Information Retrieval Effectiveness of Turkish Search Engines

Yıltan Bitirim[1], Yaşar Tonta[2], and Hayri Sever[3]

[1] Department of Computer Engineering
Eastern Mediterranean University
Famagusta, T.R.N.C. (via Mersin 10 Turkey)
yiltan.bitirim@emu.edu.tr
[2] Department of Library Science
Hacettepe University
06532 Beytepe, Ankara, Turkey
tonta@hacettepe.edu.tr
[3] Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
sever@cs.umass.edu

Abstract. This is an investigation of information retrieval performance of Turkish search engines with respect to precision, normalized recall, coverage and novelty ratios. We defined seventeen query topics for Arabul, Arama, Netbul and Superonline. These queries were carefully selected to assess the capability of a search engine for handling broad or narrow topic subjects, exclusion of particular information, identifying and indexing Turkish characters, retrieval of hub/authoritative pages, stemming of Turkish words, correct interpretation of Boolean operators. We classified each document in a retrieval output as being "relevant" or "nonrelevant" to calculate precision and normalized recall ratios at various cut-off points for each pair of query topic and search engine. We found the coverage and novelty ratios for each search engine. We also tested how search engines handle meta-tags and dead links. Arama appears to be the best Turkish search engine in terms of average precision and normalized recall ratios, and the coverage of Turkish sites. Turkish characters (and stemming as well) still cause bottlenecks for Turkish search engines. Superonline and Netbul make use of the indexing information in metatag fields to improve retrieval results.

1 Introduction

With respect to the statistics on web usage we have compiled using different Internet resources and articles [4], the number of Internet users, host sites, and documents world-wide are at least 419 millions, 120 millions, and one billion, respectively. Furthermore, it is estimated that the number of hosts and web documents are doubled in

[4] See the sites at (1) http://www.nua.com/surveys, (2) http://www.netsizer.com/index.html and (3) http://www.inktomi.com/new/press/2000/billion.html for web usage statistics.

every year [1]. These facts alone make the need for search engines to retrieve the information on web apparent, along with other types of search systems and web mining agents.

There are numerous works in literature on different perspectives of search engines, such as their functional and architectural views [2], design issues[3–5] and performance evaluation [6–10]. It is a delusive conception to regard information retrieval systems and search engines the same, just because they both strive to satisfy information needs of users. Noticeable differences for search engines emerge from raw facts such as dynamic nature (40 in every month[1]), poor information value (or spare data) [11], poor authoring quality[1, 12], duplication, meta representation of web pages. These have given rise to build up new components (or software agents in general), namely crawling and spying components (link checker, page change monitor, validator, and new page detector), and retrieval strategies (utilization of multiple evidence on relevance of web pages [4, 13]). For the above reasons, the development of search engines have their own challenges even if they continue to enjoy many solid research results of information retrieval field. A formal approach to the evaluation of Turkish search engines is, hence, an apparent need and of course valuable work for flourishing Turkish search engines.

In this paper, we discuss the method and configuration of the experiment, and summarize main results of an extensive six-month work on the four search engines, namely, Arabul, Arama, Netbul and Superonline.

2 Method

We aimed to answer several research questions with regard to the effectiveness of popular Turkish search engines in handling broad as well as narrow information needs of users, in excluding some particular information and in satisfying one- or two-word query expressions. In addition, we looked at qualitative issues such as incorporation of stemming facility for both indexing document and search terms, correct implementation of Boolean operators, incorporation of sound-like operator to properly handle accented Turkish characters, update (visit) period of crawlers for indexed (existing) links. To assess the coverage and novelty aspects of Arama, Arabul, Netbul, and Superonline engines, we picked up some terms from the top 10 list of the most frequently used search terms.

We identified the following seventeen query topics [5]:

(1) Internet and ethics,
(2) baroque music and its characteristics,
(3) information about the medicine ”Prozac” (but not about the rock group ”Prozac”)
(4) related works about the evaluation of Turkish search engines on the Internet,
(5) mp3 copies of songs of ”Barış Manço” (note the spelling)
(6) mp3 copies of songs of ”Baris Manco”,
(7) what is DPT (State Planning Organization)?

[5] We refer the reader to the site at cmpe.emu.edu.tr/bitirim/engine for further information on our work, e.g., the original Turkish query topic statements or their query expressions for search engines.

