PERGAMON

Information Processing and Management 37 (2001) 543–569

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Retrieval effectiveness on the web

Jacques Savoy [*], Justin Picard

Institut interfacultaire d'informative, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland

Received 4 April 2000; accepted 26 June 2000

Abstract

Search engines play an essential role in the usability of Internet-based information systems and without them the web would certainly break down or, at the very least would develop at a much slower rate. Our main objective is to analyze and evaluate the retrieval effectiveness of various indexing and searching strategies on a new web text collection, using a rigorous evaluation methodology. Our second aim is to describe and evaluate different preprocessing techniques that might be implemented in order to improve retrieval effectiveness. As a third objective, this paper will evaluate whether or not hyperlinks may serve as useful sources of evidence in improving retrieval algorithms. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Web search engines; Evaluation; Hyperlinks

1. Introduction

The increasing amount of information available on the web raises new and challenging problems for the IR community. Due to the huge number of pages and links, browsing cannot be viewed as an adequate searching process, even with the introduction of subject directories or classified lists (e.g., Yahoo!) (Alschuler, 1989). As a result, effective query-based mechanisms for accessing information are needed, particularly by the 85% of web users for whom search engines are a primary tool (Lawrence & Giles, 1999; Schwartz, 1998). Retrieval mechanisms currently proposed (Leighton & Srivastava, 1999; Gordon & Pathak, 1999) are based on conventional IR models (Salton, 1989) within which a centralized document index is assumed. These search engines are not able to make effective use of all available information (Lawrence & Giles, 1999), and

[*] Corresponding author.

E-mail address: jacques.savoy@seco.unine.ch (J. Savoy).

PII: S0306-4573(00)00039-X

most of them ignore hypertext links as a mean of enhancing their retrieval effectiveness. Recent works in web IR seem to acknowledge that hyperlink structures can be very valuable in locating information (Marchiori, 1997; Kleinberg, 1998; Brin & Page, 1998; Bharat & Henzinger, 1998). According to Chakrabarti, Van den Berg and Dom (1999):

> Citations signify deliberate judgment by the page author. Although some fraction of citations are noisy, most citations are to semantically related material. Thus the relevance of a page is a reasonable indicator of the relevance of its neighbors, although the reliability of this rule falls off rapidly with increasing radius on average. Secondly, multiple citations from a single document are likely to cite semantically related documents as well (Chakrabarti et al., 1999, pp. 550–551).

With small variations, similar hypotheses are also cited by other authors (Kleinberg, 1998; Bharat & Henzinger, 1998). Our previous studies on citation schemes (Savoy, 1994, 1996, 1997a; Picard, 1998) tend to suggest however that citation information might improve average precision, but only on the order of 5–10% when used with "good" retrieval schemes. Is this conclusion also valid for the web?

In Chapter 2 of this paper, we will describe and evaluate various indexing and retrieval models based on a web test collection and we will compare these results with those obtained with more traditional IR corpora. In Chapter 3, we will describe experiments that verify whether or not hyperlinks do improve retrieval effectiveness. If this paper is mainly concerned with retrieval effectiveness, we must acknowledge that discovery of web pages (web crawlers) (Bailey, Craswell, & Hawking, 2000) or problems of efficiency are also of primary concerns in commercial web-based IR systems (Lesk, 1998). Moreover, additional network costs (number of servers accessed, complexity and length of the queries, amount of data transferred, network latencies and time-outs) must also be examined. Finally, consideration must be given to the financial cost or pricing related to the access of certain collections (proprietary collections) (Lesk, 1997), or to security problems (e.g., implementation of access rights for different databases, authority controls).

2. Evaluation of various indexing and searching strategies

Automatic retrieval of information by computers can be viewed as a complex problem due to the underlying ambiguity of all natural languages. On the one hand, authors and users frequently apply different words or expressions to refer to the same meaning ("accident" may be expressed as "event", "incident", "situation", "problem", "difficulty", "unfortunate situation", "the subject of your last letter", "what happened last week", etc.) (Furnas, Landauer, Gomez, & Dumais, 1987). On the other hand, a specific term may have different (and sometimes contradictory) meanings and interpretations (e.g., polysemy relative to the word "horse" as in "Trojan horse", "light horse", "to work like a horse", "from the horse's mouth", "horse about" or relative to the word "lead" in "environment Canada plays a lead role …", "lead pollution" and "lead mining"). Moreover, other linguistic phenomena such as anaphora, ellipses, pronominal references, spelling errors etc. tend to render indexing and matching requests to documents as an imprecise, incomplete and uncertain process. Thus a computer cannot infer that a string match would always imply a match relative to a word's true sense.

This chapter will investigate whether or not various techniques used in IR do indeed improve retrieval effectiveness when applied to our web test collection. It has been debated that web engines work in radically different environments, within which classical IR techniques are largely irrelevant (Raghavan, Broder, Henzinger, Manber, & Pinkerton, 1999). Thus, the use of a stemming procedure, the presence of a stopword list, the relative importance of words appearing in the title section of a web page, and blind query expansion may become questionable techniques when dealing with web pages. In other words, search models that have relatively good average precision when used in IR test collections may not perform as well when using web pages.

In order to respond to these questions, this chapter is organized as follows. Section 2.1 presents an overview of our web test collection and Section 2.2 describes different vector-processing schemes used in this paper together and also the Okapi probabilistic model. Section 2.3 evaluates these search models using our web test collection. The preprocessing of web pages is outlined in Sections 2.4 and 2.5 describes a search technique which assigns more importance to words appearing in the title section. A more selective indexing procedure is described and evaluated in Section 2.6 and the retrieval effectiveness of the stemming procedure and stopword list is elaborated in Section 2.7. The impact of blind query expansion is analyzed in Section 2.8. Finally, Section 2.9 compares the average precision obtained from using the web test collection vs. that of the TREC test collection.

2.1. Overview of the web test collection

Some statistics describing our web test collection (WT2g of the TREC data, Hawking, Voorhees, Bailey, & Craswell, 1999b) are depicted in Table 1, while other characteristics are given in (Hawking, Craswell, Thistlewaite, & Harman, 1999a). In this corpus are found 100 queries (examples are given in Appendix B), where the requests, instead of being limited to a narrow subject range, reflect various information needs (such as "Falkland petroleum exploration", "journalist risks", "El Niño" or "piracy"). To reflect the fact that most queries on the web are relatively short (2.21 terms, Jansen, Spink, & Saracevic, 2000), our experiments are mainly based on the Title section of the requests having an average length of 2.4 terms. Of course in other environments, the length of queries submitted may be longer (e.g., in commercial IR systems, a mean query length of 14.8 search terms has been reported, Spink & Saracevic, 1997). To reflect this second type of user we also performed some computations based on the Descriptive and Narrative logical sections of our queries.

