relevant content). Therefore, a is still fully-weighted and c is half-weighted in the numerator of our precision equation. As a consequence, a search engine has a higher precision when providing links to relevant content instead of non-relevant results. The following section describes the research design underlying our empirical study.

3 Experiment Design

The objective of our experiment is to compare the precision of search engines. Therefore, each test person evaluates the relevance of search results for a given search query. Both conventional and semantic search engines are included in the experiment. Doing so, the following experiment variables have to be specified: *search engines*, *search queries*, *test persons*, and *data collection*.

Search Engines: First, the search engines to be investigated have to be selected. *Google* as the world's leading search engine has to be considered in any

case. Additionally, we selected *AltaVista* as it uses another search algorithm when compared to Google [1]. AltaVista uses the same search algorithm as Yahoo!. As a third conventional search engine we selected the meta-search engine *MetaGer*. This search engine is actually not a search engine on its own. MetaGer forwards entered search queries to various other search engines and then classifies and ranks the obtained search results [17].

Besides, we included the following semantic search engines in our experiment: *Hakia*, *Kngine*, and *WolframAlpha*. Hakia computes search queries both formulated in natural language and collections of keywords. Results are categorized, e.g., in web results, news, tweets or images [18]. Kngine is an abbreviation for "knowledge engine". Instead of indexing the web page, Kngine tries to interpret the content of web pages and organizes gained information in knowledge databases. It returns both organized, prepared information as well as conventional lists of web pages for a search query. Finally, WolframAlpha computes natural language search queries [19]. In a first step, WolframAlpha extracts relevant terms of a search query. In a second step, these terms serve as input for internal algorithms (note that almost no information is known on these algorithms). Finally, one (and only one) result is returned for a given search query.

Altogether, we investigate six search engines: two conventional non-semantic engines, one conventional meta-search engine, and three semantic search engines.

Search Queries: As explained, we want to investigate the precision of search results in our experiment. In order to compare search results from the six analyzed search engines, we use pre-defined search queries. To make sure that our experiment results are not biased by a too narrow or unfavorable selection of search queries, we use a wide range of topics and search queries. Moreover, we include both *semantic search queries* (which are formulated using natural language) and *non-semantic search queries* (which comprise a set of keywords). Specifically, we define 50 semantic and 50 non-semantic search queries. As the experiment took place in Germany, the search queries are formulated in German. Table 1 shows four exemplary search queries (translated into English).

Table 1. Sample of non-semantic and semantic search queries.

Non-semantic search queries	Semantic search queries
dollar rate	Who built the Statue of Liberty?
capital of canada	When was Wikipedia founded?

We then entered each of the 100 search queries into each of the analyzed search engines (cf. Fig. 3 - Step 1). The first ten search results delivered by each engine were copied into a separate text document (cf. Fig. 3 - Step 2); hence, six text documents belong to one search query whereas each document comprises the first ten search results of one engine. Doing so, we anonymized

and standardized the presentation of the search results in order to avoid that user ratings are biased by individual preferences, e.g., for certain search engines. During the experiment, the text documents are the basis for evaluating the search results.

Test Persons The overall number of participants in our experiment has to be high enough to ensure that our evaluation results are statistically sound. Literature suggests that at least 50 test persons should participate [20].

Data Collection: We used a web-based questionnaire to collect data from the test persons. In order to have different analysis options of the collected data, every test person had to denote some personal data: gender, age, educational background, working position, frequency of internet usage, and frequency of search engine usage. During the experiment, the test persons had to rate each search result as (1) *relevant*, (2) *not relevant*, (3) *links to relevant content*, or (4) *no result*. A *relevant* search result contains useful information for the test person with respect to the search query. Selecting *not relevant* means that the search result has no relation to the search query. If the search result itself does not contain useful information, but links to further relevant information instead, the option *links to relevant content* can be used. In order to handle WolframAlpha (remember that it does only deliver one search result with no outgoing links) the statement *no result* is also added as a possible rating.