(8) general information about ”alien”,
(9) general information about ”aliens” (note the plural form),
(10) documents mentioning former president of the Republic of Turkey, ”Süleyman Demirel” AND the current president ”Ahmet Necdet Sezer,”
(11) documents mentioning former president of the Republic of Turkey, ”Süleyman Demirel” OR the current president ”Ahmet Necdet Sezer,”
(12) approaches of presidents of the Republic of Turkey ”Süleyman Demirel” OR ”Ahmet Necdet Sezer” to TEMA,[6]
(13) information about space,
(14) information about universe,
(15) information about space OR universe,[7]
(16) the relation between ”Mustafa Kemal Atatürk” and ”Fikriye Hanım,”
(17) information about the Speaker of the Turkish Parliament ”Ömer İzgi”.

It may be worth stating that above 17 query topics were converted to structured formal queries to simplify the parsing process for each search engine. Needless to say, different search engines have different syntactic rules to parse and process search queries. For instance, each search engine treats Boolean operators *(AND, OR, NOT)* in a somewhat different way.

2.1 Evaluation of Relevance

The relevance judgment was binary, though normalized recall metric allows us to pose more relevancy levels. Documents with the same content but different web addresses (i.e., mirror pages) were considered as different ones. Duplicated information items was conceived as good as one for its relevance judgment. If the retrieved document was not accessible for some reason or in different languages other than Turkish or English, then it was considered as non-relevant information item.

2.2 Performance Measurement

The effectiveness measurements in information retrieval are typically of precision and recall, which can be defined for a user query as the proportion of retrieved and relevant documents over retrieval output and relevant documents, respectively. It has been a common practice in the evaluation of search engines to exclude recall values for obvious reason, though there were some recommendations in the literature for estimating average recall value [14] by pooling method [15]. We used, however, precision at different sizes of retrieval output, i.e., cut-off points. The precision at different cut-offs can be used to roughly see how scores of relevant documents are distributed over their ranks. Note that, in our evaluation, when the number of documents retrieved is smaller than the cut-off point at the hand, precision was calculated over total documents retrieved.

This score-rank curve is closely related to the normalized recall, say R_{norm} [16]. The normalized recall is based on the optimization of expected search length [17]. On

[6] The Turkish Foundation for Combating Soil Erosion, for reforestation and the protection of natural habitats.

[7] ”space” and ”universe” are synonymous words.

other words, It utilizes the viewpoint that a retrieval output Δ_1 is better than another one Δ_2 if the user gets fever non-relevant documents with Δ_1 than with Δ_2. We calculated normalized recall at four cut-off points for each query per search engine in order to parallel with precision values. The R_{norm} is defined as

$$R_{norm}(\Delta) = \frac{1}{2}\left(1 + \frac{R^+ - R^-}{R^+_{max}}\right)$$

where R^+ is the number of document pairs where a relevant document is ranked higher than non-relevant document, R^- is the number of document pairs where a non-relevant document is ranked higher than relevant one, and R^+_{max} is the maximal number of R^+.

We strongly believe that the effectiveness figures of search engines should be accompanied by their coverage, novelty, and recency ratios to obtain a complete picture. The coverage ratio is the proportion of the number of documents (which are previously known as relevant to the user) retrieved to the total number of documents known as relevant. The novelty ratio is the proportion of the relevant documents (which are not previously known to the user) retrieved to the relevant documents retrieved [18, 19]. The recency factor is simply the percentage of dead links.

The coverage and novelty ratios of Turkish search engines were measured using the most frequently searched five one-word search queries on Arabul search engine, namely, ”mp3”, ”oyun” (game), ”sex”, ”erotik” (erotic), and ”porno” (porn). These top search terms appear to be consistent with those of globally-known search engines [20, 21]. We used a pool of 1000 documents for each of above search terms. All these three measures, i.e., coverage, novelty, and recency, were normalized across the corresponding values of search engines.

Finally, we tested if Turkish search engines make use of metadata to retrieve documents. For each search engine, we selected a web document containing meta tags ”keyword” and ”description”. Then we used the terms that appeared in the document’s metatags and performed a search on each search engine to determine if metatags are used for retrieval purposes.

2.3 Analysis of Data

We analyzed the precision and normalized recall ratios for each search engine to determine if they significantly differ in terms of retrieval effectiveness. As the distribution of precision and normalized recall ratios were not normal, we used nonparametric Kruskal-Wallis and Mann-Whitney statistical tests that are used for ordinal data. Kruskal-Wallis (H) was used to test if precision and normalized recall ratios of each search engine in different cut-off points were different from others and if the difference was statistically significant. If different, then pair-wise Mann-Whitney tests were applied to determine which search engines engendered it. The same statistical tests were also applied to identify if there was any difference among search engines in terms of their recency values. Pearson’s (r) was used to test if there was any relationship between precision and normalized recall ratios.