For these requests, as described in Hawking et al. (1999a), the establishment of relevance judgments and evaluation methodologies are based on a rigorous approach (for example, Gordon & Pathak, 1999). Of course, we recognize that there are dozens and dozens of extremely complex and poorly understood personal and document factors that together can determine whether anyone would judge, a given document to be relevant, and this in turn leads to the conclusion that the nature of relevance is subjective (Saracevic, 1975; Robertson, Maron, & Cooper, 1982). Moreover, we must maintain that all evaluation measures are relative to the retrieved documents judged by the assessors. This indicates that not every page on the web collection would have been judged according to each query due to the limited resources.

Table 1
Web test collection statistics

Size (in MB)	2194 MB
# of web pages extracted from 969 URLs	247,491
# of distinct indexing terms in the collection	1,850,979
# of requests	100
# of distinct index terms/web page	
Mean	218.25
Standard error	326.42
Median	125
Maximum	22,722
Minimum	1
# of indexing terms/web page	
Mean	554.295
Standard error	1,402.86
Median	213
Maximum	179,303
Minimum	1
Time required to build the inverted file	
(user time)	26:28
Elapsed time	1:44:44
# of relevant web pages (100 queries)	8868
Mean	88.68
Standard error	117.48
Median	49.5
Maximum (Query #360)	828
Minimum (Query #423)	6

This corpus contains 1,171,795 hyperlinks for an average of 4.73 hyperlinks per page (used primarily for navigational purposes across the web sites). In comparison to the web, which was estimated in February 1999 to contain about 800 million web pages (Lawrence & Giles, 1999), our test collection might be viewed as being relatively small. This results consequently in the risk that a large portion of the hyperlinks between pages having different URLs (defined as the IP number) are outside of our collection and hence will be unusable. According to our computations, there were 2797 hyperlinks to pages on different hosts, representing 0.24% of the total. These findings corroborate Bray's study (1996), which reveals that almost 80% of sites contain no off-site URLs.

The data in Table 1 also shows also that the mean number of relevant pages per request is relatively low (88.68) compared to other performance studies of search engines on the web. Moreover, 50 queries have less than 50 relevant items indicating that rather strict and specific relevance assessments have been done. Given the number of queries (100 in this study) and the specific relevance interpretations, a direct comparison with other studies cannot be done. Moreover, rare topics (e.g., topics having a small number of relevant answers) do not have enough connectivity to be exploited, or in some circumstances to question the usefulness of hyperlinks.

2.2. Indexing and searching strategies

In order to define a retrieval model, we need to first explain how documents and queries are represented and then how these representations are compared, thus resulting in a ranked list of retrieved items. As a first approach, we might adopt a binary indexing scheme within which each document and request is represented by a set of keywords without any weight. To measure the similarity between the document and the request, we may count the number of terms they have in common (retrieval model denoted "doc = bnn, query = bnn", details are given Appendix A).

Binary logical restrictions may often be too restrictive for document and query indexing. It is not always clear whether or not a web page should be indexed by a given term. Often, a more appropriate answer is neither "yes" nor "no", but rather something in between. Thus, term weighting creates a distinction among terms and increases indexing flexibility. In this vein, we may assume that a term's frequency of occurrence in a document or in a query (denoted tf) can be a useful feature. Consequently, terms appearing once have a weight of 1 while terms occurring twice have a weight of 2. As previously, the similarity between a web page and the request is based on the number of terms they have in common, weighted by the component tf (retrieval model notation: "doc = nnn, query = nnn").

As a third IR model (Salton, 1989), we may also consider that those terms occurring very frequently in the collection do not help us discriminate between relevant and non-relevant items. Thus we might count their frequency in the collection, or more precisely the inverse document frequency (denoted by idf), resulting in a larger weight for sparse words and a smaller weight for more frequent ones. In this case, higher weights are given to terms appearing often in the web page (tf component) and rarely in other web pages (idf component). As such, each term does not have an equivalent discrimination power, and a match on a narrow keyword must therefore be treated as being more valuable than a match on a more common word. Moreover, using cosine normalization (retrieval model notation: "doc = ntc, query = ntc"), may prove beneficial because it normalizes each indexing weight within a range of 0–1.

We are also able to create other variants, especially when considering that the occurrence of a given term in a document is a rare event (we have 1,850,979 distinct terms in our test collection). Thus, it may be a good practice to give more importance to the first occurrence of this word as compared to its successive repeating occurrences. Therefore, the tf component may be computed as the ln(tf) + 1.0 (retrieval model notation: "doc = ltc, query = ltc") or as 0.5 + 0.5 [tf/max tf in a web page]. In this latter case, the normalization procedure is obtained by dividing tf by the maximum tf value for any term in the document (retrieval model denoted "doc = atn"). Of course, different weighting formula may be used for web pages and the request, leading to other different weighting combinations.

Finally, we may consider that the presence of a term provides stronger evidence in a short as compared to a long web page. To account for this, document length can be integrated within the weighting formula, leading to a more complex IR model; for example, the IR model denoted by "doc = Lnu" (Buckley, Singhal, Mitra & Salton, 1996) or the Okapi probabilistic search model (Robertson, Walker, & Beaulieu, 2000). In these schemes, a match on a small web page will be treated as more valuable than a match on a longer document. The question that then arises is: How will these retrieval models behave when used with our web corpus?

2.3. Evaluation of various retrieval models

As a retrieval effectiveness indicator, we may measure the precision after 10 items are retrieved (percentage of relevant items) or after 20 items. In the IR domain and in the TREC conferences, the preference has been to adopt the non-interpolated average precision as a retrieval effectiveness measure (computed on the basis of 1000 retrieved items per request) to account for both precision and recall using a single number (Voorhees & Harman, 2000). Moreover, the average precision accounts for the rank of retrieved and relevant items while the precision at 10 returns the same value if, for example, the three relevant and returned records are the first three or have ranking positions of 8–10.

