TREC-10 Web Track Experiments at MSRCN

**Jianfeng Gao*, Guihong Cao#, Hongzhao He#, Min Zhang##,
Jian-Yun Nie**, Stephen Walker*, Stephen Robertson***

* Microsoft Research, {jfgao,sw,ser}@microsoft.com
** Département d'informatique et de recherche opérationnelle, Université de Montréal,
nie@iro.umontreal.ca
#Department of Computer Science and Engineering of Tianjin University, China
State Key Lab. of Intelligent Tech. & Sys., Computer Science. &Tech. Dept, Tsinghua University, China

Abstract

In TREC-10, Microsoft Research China (MSRCN) participated in the Web track (ad hoc retrieval task and homepage finding task). The latest version of the Okapi system (Windows 2000 version) was used. We focused on the developing of content-based retrieval and link-based retrieval, and investigated the suitable combination of the two.

For content-based retrieval, we examined the problems of weighting scheme, re-weighting and pseudo-relevance feedback (PRF). Then we developed a method called *collection refinement (CE)* for QE.

We investigated the use of two kinds of link information, link anchor and link structure. We used anchor descriptions instead of content text to build index. Furthermore, different search strategies, such as spreading activation and PageRank, have been tested.

Experimental results show: (1) Okapi system is robust and effective for web retrieval. (2) In ad hoc task, content-based retrieval achieved much better performance, and the impact of anchor text can be neglected; while for homepage finding task, both anchor text and content text provide useful information contributing more on precision and recall respectively. (3) Although query expansion does not show any improvement in our web retrieval experiments, we believe that there are still potential for CE.

1. Introduction

Microsoft Research China (MSRCN) participated in the Web track (ad hoc retrieval task and page finding task) at TREC-10. We used, for the first time, the new version of the Okapi system (which is running on Windows-2000) developed at Microsoft Research Cambridge. We focused our researches on: (1) the use of traditional IR techniques (content-based retrieval) for web retrieval, (2) the use of query expansion (QE) for web retrieval, and (3) the use of link information.

In this paper, we will explore the following issues:

(1) Testing the Windows version of the Okapi system using 10GB web collection.

(2) The impact of query expansion on web retrieval. The expansion terms are chosen from the top-ranked documents retrieved using the initial queries. We used two types of collections for initial retrieval: the 10G web collection and an external collection, i.e. the MS-Encarta collection.

(3) The relative contribution of content information and link information to web retrieval. We exploit methods of combining both kinds of information to improve the effectiveness of web retrieval.

(4) The impact of link information on web retrieval. We investigate the use of two kinds of link information: link anchor text and link connection.

In the remainder of this paper, we will discuss in turn each problem together with our approaches and results of TREC experiments. The results include official runs we submitted and additional runs that we designed to help us explore the issues. Finally, we give our conclusions and present our future work.

2. The System

We used the Okapi system Windows-2000 version for our runs. The system was developed in October 2000. A detailed summary of the contributions to TREC1-9 by the Okapi system is presented in (Roberson and Walker, 2000; Roberson and Walker, 1999). In this section, we give a very brief introduction to the system.

The search engine in Okapi is called the Basic Search System (BSS). It is a set-oriented ranked output system designed primarily for probabilistic-type retrieval of textual material using inverted indexes. There is a family of built-in weighting scheme functions known as BM25 and its variants. In addition to weighting and ranking facilities, it has the usual Boolean and quasi-boolean (positional) operations and a number of non-standard set operations. Indexes are of a fairly conventional inverted type. BSS also provides functions for blind feedback.

All the TREC-10 processing was done at Microsoft Research China. Most of the experiments were run on two DELL severs. Both machines have four 500MHz Pentinum processors with 2GB RAM, and were running on Windows-2000. The network was 100Mbps Ethernet.

3. Data Processing

The collection we used in Web track is a set of web pages downloaded from the World Wide Web. The size of the original collection is more than 10GB. It is a good challenge for the new version of the Okapi system. Four query sets were used in our experiments:

(1) TREC-9 ad hoc query set of 50 queries (denoted by T9),

(2) TREC-10 ad hoc query set of 50 queries (denoted by T10),

(3) TREC-10 page finding query set of 145 queries (denoted by P10), and

(4) A page finding training set, which includes 100 queries (the query set is denoted by P9) and the relevance judgment.

3.1 Pre-processing

Our data pre-processing includes data cleaning and information extraction.