3 Experiment Results

Findings of our experiment and the analysis of results are discussed below.

3.1 The Number of Documents Retrieved by Search Engines

The number of zero retrievals (i.e., no documents retrieved) or retrievals that contain no relevant documents (i.e., the precision ratio is zero) can be used to evaluate the retrieval performance of search engines. In our experiment, Arabul could not retrieve any document for 6 of 17 queries. The same figure was 1 out of 17 for Arama. Although Netbul and Superonline retrieved at least one document for each query, Netbul could not retrieve any relevant documents for 8 of 17 queries. Superonline, Arabul, and Arama could not retrieve any relevant documents for 5, 5, and 3 queries out of 17, respectively. If we examine both zero retrievals and retrievals with no relevant documents, Arabul could not retrieve any relevant documents for 11 out of 17 queries (65%). (Netbul: 8 (47%); Superonline: 5 (29%); and Arama; 4 (24%.) The number of relevant documents retrieved for each query on each search engine is given in Table 1. The first number in the row labelled ”Total” shows the total number of relevant documents retrieved and the second one (in parentheses) shows the total number of documents retrieved by each search engine.

Query	Arabul	Arama	Netbul	Superonline
1	0	3	0	0
2	1	2	0	2
3	0	3	4	10
4	0	0	0	0
5	0	3	0	0
6	0	4	3	1
7	2	12	1	1
8	6	4	3	4
9	9	4	2	3
10	0	5	2	8
11	0	6	4	10
12	0	0	0	0
13	5	3	2	2
14	0	0	0	1
15	2	8	5	6
16	0	7	0	6
17	0	0	0	0
Total	25 (119)	64 (277)	26 (273)	54 (302)
Average	1.5	3.8	1.5	3.2

Table 1: Number of relevant documents retrieved

As Table 1 shows, Arama retrieved the highest number of relevant documents for 17 queries (64). Arama’s mean number of relevant documents per search query (3.8) was

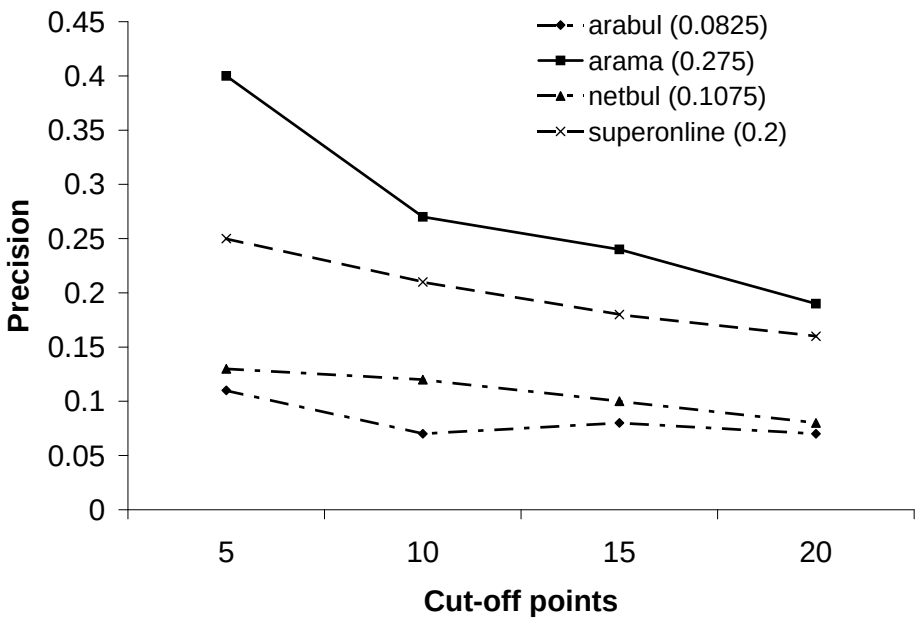

Fig. 1: Mean precision ratios

also higher than those of other search engines. Total number of documents retrieved by all search engines was 971, of which 169 were relevant. To put it differently, about 5 in 6 documents retrieved by search engines were not relevant.