A decision rule is required to determine whether or not a given search strategy is better than another. The following rule of thumb could serve this purpose: where a difference of at least 5% in average precision is generally considered significant and a 10% difference is considered material (Sparck Jones & Bates, 1977, p. A25).

For a more precise decision methodology, we might also apply statistical inference methods such as Wilcoxon's signed rank test or Sign test (Salton & McGill, 1983, Section 5.2; Hull, 1993) or hypothesis testing based on bootstrap methodology (Savoy, 1997b). More precisely, the null hypothesis H_0 states that both retrieval schemes produce similar performance. Such a null hypothesis plays the role of a devil's advocate, and this assumption will be accepted if two retrieval schemes return statistically similar means, and rejected if not. Thus, in the tables of this paper, we have underlined statistically significant differences based on a one-sided non-parametric bootstrap test based on the means with a significance level fixed at 1%.

In Table 2, we computed these three evaluation measures for a set of retrieval models using Lovins' stemming (1982), stopword elimination and case insensitive matching. As mentioned in (Tague-Sutcliffe & Blustein, 1995), these retrieval performance measures are highly correlated. For example, for the 10 retrieval models studied, the correlation coefficient between the average precision and the precision after 10 retrieved documents is 0.966 in our case, while the correlation coefficient between the mean precision and the precision after 20 retrieved documents is 0.977.

Table 2
Precision of various indexing and searching strategies

Model\query (Title only)	Precision (% change)		
	Average	Precision @ 10	Precision @ 20
doc = Okapi query = npn	26.68	44.9	39.5
doc = Lnu, query = ltc	23.38 (−12.37%)	43.00 (−4.23%)	38.65 (−2.15%)
doc = atn, query = ntc	25.67 (−3.79%)	36.60 (−18.49%)	34.90 (−11.65%)
doc = ntc, query = ntc	13.85 (−48.09%)	27.10 (−39.64%)	23.80 (−39.75%)
doc = ltc, query = ltc	13.67 (−48.76%)	18.00 (−59.91%)	17.15 (−56.58%)
doc = lnc, query = ltc	10.72 (−59.82%)	16.00 (−64.37%)	15.90 (−59.75%)
doc = lnc, query = lnc	7.24 (−72.86%)	12.00 (−73.27%)	11.25 (−7152%)
doc = anc, query = ltc	8.23 (−69.15%)	11.30 (−74.83%)	11.60 (−70.63%)
doc = nnn, query = nnn	7.12 (−73.31%)	13.70 (−69.49%)	12.45 (−68.48%)
doc = bnn, query = bnn	9.56 (−64.17%)	14.00 (−68.82%)	12.15 (−69.24%)

Thus, the average precision will be used in our further evaluations but similar conclusions can be drawn using the precision after 10 or after 20 retrieved items.

From the results shown in Table 2, we find that the Okapi probabilistic model provides the best performance, significantly better than the vector-scheme ("doc = Lnu, query = ltc"). The IR model "doc = atn, query = ntc" performs well, but still not as well as the Okapi search approach. However, based on the bootstrap test, the difference cannot be viewed as significant (significance level of 1%). The traditional tf–idf weighting scheme ("doc = ntc, query = ntc") does not produce very satisfactory results. Finally, the simple term-frequency weighting scheme ("doc = nnn, query = nnn") or the simple coordinate match ("doc = bnn, query = bnn") results in poor retrieval performance.

The impact of query length on the improvement of search performance is listed in Table 3, showing three different query formulations: (1) only the Title section, (2) both the Title and Descriptive sections or (3) all three sections (Title, Descriptive and Narrative). Table 3 shows that retrieval effectiveness is significantly enhanced when topics include more search terms. This finding does not hold however when there is a simple coordinate match ("doc = bnn, query = bnn") or a simple term-frequency weighting scheme ("doc = nnn, query = nnn"). This phenomenon seems to demonstrate that search terms extracted from the Descriptive or Narrative sections have less discrimination ability.

The use of overall statistics such as the average precision may hide performance irregularities among requests when comparing two or more retrieval schemes. Based on 100 queries, Table 4 depicts average precision, and for each retrieval scheme the number of best individual runs on a per query basis. Thus, for 43 queries out of 100, the best choice is the Okapi strategy. From using this search model, 99 queries resulted in average precision over the median, leading to the conclusion that this retrieval model has good overall performance. It is interesting to note that the

Table 3
Average precision of various search models using different query formulations

Model		Average precision (% change)		
	Query:	Title	Title and Descriptive	Title, Descriptive and Narrative
	Mean search terms:	2.4 terms	5.71 terms	14.71 terms
doc = Okapi, query = npn		26.68	31.83 (+19.30%)	35.32 (+32.38%)
doc = Lnu, query = ltc		23.38	28.26 (+20.87%)	30.22 (+29.26%)
doc = atn, query = ntc		25.67	28.25 (+10.05%)	28.53 (+11.14%)
doc = ntc, query = ntc		13.85	15.02 (+8.45%)	15.97 (+15.31%)
doc = ltc, query = ltc		13.67	15.58 (+13.97%)	18.10 (+32.41%)
doc = lnc, query = ltc		10.72	13.62 (+27.05%)	17.32 (+61.57%)
doc = lnc, query = lnc		7.24	9.42 (+30.11%)	12.24 (+69.06%)
doc = anc, query = ltc		8.23	10.55 (+28.19%)	13.54 (+64.52%)
doc = nnn, query = nnn		7.12	3.64 (−48.88%)	2.86 (−59.83%)
doc = bnn, query = bnn		9.56	8.18 (−14.44%)	3.96 (−58.58%)
Search time (user + system)/request		0.3033	0.5279	0.8185

Table 4
Characteristics of individual retrieval schemes

Model	Average precision	Best IR scheme for #Queries	#Queries over the median
doc = Okapi, query = npn	26.68	43	99
doc = Lnu, query = ltc	23.38	13	97
doc = atn, query = ntc	25.67	27	97
doc = ntc, query = ntc	13.85	5	56
doc = ltc, query = ltc	13.67	2	48
doc = lnc, query = ltc	10.72	0	27
doc = lnc, query = lnc	7.24	0	11
doc = anc, query = ltc	8.23	1	7
doc = nnn, query = nnn	7.12	8	28
doc = bnn, query = bnn	9.56	1	30

best vector-space approach (“doc = atn, query = ntc”) provides the best results for 27 queries out of 100 and has good overall performance (97 queries have an average performance over the median). This data also shows that even a simple retrieval scheme such as the vector-space model, based on a simple frequency count (“doc = nnn, query = nnn”) represents the best scheme for 8 out of 100 requests. Finally, other vector processing schemes (“doc = ntc, query = ntc” or “doc = ltc, query = ltc”) do not perform well in general.