When performing the experiment, each test person received an e-mail containing a short description explaining the experiment and its goals. The e-mail includes the prepared text documents (results) (cf. Fig. 3 - Step 3) and a link to the web-based questionnaire (cf. Fig. 3 - Step 4).

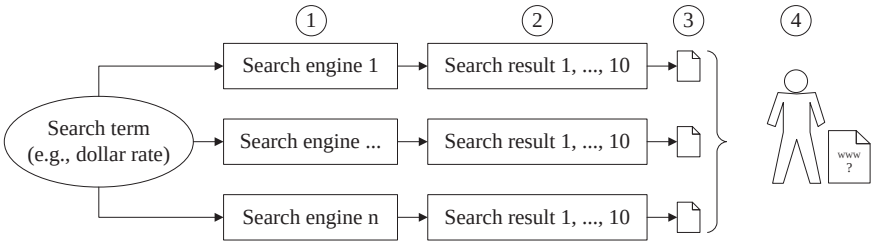

Fig. 3. Performing the experiment - Step 1 - 4.

4 Experiment Results

In our experiment, 77 people participated. Most participants (58%) were between 16 and 25 years old. Another 34% were between 26 and 35 years old and 4%

were between 36 and 45 years old. Only 1% of the participants was between 46 and 55 years old. The rest of the participants (3%) were older than 55.

We also asked for the frequency of internet usage. The majority (61%) told us that they use the internet more than 3 hours a day. Another 35% use it up to 3 hours a day. Only a minority of 4% is online less than 5 hours in a week.

We also wanted to know, how often the participants use search engines. The majority of participants (62%) use search engines more than 3 times a day. Another 25% use respective engines up to 3 times a day. Only 12% use search engines only up to 5 times in a week. The rest of the participants (1%) use search engines less than once in a week.

Moreover, in our experiment the participants evaluated 770 semantic and 770 non-semantic search results for each search engine except for WolframAlpha with only 77 search results. To ensure comparability with the other search engines the results of WolframAlpha were extrapolated (cf. Fig. 4). Table 2 and Table 3 show the raw data collected during the experiment.

Table 2. Raw data "non-semantic".

Search engine	**relevant**	**links to**	**not relevant**	**no result**
Google	362	142	237	29
Hakia	196	112	367	95
AltaVista	228	106	343	93
Kngine	281	118	327	44
MetaGer	192	99	426	53
WolframAlpha	30	0	32	15

Table 3. Raw data "semantic".

Search engine	**relevant**	**links to**	**not relevant**	**no result**
Google	372	110	246	42
Hakia	162	73	410	125
AltaVista	211	70	443	46
Kngine	245	85	385	55
MetaGer	171	84	441	74
WolframAlpha	28	0	30	19

Figure 4 compares the total number of identified relevant results, links to relevant content, non-relevant results and no results for both non-semantic (cf.

Fig. 4A) and semantic (cf. Fig. 4B) search queries. Figure 4 shows that Google delivers the best results for both semantic and non-semantic search queries.

Figure 5 additionally shows that relevant results differ between semantic and non-semantic search terms. All investigated search engines except for Google provide a smaller number of relevant results for semantic search queries than for non-semantic ones (cf. Fig. 5A).

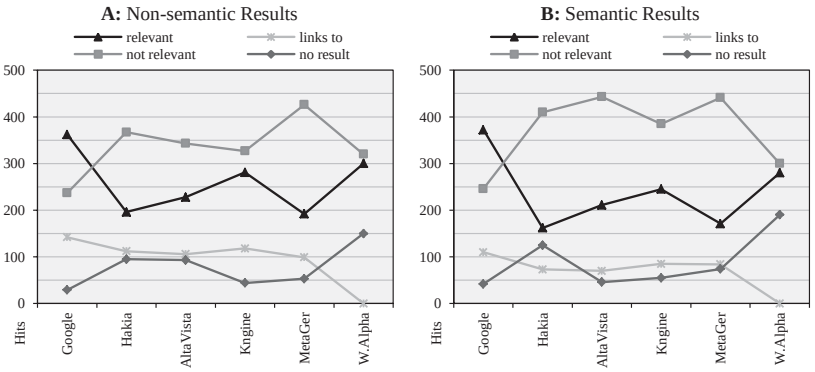

Fig. 4. Comparison I - total number of search results.