We first removed junk from the collection. The junk includes mismatched tags, and files which contain non-text material (i.e. compressed data, etc). All lines starting with “Sever:” and “Content-type” etc. were also removed. The resulting collection is of size 6GB.

We then used an HTML parser developed at Microsoft to extract logical fields, including Title <T>, Subtitle <H1>, <H2> and <H3>, and Passage delimited by tags <P> and </P>. We also, from the collection, established two tables. One table contains the link connection information; each entry of the table is a page-pair connected by a link. The other contains link anchor text information; each entry includes an anchor text, the page containing the anchor text, and the page pointed by the anchor text. Title, Subtitle and Passage were used for content-based retrieval while the link connection and link anchor text were used for link-based retrieval. We will describe both retrieval methods in detail later.

3.2 Indexing/Query processing

For query processing, we first performed stemming using the Okapi stemmer. Stop words were then removed. We used a stop word list of 222 words (Roberson and Walker, 1999). For four query sets, we used title-only queries in our experiments.

For each web page, before indexing, all words were stemmed, and stop words were removed. The term weight is BM2500. It is a variant of BM25 and has more parameters that we can tune. BM2500 is of the form:

$$\sum_{T \in Q} w^{(1)} \frac{(k_1 + 1)tf \ (k_3 + 1)qtf}{(K + tf)(k_3 + qtf)} \tag{1}$$

where *Q* is a query containing key terms *T*, *tf* is the frequency of occurrence of the term within a specific document, *qtf* is the frequency of the term within the topic from which *Q* was derived, and $w^{(1)}$ is the Robertson/Spark Jones weight of *T* in *Q*. It is calculated by Equation (2):

$$\log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{2}$$

where *N* is the number of documents in the collection, n is the number of documents containing the term, *R* is the number of documents relevant to a specific topic, and *r* is the number of relevant documents containing the term. In Equation (1), *K* is calculated by Equation (3):

$$k_1((1-b)+b\times dl/avdl) \tag{3}$$

where *dl* and *avdl* denote the document length and the average document length measured in some suitable unit, such as word or a sequence of words.

Parameters k_1, k_3, *b*, and *avdl* are tuned by experiments to optimize the performance. In our experiments, we set k_1=1.2, k_3=1000, *b*=0.75, and *avd*=61200.

4. Basic Content-based Retrieval

For basic content-based retrieval, only initial retrievals (i.e. without QE) were performed. Only those words in fields of Title/Subtitle/Passage were indexed. The initial retrieval results are summarized in Table 1 and 2. We can see that the ad hoc retrieval results for TREC-9 query set are very promising. It is favorably comparable to the best effectiveness achieved in the previous web track experiments. This indicates the robustness and effectiveness of our new version of the Okapi system. The results in Table 1 and 2 will also serve as the baseline in all experiments described below. The evaluation metric of ad hoc task is non-interpolated average precision. The evaluation metrics of page finding task includes average reciprocal rank, top-10 precision, and not-found rate.

Query set	Avg. P.
T9	22.08%
T10	19.42%

Table 1: Baseline results of ad hoc task

Query set	Average reciprocal rank	Top-10 precision	Not-found rate
P9	19.68%	34.00%	25.00%
P10	22.46%	44.10%	25.52%

Table 2: Baseline results of page finding task

5. Query Expansion

The average length of title-only queries is less than 3 words (non stop word). It seems that query expansion is needed to deal with word mismatching problem for web retrieval. We performed query expansion experiments on ad hoc retrieval. The procedure works as follows:

(1) For each query, retrieve 10 top ranked documents by an initial retrieval;

(2) Choose 10 expansion terms from the top ranked documents. First, stop words were discarded. Then expansion terms were ranked in decreasing order of a term selection value (TSV) of the form

$$TSV = w^{(1)} * r / R \tag{4}$$

where $w^{(1)}$, R, and r are same elements described in Equation (1) and (2). The top-10 terms were added to the initial query.

As shown in Table 3, the conventional query expansion (i.e. pseudo-relevance feedback (PFB)) result is not good. We think that there might be two reasons. First, the topics of web pages are diverse. Although the expansion terms were chosen from top ranked documents, the ranking of these terms was based on the statistics over the whole collection as indicated by Equation (4). Second, the quality of documents in the web collection is highly mixed.

We adopted two methods to solve the abovementioned two problems..