Moreover, the Okapi probabilistic model performs better than the “doc = Lnu, query = ltc” for 76 queries from a total of 100. In the same way, the Okapi approach results in better average precision for 60 requests over the vector-processing scheme “doc = atn, query = ntc”.

Tables 2 and 3 summarizes our various experiments. The average precision depicted in these tables is however based on different methods of preprocessing, which will be discussed in the following sections.

To obtain a better view of how each object (query or search model) corresponds to the composite score, we may use correspondence analysis (Greenacre, 1984). To display all the relationships in the current case among the retrieval models, we need a nine-dimensional space, but we may project them into a two-dimensional space that contains 67.6% of all the information. However, when we try to understand the geometry of a set of numerous dimensional points through the use of an approximated two-dimensional plane, we must be careful to interpret the figure correctly. An object near the center of the graph which scored similarly for all aspects (e.g., “doc = Lnu, query = ltc” has a retrieval effectiveness “similar” to the average profile). An object drawn away from the center differs from the centroid (graph center) and differs from objects drawn away from the center in another direction. However, the distances between points do not allow a straightforward interpretation (see Fig. 1).

2.4. Preprocessing of web pages

From the original web pages, we retained only the following logical sections: <TITLE>, <H1>, <CENTER>, <BIG>, with the most common tags <P> (or <p>, together with </P>, </p>) being removed. Text delimited by the tags <DOCHDR>, </DOCHDR> were also

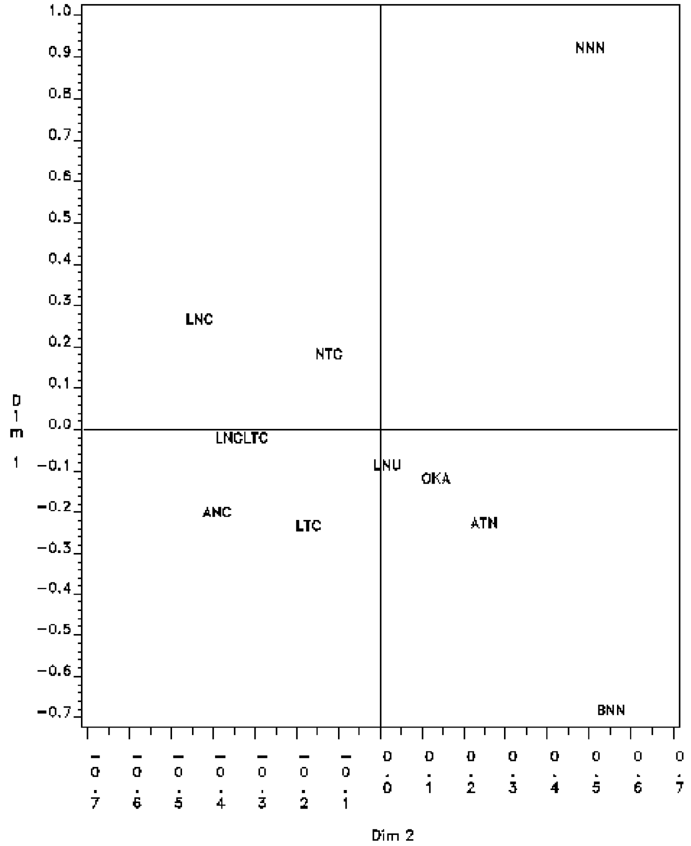


Fig. 1. Synthetic view of each search model.

Table 5
Average precision with and without preprocessing of web pages

Model\query (Title)	Average precision (% change)	
	Original	Preprocessed
doc = Okapi, query = npn	21.56	26.68 (+23.75%)
doc = Lnu, query = ltc	18.63	23.38 (+25.50%)
doc = atn, query = ntc	21.18	25.67 (+21.20%)
doc = ntc, query = ntc	11.7	13.85 (+18.38%)
doc = ltc, query = ltc	12.55	13.67 (+8.92%)
doc = lnc, query = ltc	10.31	10.72 (+3.98%)
doc = lnc, query = lnc	7.33	7.24 (−1.23%)
doc = anc, query = ltc	8.41	8.23 (−2.14%)
doc = nnn, query = nnn	5.56	7.12 (+28.06%)
doc = bnn, query = bnn	7.96	9.56 (+20.10%)

removed. As shown in Table 5, preprocessing the web pages significantly improves retrieval effectiveness for most retrieval strategies. For some retrieval models, the boostrap approach cannot detect a significant difference. Moreover, the size of the source text file is reduced by around 32%.

2.5. Impact of Title and H1 logical sections on retrieval

From the resulting web pages, we might consider that some logical sections contain more valuable indexing terms than others. In order to evaluate such a hypothesis, we assume the Title and H1 sections to be potentially useful sections. Table 6 depicts the retrieval effectiveness achieved when doubling the occurrence frequency of words appearing in the Title or in H1 logical section.

Of course, varying the importance of terms appearing in the Title or H1 sections does not modify the retrieval effectiveness achieved by the simple coordinate match ("doc = bnn, query = bnn"), for which the retrieval status value is computed according to the number of terms in common between the request and web pages. In fact, for most retrieval models, according more importance to keywords appearing in the Title or H1 logical sections does not have a significant effect on average precision. However, the bootstrap method and the Sign test detect a significant difference in two cases. For both, the variation is rather small but appears on many queries (61 times for the "doc = lnc, query = ltc" strategy with seven requests having the same retrieval performance; 50 times for the "doc = nnn, query = nnn" scheme with 29 queries presenting the same average precision).

2.6. Indexing limited to specific logical sections

For reasons of efficiency, when retrieving information from the web, it can be beneficial to index a given page based only on the Title field, or to use both the Title and H1 logical sections. Data depicted in Table 7 shows that the inverted file is reduced by a factor of 10. Such an indexing approach does however reduce the retrieval performance to a large extent, as shown in Table 7.