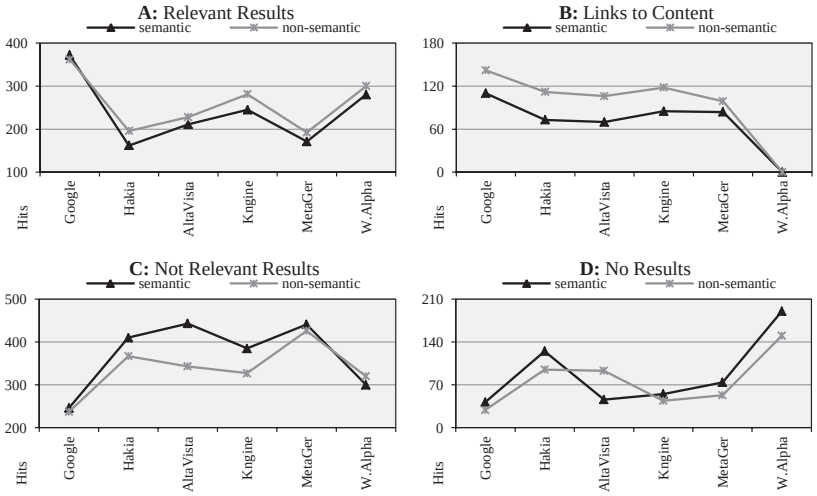

Fig. 5. Comparison II - total number of search results.

In order to determine the quality of the investigated search engines in detail, we use equation (2) from Section 2.2. The following results were obtained for

the precision (cf. Table 4): Google has a precision of $p = 0.555$ for semantic search queries and a precision of $p = 0.562$ for non-semantic search queries. AltaVista has a precision of $p = 0.319$ for semantic search queries and a precision of $p = 0.365$ for non-semantic search queries. Kngine, as the best semantic search engine, delivers better results compared to AltaVista with a precision of $p = 0.373$ for semantic search queries and a precision of $p = 0.442$ for non-semantic search queries. Table 4 shows the results for all search engines in detail.

Table 4. Comparison: Precision "non-semantic" and "semantic".

Search engine	***p* "non-semantic"**	***p* "semantic"**	***p* "total"**
Google	0.562	0.555	0.559
Hakia	0.327	0.260	0.294
AltaVista	0.365	0.319	0.342
Kngine	0.442	0.373	0.408
MetaGer	0.314	0.280	0.297
WolframAlpha	0.390	0.364	0.377

Altogether, all search engines are delivering less relevant search results for semantic search queries - even the semantic search engines. A first reason might be that semantic search queries contain some unnecessary copulas; not all words which must be "understood" to deliver a relevant search result might be identified. A second reason might be that the search engines had problems in handling the German language and the correct interpretation of the (German) search queries. Especially the semantic search engines had significant problems in this respect. These difficulties might lead to search results the user does not consider as relevant. Interestingly, WolframAlpha had the biggest problems.

In summary, best results (cf. Fig. 6) are achieved by Google with an overall precision p of 0.559 followed by Kngine with a precision p of 0.408. Third best is AltaVista ($p = 0.342$) followed by MetaGer ($p = 0.297$) and Hakia ($p = 0.294$). WolframAlpha which has to be separately evaluated, has a precision p of 0.377.

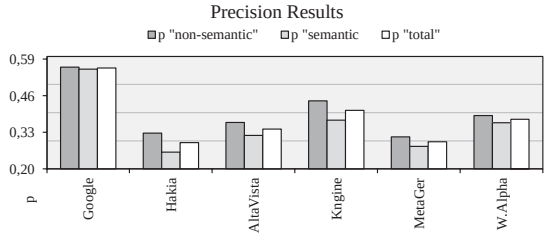

Fig. 6. Comparison III - precsion of search engines.