First, we introduced the local context analysis (LCA) technique as proposed in (Xu and Croft, 1996). We used statistics of documents and terms from local collection (i.e. top-10 ranked document collection retrieved by an initial retrieval) to estimate the TSV for each expansion terms. That is, we set R=10, and r is the number of relevant documents containing the term in the local collection (r<10). As shown in Table 3, although the result is a little better than PRF, it is still worse than the initial retrieval.

Second, we introduced the idea of collection enhancement (CE), which was successfully applied for TREC cross language information retrieval experiments (Kwok et al., 2000). The basic idea is: if we can refine web queries by QE using documents from an external high-quality and well-organized collection, we may able to improve the web retrieval. We used MS-Encarta collection as the external collection. That is, in our experiments, the initial retrieval was performed using MS-Encarta collection. The expansion terms were chosen from the top-1 Encarta document. Notice that we did not use top-10 documents because the MS-Encarta collection is relatively small (i.e. less than 200MB). More importantly, MS-Encarta is a well-written encyclopedia with each document discussing one specific topic. So terms from multiple documents are likely of different topics and not relevant. Documents in MS-Encarta are categorized by a set of pre-define keywords and are well-organized under a domain hierarchy structure. We think that such information will be helpful for navigating the web collection, but we have not found an effective way to make use of them.

The preliminary result shown in Table 3 is not encouraging. We found that it is largely due to the difference between the Encarta collection and the web data. But we do believe it has potential if we can make good use of rich information imbedded in MS-Encarta collection (i.e. pre-defined keywords, domain hierarchy, etc.) or figure out an effective way to fuse the web collection with MS-Encarta collection.

Initial retrieval	PRF	LCA	CE
22.08%	20.89%	21.55%	18.24%

Table 3: QE results of TREC-9 ad hoc retrieval

6. Link-based Retrieval and Content-based Retrieval

Recently, the research of web retrieval has focused on link-based ranking methods. However, none had achieved better results than content-based methods in TREC experiments. We investigated the use of two kinds of link information: link anchor and link connection. Our focus was on finding the effective ways for combining link-based retrieval with content-based retrieval.

6.1 Using anchor text

We assumed that the anchor text of a link describes its target web page (Craswell et al., 2001). We then, for each web page, built an anchor description document containing all the anchor texts of a page’s incoming links.

We observed that plenty of anchor texts are names of home pages (i.e. URL, or URL-like terms), which are reasonable. Therefore intuitively, they could be very effective for page finding task, in which most queries are also a bunch of URLs or URL-like terms. Our results on TREC-10 page finding tasks confirmed the intuition. In Table 4, row 1 and row 2 show that anchor-text-based retrieval achieved much better performance than content-based retrieval, i.e. more than 96% improvements on average reciprocal rank and 48% improvements on top-10 precision.

	Average reciprocal rank	Top 10 precision	Not found	Method
1	22.46%	44.10%	25.52%	Content-based retrieval
2	44.06%	65.50%	25.52%	Anchor-text-based retrieval
3	42.40%	65.50%	13.10%	Content + anchor text (Comb-1)
4	50.50%	69.00%	15.20%	Content + anchor text (Comb-3)

Table 4: Page finding results of TREC-10 query set (P10)

We then performed experiments on ad hoc retrieval using anchor description only for indexing. The results are much worse than content-based retrieval as shown in Table 5 (row 1-3). This is due to the data sparseness problem. As indicated in Figure 1, statistics showed that about 28% web pages in the web collection have no anchor description at all. Totally 75% web pages have anchor description documents with less than 10 words. Therefore, the information (in terms of keywords) that anchor text provided for ad hoc retrieval is very limited. It is most unlikely that the title-only queries have chances to match words contained in such a short description. So even with query expansion (e.g. chose 30 terms from top-5 ranked documents), anchor-text-based retrieval was still much worse than content based retrieval although it was much better than the result without query expansion as shown in row 1-3 of Table 5.

	Avg. P using T9	Avg. P. using T10	Method
1	20.08%	19.42%	Content-based retrieval
2	3.12%	--	Anchor-text-based retrieval
3	4.85%	--	2 + query expansion
4	22.23%	19.13%	Content + anchor text (Comb-1)
5	23.27%	18.64%	Content + anchor text (Comb-3)