2.7. Stopword list and stemming procedure

Some search engines available on the web do not use a stopword list or do not employ an automatic stemming procedure. The stopword list contains non-significant words that are

Table 6
Average precision when varying <Title> and <H1> importance

Model\query (Title only)	Average precision (% change)		
	Title 1×	Title 2×	H1 2×
doc = Okapi, query = npn	26.68	26.56 (−0.45%)	26.66 (−0.07%)
doc = Lnu, query = ltc	23.38	22.64 (−3.17%)	23.37 (−0.04%)
doc = atn, query = ntc	25.67	25.87 (+0.78%)	25.74 (+0.27%)
doc = ntc, query = ntc	13.85	13.71 (−1.01%)	13.76 (−0.65%)
doc = ltc, query = ltc	13.67	13.55 (−0.88%)	13.59 (−0.58%)
doc = lnc, query = ltc	10.72	10.58 (−1.31%)	10.65 (−0.65%)
doc = lnc, query = lnc	7.24	7.09 (−2.07%)	7.23 (−0.14%)
doc = anc, query = ltc	8.23	8.20 (−0.36%)	8.21 (−0.24%)
doc = nnn, query = nnn	7.12	7.16 (+0.56%)	7.14 (+0.28%)
doc = bnn, query = bnn	9.56	9.56 (0.00%)	9.56 (0.00%)

Table 7
Average precision using different logical sections of web pages

Model\query (Title only)	Average precision		
	All sections	Title only	Title & H1
doc = Okapi, query = npn	26.68	4.03	4.26
doc = Lnu, query = ltc	23.38	3.64	3.94
doc = atn, query = ntc	25.67	3.57	4.30
doc = ntc, query = ntc	13.85	3.35	3.75
doc = ltc, query = ltc	13.67	3.35	3.69
doc = lnc, query = ltc	10.72	3.28	3.53
doc = lnc, query = lnc	7.24	2.56	2.82
doc = anc, query = ltc	8.23	3.27	3.47
doc = nnn, query = nnn	7.12	2.29	1.58
doc = bnn, query = bnn	9.56	2.45	2.60
Inverted file size (MB)	465.02	49.00	50.10

removed from a web page or a request before beginning the indexing process. The stemming procedure tries to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root (e.g., “kings” is reduced to “king”, “banking” to “bank”, “reduced” to “reduce”). The usefulness of a stopword list or an automatic stemming procedure is sometimes questionable (e.g., in a query like “Vitamin A”, the word “A” is clearly important).

The data in Table 8 indicates that a stopword list is a good technique both for reducing the inverted file size and for improving retrieval effectiveness. The Lovins’ (1982) stemming procedure used in our experiments seems to be beneficial for most retrieval models except for “doc = nnn, query = nnn” and “doc = bnn, query = bnn” (for which the bootstrap method may find a significant difference only when fixing the significance level at 5%). A second advantage of stemming

Table 8
Average precision without stopword list or without stemming procedure

Model	Average precision (% change)		
	With stopword list & stemming	Without stopword list	Without stemming
doc = Okapi, query = npn	26.68	22.82 (−14.47%)	21.12 (−20.84%)
doc = Lnu, query = ltc	23.38	20.21 (−13.56%)	18.29 (−21.77%)
doc = atn, query = ntc	25.67	22.82 (−11.10%)	22.46 (−12.50%)
doc = ntc, query = ntc	13.85	13.69 (−1.16%)	12.94 (−6.57%)
doc = ltc, query = ltc	13.67	13.69 (+0.14%)	13.80 (+0.95%)
doc = lnc, query = ltc	10.72	10.04 (−6.34%)	10.23 (−4.57%)
doc = lnc, query = lnc	7.24	6.63 (−8.42%)	6.96 (−3.87%)
doc = anc, query = ltc	8.23	7.51 (−8.75%)	7.43 (−9.72%)
doc = nnn, query = nnn	7.12	7.02 (−1.40%)	8.32 (+16.85%)
doc = bnn, query = bnn	9.56	9.60 (+0.42%)	10.41 (+8.89%)
Inverted file size (MB)	403.8	523.8	607.1

is that it reduces the inverted file size (as depicted in the last row of Table 8). Moreover, these findings corroborate other studies, even those based on other languages such as French (Savoy, 1999).

2.8. Pseudo-relevance feedback

It is recognized that pseudo-relevance feedback (blind expansion) is a useful technique for enhancing retrieval effectiveness. For example, we evaluated the Okapi search model with and without query expansion in order to verify whether or not this technique might improve retrieval performance when faced with different query formulations (a technique that is known to be time-consuming). In this study, we adopted Rocchio's approach (Buckley, Singhal, Mitra, & Salton, 1996) with $\alpha = 0.75$, $\beta = 0.75$, where the system is allowed to add to the original query 17 search terms which are extracted from the 30-best ranked documents. The resulting retrieval effectiveness is depicted in Table 9.

Pseudo-relevance feedback results in satisfactory and significant enhancement over baseline performance. This improvement is more important when dealing with short queries (2.4 search terms in average), and is clearly effective in terms of retrieval performance. However, as shown in Table 9, its expensive computation cost may render this approach less attractive in the web context.

2.9. Comparison of search strategies on the web and on TREC-8 corpora

The web is in many respects very different from standard text collections. This has led some authors to affirm that the techniques developed for "classic" information retrieval applications are not well adapted to deal within the web environment. For example:

> Hence, it is mostly by chance that any classic IR method works on the web – most don't (Raghavan et al., 1999, p. 683).

Table 9
Average precision and search time with blind query expansion

Model		Average precision (% change)		
	Query:	Title	Title and Descriptive	Title, Descriptive and Narrative
	Mean search terms:	2.4 terms	5.71 terms	14.71 terms
doc = Okapi, query = npn		26.68	31.83	35.32
With query expansion		32.77 (+22.8%)	35.92 (+12.8%)	37.32 (+5.7%)
		Search time in s (% change)		
Search time (original)/request		0.3033	0.527	0.8185
Search time (expand)/request		4.570 (+1406%)	4.748 (+999%)	5.138 (+527%)

However, the experimental evidence we have gathered so far in this study seems to contradict this view: established techniques such as using a stopword list, a stemming procedure or blind query expansion, amongst others, also appear to be effective with web pages. In this section, we will push this analysis further by comparing the 10 search models on both the web and on TREC-8 corpora. This test collection contains documents extracted from the Los Angeles Times, the Financial Times, the Federal Register and the Foreign Broadcast Information Service, representing around 2 GB of text (see Appendix C for details about this corpus).