5 Discussion

Our experiment results show very good results for Google when compared to all other search engines. The results indicate that Google is currently indeed the top of the notch search engine. Semantic search engines generally showed relatively poor results. Reason might be, as aforementioned, linguistic deficits in the translation of our queries (which were formulated in German). However, note that all search engines are classified as multilingual, i.e., handling different languages actually should not be a problem.

To better understand the needs of search engine users, we asked the test persons (using the web-based questionnaire) to give feedback on the three most important characteristics of a search engine. Figure 7 shows that the quality and the actuality of results as well as the speed of a search engine regarding the processing of search queries are considered as particularly important.

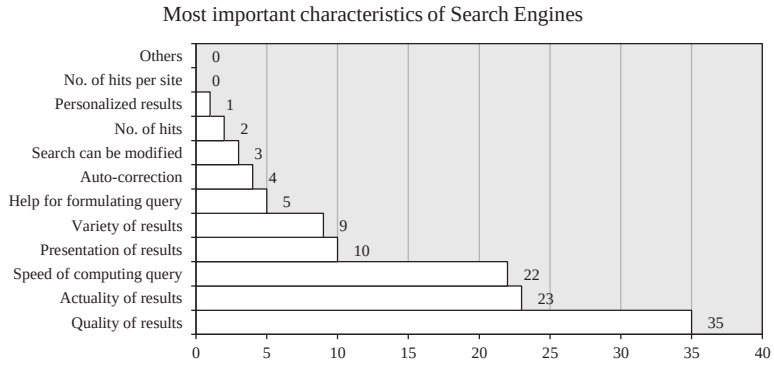

Fig. 7. Most important search engine characteristics.

6 Related Work

There are other studies dealing with the comparison of conventional and semantic search engines. Empirical and interdisciplinary studies combining semantic web and conventional information retrieval approaches are provided by several authors. Hendler [21], for example, investigates the capabilities of semantic technologies towards their ability to increase the value of content (or search results) through the linking of content. Our experiment, by contrast, only considers retrieved search results (and therewith the precision). The work by Xu [22] provides interesting insights on the impact of collaborative filtering (based on Web 2.0 approaches) on the quality of search results.

Further relevant studies stem from Finin [23] and Ding [24, 25]. The relevance of search results of conventional search engines is also investigated by Brin [26],

Page [27], and Silverstein [4]. Silverstein [4], for example, analyzed the AltaVista search engine query log comprising approximately one billion entries for search requests over a period of six weeks.

7 Summary and Outlook

This paper investigates and compares the precision of both conventional and semantic search engines based on a controlled experiment with 77 participants. The six search engines Google, AltaVista, MetaGer, Hakia, Kngine, and WolframAlpha are investigated. Best results are achieved by Google with an overall precision of $p = 0.559$ followed by Kngine with an overall precision of $p = 0.408$. Third best search engine is AltaVista ($p = 0.342$) followed by MetaGer ($p = 0.297$) and Hakia ($p = 0.294$). WolframAlpha (which is evaluated differently) shows an overall precision of $p = 0.377$. In summary, semantic search engines (e.g., Kngine) do not yet achieve the same relevance ratings as conventional search engines (e.g., Google).

Future work will include further controlled experiments in order to evaluate other criteria determining the performance of both conventional and semantic search engines (e.g., further investigation on the recall will be done). Additional research will be also done in the context of enterprise search. Enterprise search engines will be investigated and compared regarding their performance.