3.2 Precision Ratios

Mean precision values of search engines in various cut-off points (for first 5, 10, 15, and 20 documents retrieved) are shown in Figure 1. Arama has the highest precision ratios on all cut-off points (mean 28%). Then comes Superonline (20%), Arabul (15%), and Netbul (11%).

Tests showed that there was no statistically meaningful difference between precision values of search engines while cut-off points were 10, 15, and 20. Yet, a statistically meaningful difference was observed when the cut-off point was 5: Arama retrieved more relevant documents per search query than those of Arabul and Netbul. Otherwise, search engines scored similar precision ratios in higher cut-off points.

3.3 Normalized Recall Ratios

As we pointed out earlier, the normalized recall ratio measures if search engines display relevant documents in the top ranks of the retrieval outputs. If a search engine could not retrieve any documents for a search query, the normalized recall value for that query will be zero. Mean normalized recall values of search engines in various cut-off points (for first 5, 10, 15, and 20 documents retrieved) are shown in Figure 2.

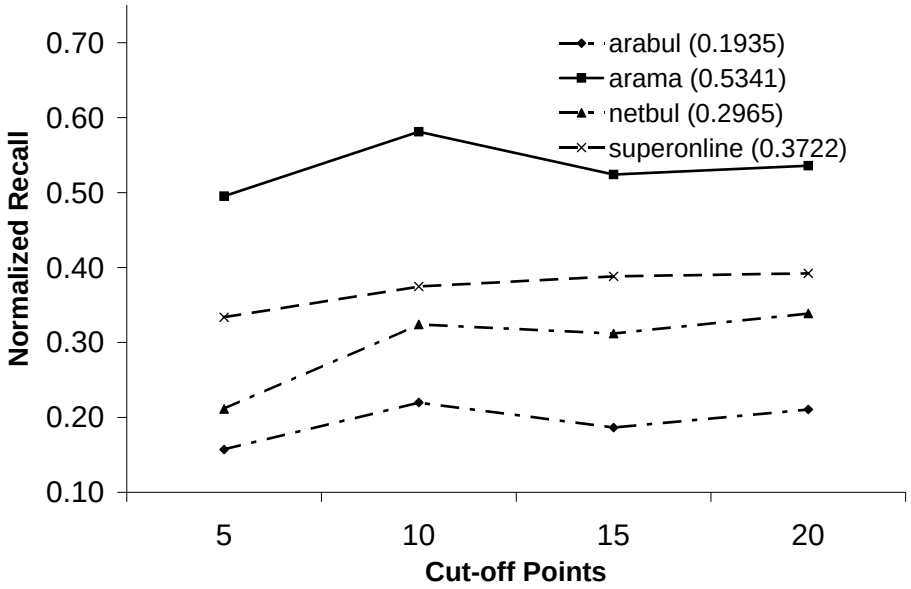

Fig. 2: Mean normalized recall ratios

As Figure 2 shows, Arama has the highest (mean 54%) normalized recall values in all cut-off points. The mean normalized recall value of Arabul is the lowest as it could not retrieve any documents for 6 of 17 queries.

Tests showed that there was no statistically meaningful difference between normalized recall values of search engines while the cut-off point was 20. In other words, none of the search engines displayed relevant documents in distinguishably higher ranks than others in general. Yet, Arama scored better than Arabul (in cut-off points 5, 10, and 15) and Netbul (in cut-off point 10), and the differences were statistically significant.

3.4 The Relation between Precision and Normalized Recall Ratios

The relationship between mean precision and mean normalized recall ratios was statistically significant. In other words, when the mean precision value was high, the mean normalized recall value was high, although the relationship got weakened as the cut-off point had increased (cut-off(5)$\rightarrow$ Pearson's $r = .97$, cut-off(10)$\rightarrow$ $r = .89$, cut-off(15)$\rightarrow$ r=.70, cut-off(20)$\rightarrow$ r=.61). It appears that the number of relevant documents in the retrieval output tends to decrease as one goes down the list.