Indeed, this comparison is reliable because we use the same set of requests (TREC Topics #351–#450), with relevance assessments having been done by the same person for both corpora. From a comparison of Table 1 and Appendix C, one can see that the mean number of relevant items par request is roughly equal (88.68 for the web, 94.02 for TREC-8). We preprocessed the TREC-8 collection with the same stopword list and stemming procedure, and we applied the same indexing and retrieval schemes. Finally, the same evaluation methodology was used for both test collections.

Table 10 shows the average precision obtained with our 100 short requests (Title only), on the web and on the TREC-8 collections. From this data, one can see that the three best search strategies (Okapi, "doc = atn, query = ntc" and "doc = Lnu, query = ltc") are ranked in the same order under both corpora.

The linear correlation coefficient ρ between the average precision obtained with each search strategy on both collections is depicted in the last column of Table 10. The smallest correlation values were obtained with two poor retrieval strategies ("doc = nnn, query = nnn" and "doc = bnn, query = bnn"). This is not very surprising since these approaches tend to behave more randomly. On the other hand, for the three best strategies, the correlation value is of 0.63. For all these correlation coefficients, the null hypothesis H_0 stating that $\rho = 0$ (the average precision is not correlated) must be rejected (two-sided test, significance level of 0.05).

In order to better analyze the web vs. TREC-8 performance, we performed a query by query comparison using the Okapi probabilistic model. Each point on Fig. 2 corresponds to the average

Table 10
Average precision of the web and TREC-8 collections

Model\query (Title only)	Average precision		
	Web	TREC-8	Correlation ρ
doc = Okapi query = npn	26.68	24.47	0.608
doc = Lnu, query = ltc	23.38	21.34	0.591
doc = atn, query = ntc	25.67	22.17	0.695
doc = ntc, query = ntc	13.85	11.96	0.610
doc = ltc, query = ltc	13.67	14.36	0.692
doc = lnc, query = ltc	10.72	15.07	0.511
doc = lnc, query = lnc	7.24	10.07	0.446
doc = anc, query = ltc	8.23	13.50	0.519
doc = nnn, query = nnn	7.12	2.33	0.231
doc = bnn, query = bnn	9.56	7.11	0.373

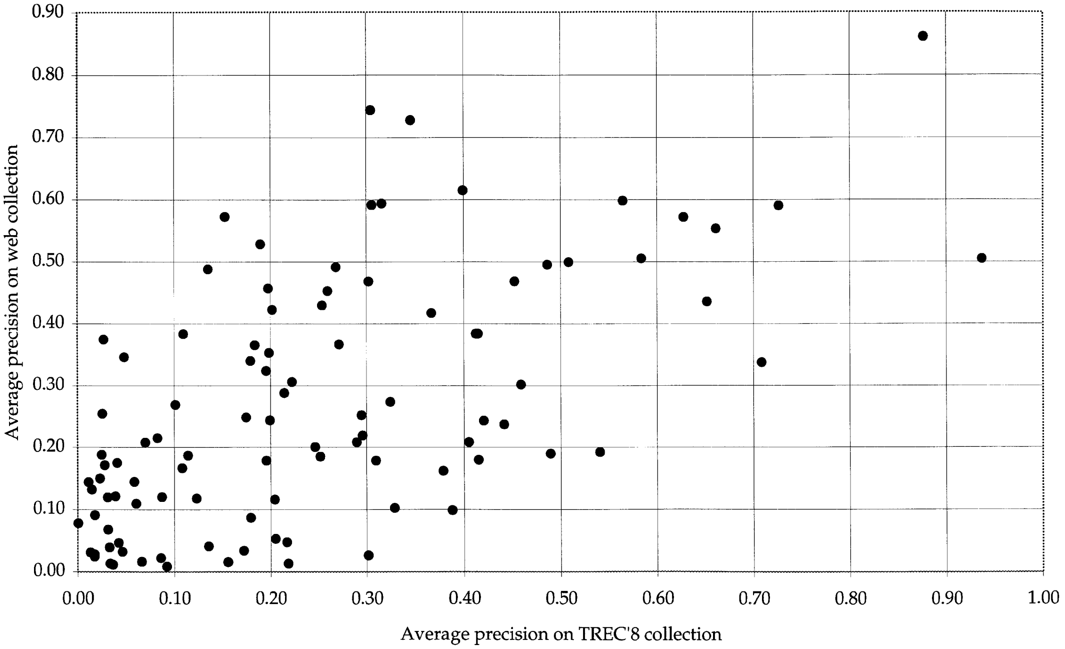


Fig. 2. Average precision obtained from web vs. TREC-8 collections, based on Okapi search strategy.

precision of one request obtained from the TREC-8 and web corpora. This graph depicts the general trend ($\rho = 0.608$) when comparing performances obtained from both collections.

As a conclusion, we may infer that search strategies having a good overall retrieval effectiveness on classical test collections will also produce good average precision on the web. However, the exact value of this performance cannot be directly known due to the subjective nature of relevance (Saracevic, 1975) and to the fact that some inherent characteristics of a test collection, often unknown, may favor one retrieval scheme to the detriment of another.

So far, our IR models do not take hyperlinks into account when attempting to improve their retrieval effectiveness. The question that then arises is; How can a retrieval scheme take this additional source of information into account and in order to enhance its performance? The next chapter will examine this question.

3. Hyperlinks and retrieval effectiveness

Relationships between web pages may play an important role in retrieving relevant information. As an analogy, we know that lawyers justify their conclusions by using precedential reasoning. In order to prepare a speech for the defense (in common law), a precedent is cited as authority in drawing a legal conclusion in favor of a client's argument. Thus it helps to know which other rulings might favor that of the defendant, and consequently which previous cases may have been affirmed, questioned, refined, modified, or overruled by another court.

This concept has been used in the IR domain to identify relationships between scientific papers by using bibliographic reference information (Garfield, 1983). Moreover, based on bibliographic links, the system may automatically derive other relationships such as bibliographic coupling similarity (Kessler, 1963) and co-citation relationships (Small, 1973).

In this chapter we will compare the relative merits of bibliographic references and hyperlinks. In Section 3.1, we will list the advantages and limitations of bibliographic references. Search models using bibliographic or hyperlinks will be described in Sections 3.2 and 3.3 will describe techniques used to evaluate the quality of hyperlinks, and to hopefully improve our search model's performance. Finally, Section 3.4 evaluates the retrieval effectiveness of some of these search models.