References

1. Levene, M.: An Introduction to Search Engines and Web Navigation - Second Edition. John Wiley & Sons, Inc., Hoboken, New Jersey (2010)
2. Cambazoglu, B.B., Beaze-Yates, R.: Scalability Challenges in Web Search Engines. in: Book of Advanced Topics in Information Retrieval, Springer Verlag, pp. 27-49 (2011)
3. Gordon, M., Pathak, P.: Finding Information on the World Wide Web: the Retrieval Effectiveness of Search Engines. in: J. Information Processing and Management, 35(2), pp. 141-180 (1999)
4. Silverstein, C., Heinzinger, M., Maires, H., Moricz, M.: Analysis of a Very Large Web Search Engine Query Log. in: J. ACM SIGIR FORUM, 33(1), pp. 6-12 (1999)
5. Risvik, K.M., Michelsen, R.: Search Engines and Web Dynamics. in: J. of Computer Networks, 39(3), pp. 289-302 (2002)
6. Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S.: Searching the Web. in: J. of ACM Transactions on Internet Technology (TOIT), 1(1), pp. 2-43 (2001)
7. Langville, A.N., Meyer, C.D.: Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton Univers. Press, Princeton and Oxford (2006)
8. Mika, P.: Ontologies are us: A Unified Model of Social Networks and Semantics. in: J. of Web Semantics: Science, Services and Agents on the World Wide Web, 5(1), pp. 5-15 (2007)
9. Mayfield, J., Finin, T.: Information Retrieval on the Semantic Web: Integrating Inference and Retrieval. in: Proc. of the Int'l Workshop on the Semantic Web at the 26th Int'l ACM SIGIR Conf.e on Research and Development in Information Retrieval, Toronto, Canada (2003)

10. Esmaili, K.S., Abolhassani, H.: A Categorization Scheme for Semantic Web Search Engines. in: Proc. of the 2006 IEEE Int'l Conf. of Computer Systems & Applications, pp. 171-178 (2006)
11. Vaughan, L.: New Measurements for Search Engine Evaluation Proposed and Tested. in: J. Information Processing and Management, 40(4), pp. 677-691 (2004)
12. Lewandowski, D.: Web Information Retrieval: Technologien zur Informationssuche im Internet. Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI), Frankfurt am Main (2005)
13. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. in: J. of Information Process Management, 28(4), pp. 467-490 (1992)
14. Mizzaro, S.: How Many Relevances in Information Retrieval. in: J. Interacting with Computers, 10(3), pp. 303-320 (1998)
15. Cooper, W.S.: A Definition of Relevance for Information Retrieval. in: J. Information Storage and Retrieval, 7(1), pp. 19-37 (1971)
16. Crestani, F., Lalmas, M.: Logic and Uncertainty in Information Retrieval. in: Lect. on Information Retrieval, Springer Verlag, pp. 179-206 (2001)
17. Yang, X., Zhang, M.: Necessary Constraints for Fusion Algorithms in Meta Search Engine Systems. in: Proc. of the Int'l Conf. on Intelligent Technologies, pp. 409-416 (2000)
18. Campesato, O., Nilson, K.: Web 2.0 Fundamentals with AJAX, Development Tools, and Mobile Platforms. Jones and Barlett Publishers LLC (2011)
19. Weikum, G.: Search for Knowledge. in: Proc. SeCO Workshop on Search Computing Challenges and Directions, pp. 24-39 (2009)
20. Buckley, C., Voorhees, E.M.: Evaluating Evaluation Measure Stability. in: SIGIR '00 Proc. of the 23rd annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 33-40 (2000)
21. Hendler, J., Golbeck, J.: Metcalfe's Law, Web 2.0, and the Semantic Web. in: J. of Web Semantics: Science, Services and Agents on the World Wide Web, 6(1), pp. 14-20 (2008)
22. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the Semantic Web: Collaborative Tag Suggestions. in: Proc. of the Collaborative Web Tagging Workshop at the WWW 2006 (2006)
23. Finin, T., Ding, L.: Search Engines for Semantic Knowledge. in: Proc. of XTech 2006: Building Web 2.0 (2006)
24. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Pen, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. in: Proc. of the 13th Int'l Confernce on Information and Knowledge (CIKM'04), pp. 652-659 (2004)
25. Ding, L., Finin, T., Joshi, A., Peng, Y., Pan, R., Reddivari, P.: Search on the Semantic Web. in: J. Computer, IEEE Computer Society, 38, pp. 62-69 (2005)
26. Brin, S., Page, L.: The Anatomy of a Large Scale Hypertextual Web Search Engine. in: J. Computer Networks ISDN Syst., 30(1-7), pp. 107-117 (1998)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. in: Stanford InfoLab, Technical Report (1999)