Table 5: Ad hoc retrieval results using anchor text

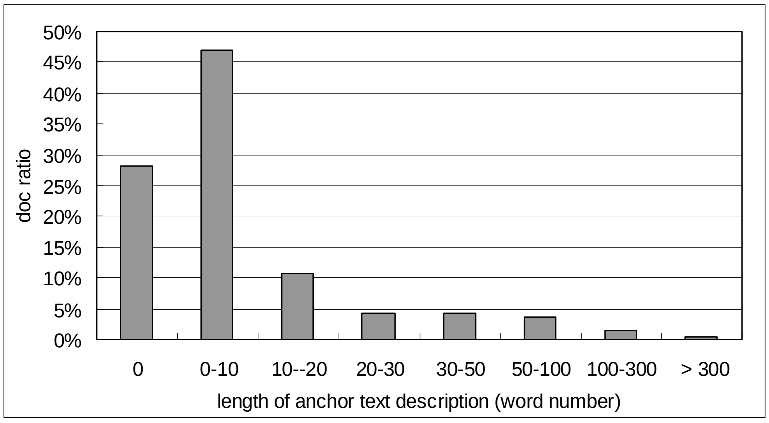

Figure 1: Anchor text description length vs. number of documents

In what follows, we examine three different ways to make use of anchor text and content text.

First, we simply combined the content text and anchor text for indexing (denoted by Comb-1 in Table 4 and Table 5).

Second, we merged the two ranking lists obtained by content-based retrieval and anchor-text-based retrieval respectively. The new score, *s*, of a retrieved page is estimated by Equation (5).

$$s = \lambda * sc + (1-\lambda)*sa \qquad (5)$$

where *sc* and *sa* are, respectively, the scores of content-based retrieval and anchor-text-based retrieval, and λ ($0 \leq \lambda \leq 1$) is the interpolation weight tuned on a test set.

The last method we used is to re-rank the results of content-based retrieval according to the results of anchor-text-based retrieval (denoted by Comb-3 in Table 4 and 5). For each retrieved page in the ranking list of content-based retrieval, if it is also included in the ranking list of anchor-text-based retrieval, we set a new score by Equation (6), where $\lambda \geq 1$.

$$s = \lambda * sc \qquad (6)$$

The page finding results are summarized in Table 5. The ad hoc retrieval results are summarized in Table 6. Let us discuss the frustrating ad hoc retrieval results first.

As we expected, since the addition indexing words provided by anchor text are very limited, the impact of combination is neglectable, as shown in row 4 of Table 5. Similarly, in the second method, we found that the best result is obtained when λ=1. This indicates again the neglectable impact of anchor text information on ad hoc retrieval. For Comb-3, we still found that the best results are obtained when λ approached 1.

The reason that anchor text is not helpful to ad hoc retrieval is largely due to the sparseness problem we discussed above. The following Figure 2 shows the results of query by query analyses on TREC-10 ad hoc retrieval task. Because most of the anchor texts are too short, they are submerged in the content data. At the same time, what most commonly happens is that the query and the anchor text are mismatched for both of them are extremely short. Since no good result can be achieved by using anchor description of the document only, no improvement may be obtained by combining anchor text retrieval result and content text retrieval results.

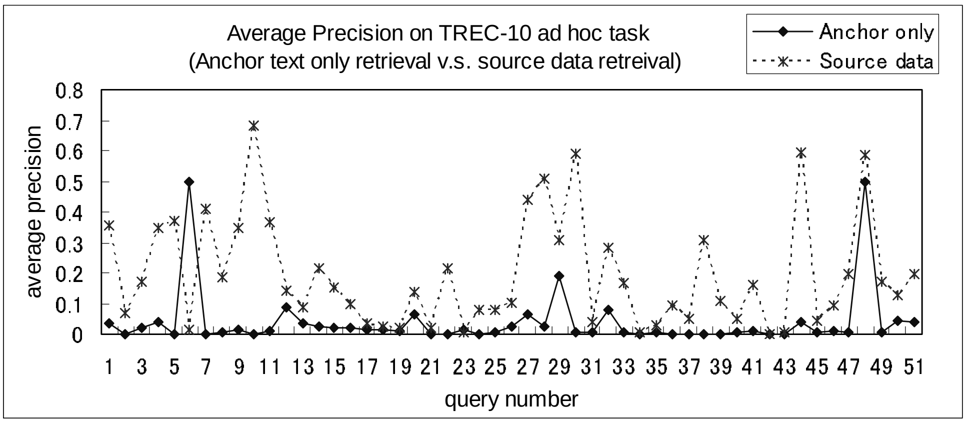

Figure 2: Average precision on TREC-10 ad hoc task (using anchor text v.s. source data)

Now let's look at the good results achieved in page finding task by using both content text and anchor text.