3.1. Direct citation indexing

Citation indexing (Garfield, 1983; Liu, 1993) is based on the bibliographic references contained in a document and it is already being used in libraries and library studies (Egghe & Rousseau, 1990). The underlying hypothesis of this indexing scheme is that bibliographic references give credit to related works that have clearly influenced the current paper. There are various advantages to this approach. First, citation indexing is independent of words and language, and thus may remove the underlying ambiguity of all natural languages. For example, the majority of subject indexers are specialists in a given subject, and documents may contain information related to more than one specific domain of knowledge. Therefore, with the use of citation information, documents having inappropriate indexing keywords can still be retrieved. Second, the words or phrases describing the semantic content of a document are vulnerable to scientific and technological obsolescence. Third, citation indexing is more easily managed by computers because, being very precise and following strict given patterns, such as the URL syntax used in our case (URLs may however appear in their relative forms).

Finally, citation schemes add new and unique capabilities. For example, when using citation systems in science, we may easily see whether the basic concept described has been applied elsewhere, or whether the underlying theory has been confirmed, improved or contradicted, or if a new overview exists for some old compound (Weinstock, 1971).

In the case of bibliographic references however research articles do not always cite other pertinent documents. A study by Cleverdon, Mills and Keen (1966) reveals that 36% of cited articles were judged by the authors as not relevant, and this rate increases to 52% if we include articles judged marginally relevant.

> It may be concluded that about half the references in an author's paper are not included in connection with the main problem of the paper, a fact which may assist examination of the possibilities and limitations, of the bibliographic coupling and citation indexing. (Cleverdon, Mills, & Keen, 1966, Volume I, p. 30).

Other explanations may be formulated. When questioning the authors, Cleverdon found the following remarks: "[these citations were included] to relate this report to my own previous work", "included to amplify certain details in the text … to save time and words in this report", "was included merely in order to satisfy anyone who wanted a long list", or "[this reference] did not come to hand until after the work was completed and the report nearly so".

Liu (1993) corroborates such findings and concludes that the motivations behind citations are more complex (in addition to the author's point of view, there are those of the editors and referees). Thus, relating citation to content analysis is not obvious and various motivations may invalidate the underlying hypothesis of citation.

3.2. Related work

When trying to exploit bibliographic reference information and other inter- or intra-document relationships, the first general approach was to extend the document surrogate by including index terms provided by related papers (Kwok, 1988; Turtle & Croft, 1991). However, the integration of these new terms represents weak information while original terms supply more valuable information.

A second principle is described in the work of Fox, Nunn and Lee (1988), in which an extended vector-processing model is presented that incorporates evidence from indexing terms, external attributes (e.g., author's name), and various citation schemes. While indexing terms are among the most important predictors of relevance, they also found that considering various bibliographic relationships might increase the system's performance.

As a third approach, Michelson, Amreich, Grissom and Ide (1971) compared relevance feedback based only on terms vs. relevance feedback using only authors, cited authors and citations. Michelson's study tends to confirm the usefulness of bibliographic reference relationships for relevance feedback.

As a fourth scheme various authors have suggested IR models that work in two stages. First, the retrieval engine defines a ranked list of records according to their similarity with the request, using only indexing terms. In the second stage, the retrieval system takes into account the incoming or outgoing hyperlinks related to the first m best-ranked documents (with m varying from 5 to 200). Thus the system does not have to analyze all links but only a few ones, reducing search space to a large extent. In this second stage, (Croft & Thompson, 1987; Frisse, 1988; Savoy, 1992, 1994; Frei & Stieger, 1995; Savoy, 1997a; Marchiori, 1997; Crestani & Lee, 2000) suggest using the spreading activation approach (Cohen & Kjeldsen, 1987) to propagate a fraction representing the best-ranked records' similarity to their linked documents.

Finally, recent works in IR on the web seem to acknowledge that hyperlink structures can be very valuable in locating information, and Kleinberg's paper (1998) explains how we can compute, for each web page, a hub and authoritative score. In this scheme, a web page that points to many others must be viewed as a "good" hub, and a document with many web pages pointing to it is a "good" authority. By transitivity, a document that points to many "good" authorities is an even better hub while a web page pointed to by many "good" hubs is an even better authority. Underlying this computation, the hypothesis states that documents with high authority scores contain relevant content, whereas web pages having high hub scores are assumed to point to relevant web pages, leading to an indirect use of the citation scheme for improving search algorithms.

Bharat and Henzinger (1998) indicate that Kleinberg's approach (1998) may work well for some queries and poorly for several other requests. They found three problems, namely:

(a) mutual reinforcement between hosts (e.g., when a set of pages on one host points to a single page on a second host); (b) automatically generated links which do not express human judgment, and (c) well connected non-relevant nodes which may cause the search to drift in the wrong direction.

Brin and Page (1998) suggest a search model that produces highly precise results, using the PageRank scheme based on the citation importance of retrieved documents. As for the spreading activation approach, the PageRank algorithm reranked the retrieved pages. In this case, a web page will have a higher score if many web pages point to it, or if some highly scoring documents point to it.

Dean and Henzinger (1999) suggest using a variation of Kleinberg's HITS algorithm (1998), where the similarity, relative to the content of both web pages and queries, is taken into account. The authors suggest exploiting the order of the links in the page, favoring those that precede the link to the given web page.

However, as mentioned by Marchiori (1997), treating the visibility of a web object as being synonymous to its popularity may be questionable, since the visibility of a given web page is not necessarily equivalent to either its informative content or its quality.

3.3. Relationships between hyperlinks and relevance

The hypothesis underlying our experiments is that hyperlinks may contain a certain amount of information about relevance. Before starting experiments, it is therefore advisable to have a better understanding of how and to what degree links are sources of evidence about relevance. This can be enlightening and can help in determining which techniques or parameter values better fit the particular situation at hand.

Our main purpose for using hyperlinks is that they may very well propagate some score or probability. But when should a link propagate information to other documents? Clearly if the document is not relevant, the link will not tell us much about the linked documents. However if it is relevant, one should expect that there is some probability that the linked documents will also be relevant, or in other words that the link is "valid". Obviously, the higher the probability of document relevance, the greater the link's information about relevance. It might therefore prove worthwhile to estimate this probability in order to get an idea of what can (and cannot) be expected from links, and eventually to use this probability estimate in our experiments.