3.5 Turkish Character Usage in Search Engines

We used the fifth and sixth queries to examine how Turkish search engines respond to search queries that contain Turkish characters in them. Both queries were on mp3 copies of songs of the late Turkish pop singer, Barış Manço. His name was spelled in

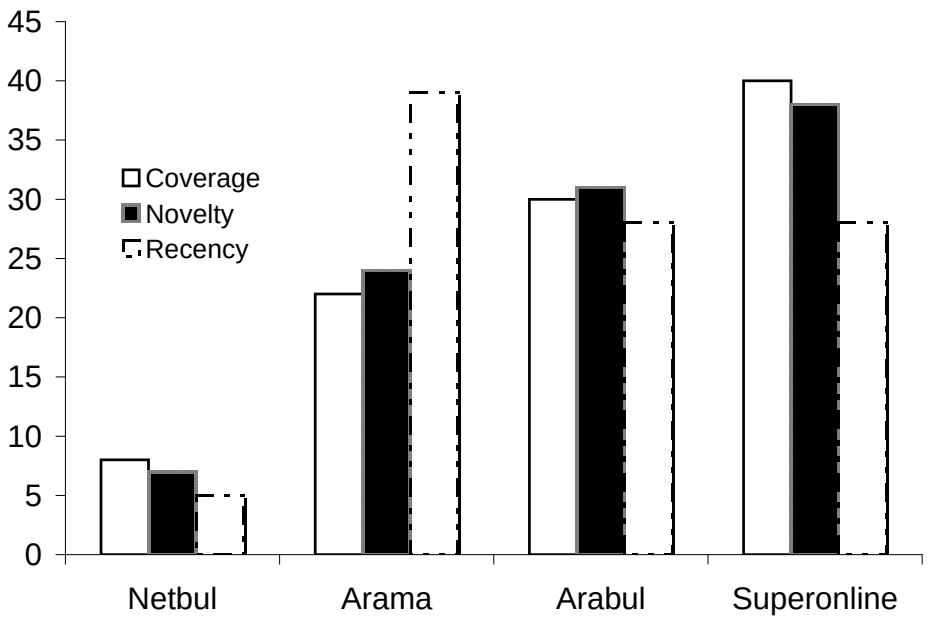

Fig. 3: Macro average of normalized coverage and novelty measures of search engines for a pool of 1000 documents for each of top five search terms. Normalized recency was calculated over seventeen queries.

two different forms: one with proper Turkish characters ”ı”, ”ş” and ”ç”, the other with the accented versions of the same characters i, s and c. As shown in Table 1, Arama, Netbul, and Superonline handled fifth and sixth query expressions in different ways. In case of Arabul, it differed on corresponding retrieval sets, which is apparently an indication of discriminative responses of the system against these two queries.

3.6 Coverage, Novelty and Recency Ratios

Four search engines retrieved a total of 9944 unique relevant documents for all five queries (”mp3”, ”oyun” (game), ”sex”, ”erotik” (erotic), and ”porno” (porn)), which were collected in the pool. Superonline retrieved about 40% of all unique documents (Arabul 31%, Arama 24% and Netbul 7%). As shown in Figure 3, Superonline’s coverage ratio was much higher than that of other search engines for the most frequently searched queries on the Turkish search engines [8]. Furthermore, it was evident that Netbul had poor coverage of web domain. We see the same trend for novelty ratios of search engines with slightly changed values.

As shown in the same figure, Arama has the largest ratio (39%) of all broken links. Search engines Arabul and Superonline follow Arama with 28% each. Netbul has the

[8] Arama is the indisputable leader in covering documents with Turkish addresses, though we did not publish the results in regard to .tr domain because of space limitation.

lowest ratio 5%. The average number of broken links per query for search engines were as follows: Arama: 5.1, Superonline: 2.8, Arabul: 1.4, and Netbul: 0.7. In general, one in six (17%) documents retrieved by search engines contained broken links.

3.7 Metadata Usage

Web documents contain indexing information in their metatag fields (i.e., metatags for ”author”, ”description”, ”keywords”, and so on). Some search engines make use of indexing information that appears in metatag fields of documents to increase the likelihood of retrieving more relevant documents. To test the use of metadata for retrieval purposes by Turkish search engines, we first identified certain documents with metatag ”keywords” field filled. We further checked if these documents were already indexed by each search engine. We found one document for each search engine satisfying both conditions. We then used the key words that appeared in metatag field ”keywords” as search queries and tested if each search engine was able to find the document in question. It appears that Arama and Arabul have not benefited from metadata for retrieval purposes whereas Netbul and Superonline took advantage of the indexing information contained in metatags to upgrade the retrieval status of documents.