We applied Comb-1 and Comb-3. Unlike the ad hoc retrieval, experimental results on page finding task are much more encouraging as shown in Table 4. Row 3 shows that when using Comb-1, although we did not get any improvements on average reciprocal rank and top-10 precision, the not found rate dropped dramatically by more than 48%. As shown in row 4, by using Comb-3, we obtained even better performance. We achieved approximately 15% improvement on average reciprocal rank, and 56% improvement on top-10 precision. The not found rate also dropped substantially by more than 40%.

We give the similar query by query analyses for TREC-10 homepage finding task, the first 50 queries of which are shown in Figure 3. That is to say, we evaluate the retrieval result by non-interpolated 11 points average precision metric, which shows the performance in terms of both precision and recall. For the remaining queries, the results are most similarly. On a whole, there are 90 queries that can get better performance by using anchor description than using source data for indexing; only 38 queries are worse than source data retrieval; and for the remaining 17 queries, both anchor description retrieval and source date retrieval get the same results. Since anchor text takes limited but precise information of a homepage, especially the URL feature of the page, it can get better performance. Then it is reasonable to make improvements while combining two different ranked lists of retrieval results.

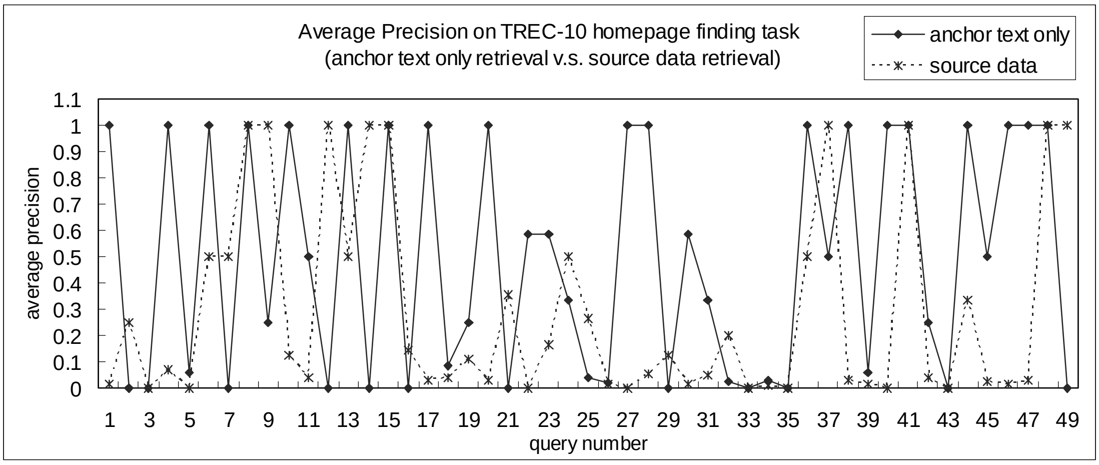

Figure 3: Average precision on TREC-10 homepage finding task (using anchor text v.s. source data)

The results indicate that (1) anchor text containing less but URL-like terms which contributed more to the precision of page finding; (2) content text with more terms might contribute more to the recall; and (3) when we combined anchor text and content text for indexing, both kinds of information really complemented each other, and achieved a better tradeoff between precision and recall.

6.2 Using link connection

We assumed that links between web pages indicate useful semantic relationships between related web pages. Especially, we tried spreading activation (SA) approach (Crestani and Lee, 2000; Savoy and Rasolofo, 2000) for ad hoc task using TREC-9 query set. In SA method, the degree of match between a web page D_i and a query Q, as initially computed by the IR system (denoted $SIM(D_i,Q)$), is propagated to the linked documents through a certain number of cycles using a propagation factor. Savoy and Rasolofo (2000) used a simplified version with only one cycle and a fixed propagation factor λ for k-best incoming links and k-best outgoing links. Our experiments showed that considering outgoing links negatively affects the retrieval results. Therefore only the top-1 similar incoming link is considered in our methods. In this case, the final retrieval value of a document D_i with m incoming linked documents is computed as:

$$\text{SAscore}(D_i) = SIM(D_i, Q) + \lambda \cdot \max\{SIM(D_j, Q) \mid j = 1, \ldots m\} \tag{7}$$

Unfortunately, we found that the best result could be obtained only when λ approached 0.

In addition to SA, different search strategies, such as PageRank etc have been tested. However, none was able to improve the retrieval effectiveness. This result confirmed the previous results in TREC using link connections.

7. Summary of Official Retrieval Results

In TREC-10, we submitted 5 official runs for ad hoc task, and 4 runs for page finding task. In both tasks, title-only query sets were used. Table 6 and 7 show the results as well as methods we used for our submitted runs.

Run #	Avg. P	Method
Msrcn1	19.42%	Content-based retrieval
Msrcn2	19.13%	Content + anchor text (Comb-1)
Msrcn3	18.64%	Content + anchor text (Comb-3)
Msrcn4	17.79%	Content + anchor text (Comb-3) +LCA
Msrcn5	18.80%	Content + anchor text (Comb-1) + PRF

Table 6: Ad hoc official results of submitted runs

Run#	Average reciprocal rank	Top 10 precision	Not found	Method
Msrcnp1	22.46%	44.10%	25.52%	Content-based retrieval
Msrcnp2	42.40%	65.50%	13.10%	Content + anchor text (Comb-1)
Msrcnp3	44.06%	65.50%	25.52%	Anchor-text-based retrieval
Msrcnp4	50.50%	69.00%	15.20%	Content + anchor text (Comb-3)

Table 7: Page finding official results of submitted runs

8. Conclusions and Future Work

In this paper, we described our work in the Web track (ad hoc retrieval task and page finding task) evaluated at TREC-10. We used the latest version of the Okapi system (Windows 2000 version), and focused our researches on: (1) the use of traditional IR techniques (content-based retrieval) for web retrieval, (2) the use of query expansion, and (3) the use of link information.

Several conclusions are suggested by our experiments.

(1) The new version of the Okapi system was shown to be very robust and effective for web retrieval.

(2) In ad hoc task, content-based retrieval achieved much better performance than link-based retrieval. This confirmed again the previous Web track results at TREC.

(3) In ad hoc task, the impact of anchor text could be neglected. This might be due to the problem of the nature of the TREC web collection, such as the sparseness problem mentioned in Section 6. Other groups have reported that similar methods can achieve improvements on other web collection than TREC collection (Dumais and Jin, 2001).

(4) In page finding task, anchor-text-based retrieval achieved much better results than content-based retrieval in spite of much less terms contained in the anchor description. This is perhaps because terms in page finding queries and anchor text are very similar (i.e. URL, or URL-like terms).

(5) Both anchor text and content text provided useful information for page finding. In particular, anchor text contributed more to the precision of page finding, while content text contributed more to the recall. Both kinds of information complemented each other. The combination thus achieved a better tradeoff between precision and recall.

(6) Although query expansion did not show any improvement in our web retrieval experiments, we think that there are still potential for CE if we can make good use of other rich information imbedded in the well-organized high-quality external collection (MS-Encarta) or figure out an effective way to combine the web collection with the external collection.

Our future work includes

(1) Study the nature of the web collection, and exploit the use of link information on a more ‘complete’ web collection.

(2) Enrich the anchor description by using context information of the anchor. The context information can be a sentence or a passage that contains the anchor text. The context information may enhance the anchor description from two aspects: (1) providing clues to evaluate the relevance between the anchor text and its target web page; (2) providing richer description of the target web page.

(3) Exploit the use of the external collection for QE including the use of information of domain hierarchy, pre-defined keywords, and topics etc, and the effective combination of external collection with web collection, etc.

References

Bhrat, K., and Henzinger, M. R., (1998). Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *SIGIR-98*.

Brin, S., and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *WWW7*.

Craswell, N., Hawking, D., and Robertson, S.E. (2001). Effective site finding using link anchor information. In *SIGIR-01*.

Crestani, F., and Lee, P. L. (2000). Searching the web by constrained spreading activation. *Information Processing & Management*, 36(4), page 585-605.

Dumais, S., and Jin, R. (2001). Probabilistic combination of content and links. In *SIGIR-01*.

Kowk, K. L., Grunfeld, N., and Chan, M. (2000). TREC-9 cross language, web and question-answering track experiments using PIRCS. In *TREC-9*.

Robertson, S. E., and Walker, S. (1999). Okapi/Keenbow at TREC-8. In *TREC-9*.

Robertson, S. E., and Walker, S. (2000). Microsoft Cambridge at TREC-9: Filtering track. In *TREC-9*.

Savoy, J., and Rasolofo, Y. (2000). Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections. In *TREC-9*.

Xu, J. and Croft, W. (1996). Query expansion using local and global document analysis. In *SIGIR-96*.