4 Conclusions

Major findings of our research can be summarized as follows: On the average, one in six documents retrieved by Turkish search engines was not available due to dead or broken links. Netbul retrieved fewer documents with dead or broken links than other search engines did. Some search engines retrieved no documents (so called ”zero retrievals”) or no relevant documents for some queries. On the average, five in six documents retrieved by search engines were not relevant. Average precision ratios of search engines ranged between 11% (Netbul) and 28% (Arama) (Superonline being 20% and Arabul 15%). Arama retrieved more relevant documents than that of Arabul and Netbul in the first five documents retrieved. Search engines do not seem to make every efforts to retrieve and display the relevant documents in higher ranks of retrieval results. Average normalized recall ratios of search engines ranged between 20% (Arabul) and 54% (Arama) (Superonline being 37% and Netbul 30%). Arama retrieved the relevant documents in higher ranks than that of Arabul and Netbul. The strong positive correlation between the precision and normalized recall ratios got weakened as the cut-off value increased. Search engines were less successful in finding relevant documents for specific queries or queries that contained broad terms. Although nonrelevant documents were higher in number, search engines were more successful in single-term queries or queries with Boolean ”OR” operator. The use of Turkish characters such as ”ç”, ”ı”, and ”ş” in queries still creates problems for Turkish search engines as retrieval results differed for such queries. For retrieval purposes, Netbul and Superonline seem to index and make use of metadata fields that are contained in HTML documents under ”keywords” and ”description” meta tags. We did not encounter any anomaly case for implementation of Boolean operators.

References

1. M. Kobayashi and K. Takeda. Information retrieval on the web. *ACM Computing Surveys*, 32(2):144–172, June 2000.
2. J. Jansen. Using an intelligent agent to enhance search engine performance. *First Monday*, 1996. http://www.firstmonday.dk/issues/issue2_3/jansen/index.html.
3. W. Mettrop and P. Nieuwenhuysen. Internet search engines: Fluctuations in document accessibility. *Journal of Documentation*, 57:623–651, 2001.
4. W.B. Croft and H. Turtle. A retrieval model for incorporating hypertext links. In *Proceedings of ACM Hypertext Conference*, pages 213–224, New Orleans, LA, November 1989.
5. V.N. Gudivada, V.V. Raghavan, W.I. Grosky, and R. Kasanagottu. Information retrieval on the world wide web. *IEEE Internet Computing*, 1(5):58–68, 1997.
6. H. Chu and M. Rosenthal. Search engines for the world wide web: A comparative study and evaluation methodology. In Steve Hardin, editor, *Proceedings of the 59th ASIS Annual Meeting*, pages 127–135, Baltimore, Maryland, October 1996.
7. H.V. Lerghton and J.V. Srivastava. First 20 precision among WWW search services. *Journal of the American Society for Information Science*, 50:870–881, 1999.
8. C. Oppenheim, A. Morris, and C. McKnight. The evaluation of WWW search engines. *Journal of Documentation*, 56:190–211, 2000.
9. J. Savoy and J. Picard. Retrieval effectiveness on the web. *Information Processing and Management*, 37:543–569, 2001.
10. M. Gordon and P. Pathak. Finding information on the WWW: The retrieval effectiveness of search engines. *Information Processing and Management*, 35:141–180, 1999.
11. J. S. Deogun, H. Sever, and V. V. Raghavan. Structural abstractions of hypertext documents for web-based retrieval. In Roland R. Wagner, editor, *Proceedings of Ninth International Workshop on Database and Expert Systems Applications, (in conjunction with DEXA'98)*, pages 385–390, Vienna, Austria, August 1998.
12. M.E. Küçük, B. Olgun, and H. Sever. Application of metadata concepts to discovery of internet resources. In Tatyana Yakhno, editor, *Advances in Information Systems (ADVIS'00)*, volume 1909, pages 304–313. Springer Verlag, Berlin, GR, October 2000.
13. J.M. Kleinberg. Authoritative source in a hyperlinked environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, 1998.
14. S. C. Clarke and P. Willet. Estimating the recall performance of web search engines. *Aslib Proceedings*, 49(7):184–189, July/August 1997.
15. D. Hawking, N. Craswell, P. Thislewaite, and D. Harman. Results and challenges in web search evaluation. In D. Harman, editor, *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, November 1999.
16. Y.Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46:133–145, 1995.
17. W.B. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19:30–41, 1968.
18. S. Lawrence and C.L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 3 April 1998. http://www.neci.nec.com/ lawrence/science98.html.
19. R.R. Korfhage. *Information Storage and Retireval*. Wiley, New York, NY, 1997.
20. B. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.
21. C. Silverstein, M. Henziger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.