A possible technique for estimating this link probability is the following. Based on a set of queries along with their relevance assessments, we compute for each relevant document the fraction of linked documents that are themselves relevant, and then we compute the average of this fraction for all queries (Algorithm 1). An objection to this method is that some documents are linked to more than one relevant document, and thus will have a higher probability of being relevant. To avoid an optimistically biased estimate, we exclude these documents from the computation, and compute the probability in the same way as Algorithm 1 (Algorithm 2). Finally, the link probability might vary largely between queries, mostly because the number of relevant documents can also vary by one or even two orders of magnitude (min: 6 relevant items; max: 828 relevant web pages, see Table 1). In order to keep a few queries from

Table 11
Probability estimation of hyperlinks

Estimation method	Incoming links	Outgoing links
Algorithm 1	0.145	0.106
Algorithm 2	0.066	0.090
Algorithm 3	0.062	0.051

dominating the computation, we take Algorithm 2 but compute the median instead of the mean (Algorithm 3).

From Table 11, one can find that depending on the algorithm used, the estimate may vary greatly. The experiments presented below make direct use of this probability, and work better for the smallest estimates found with Algorithm 3. This finding suggests that this value is a better estimate of the link's probability. It is lower than equivalent estimates found with the CACM collection (based on bibliographic references rather than hyperlinks, Picard, 1998). This seems to suggest that a web hyperlink might convey less information about relevance than does a bibliographic reference.

3.4. Evaluation of hyperlinks

Citation indexing seems useful and various previous research projects demonstrate that retrieval effectiveness may be improved when dealing with bibliographically based information. However, this improvement cannot be considered as being very important (in the range of 5–10%).

Spreading activation techniques, probabilistic argumentation systems (PAS) and an adaptation of Kleinberg's algorithm for computing hub and authority scores were included, along with each of the weighting schemes used so far, except for the latter technique. For spreading activation and PAS, we only considered links from/to the 50 best-ranked documents that are the 50 "strongest" sources of evidence about relevance. We took the initial rank and score for each document and computed a retrieval status value (spreading activation) or a degree of support (PAS), after the integration of the link information. Documents were then reranked according to this new score/degree of support.

3.4.1. Spreading activation

We first experimented with the simple technique of spreading activation. In that method, the degree of match between a document D_i and a query, as initially computed by the IR system (denoted $\mathrm{SIM}(D_i, Q)$), is propagated to the linked documents through a certain number of cycles using a propagation factor. We used a simplified version with only one cycle and a fixed propagation factor λ for all links. In that case, the final retrieval status value of a document D_i linked to m documents is

$$\mathrm{RSV}(D_i) = \mathrm{SIM}(D_i, Q) + \lambda \cdot \sum_{j=1}^{m} \mathrm{SIM}(D_j, Q).$$

Table 12
Average precision with spreading activation, best-ranked neighbor

Model	Average precision (% change)			
	Baseline	Best incoming link	Best outgoing link	Combined
doc = Okapi, query = npn	26.68	26.68 (0.0%)	26.68 (0.0%)	26.68 (0.0%)
doc = Lnu, query = ltc	23.38	23.52 (+0.60%)	23.49 (+0.47%)	23.64 (+1.11%)
doc = atn, query = ntc	25.67	25.67 (0.0%)	25.75 (+0.31%)	25.46 (−0.82%)
doc = ntc, query = ntc	13.85	13.85 (0.0%)	13.85 (0.0%)	13.68 (0.0%)
doc = ltc, query = ltc	13.67	13.69 (+0.15%)	13.76 (+0.66%)	13.77 (+0.73%)
doc = lnc, query = ltc	10.72	10.80 (+0.75%)	10.89 (+1.59%)	10.81 (+0.84%)
doc = lnc, query = lnc	7.24	7.34 (+1.38%)	7.43 (+2.62%)	7.38 (+1.93%)
doc = anc, query = ltc	8.23	8.30 (+0.85%)	8.62 (+4.74%)	8.43 (+2.43%)
doc = nnn, query = nnn	7.12	7.12 (0.0%)	7.12 (0.0%)	7.12 (0.0%)
doc = bnn, query = bnn	9.56	9.68 (+1.26%)	9.91 (+3.66%)	10.26 (+7.32%)

Using all the incoming and outgoing links separately, and for different values of the parameter λ, in most cases did not result in retrieval improvement. This tends to show that simple and intuitive techniques, which have produced satisfactory results in other retrieval environments, do not seem to perform well in this situation. It is our opinion that hyperlinks seem to provide less information than do the bibliographic references or co-citation schemes used in our previous studies.

We have however obtained better results in some cases by considering not all the 50 best-ranked documents but only the best-ranked one. The reason for choosing the best source of evidence is that when a document is already linked to one of the best-ranked documents, the other linked documents only have a marginal effect on its relevance probability.

The results shown in Table 12 were computed using the probability estimates of 0.06 and 0.05 (Algorithm 3 in Table 11), for respectively the incoming and outgoing links. The last column of Table 12 shows the results, as obtained using both the best incoming and the best outgoing document.

3.4.2. Probabilistic argumentation systems

In a second set of experiments, we used PAS for which documents are ranked by a decreasing degree of support (Picard, 1998). For this study, we used a simplified version of our approach whereby a document's degree of support can be affected only by its direct neighbors. In that case we do not need to keep track of inferences, and can derive a simple formula which can be understood as a more refined way of spreading activation. Instead of propagating a document's score, we propagated its probability of being relevant. This probability was multiplied by the probability of the link, denoted $p(\text{link})$, and then assessed according to Table 11. To compute the relevance probability of a document D_i given its rank $p(D_i \mid \text{rank})$, we fitted a logistic regression (Bookstein, O'Neil, Dillon, & Stephens, 1992) to its rank on the available topics (Le Calvé & Savoy, 2000). For the probabilities of the links, we used the probability estimates found with Algorithm 3 in Table 11.

The individual contribution of a document D_i is then $[p(D_i \mid \text{rank}) \cdot p(\text{link})]$, instead of $[\text{SIM}(D_i, Q) \cdot \lambda]$ used with the spreading activation technique. Just as we did for spreading

Table 13
Average precision with PAS, two best-ranked neighbors

Model	Average precision (% change)			
	Baseline	Best incoming link	Best outgoing link	Combined
doc = Okapi, query = npn	26.68	26.68 (0.0%)	26.73 (+0.18%)	26.68 (0.0%)
doc = Lnu, query = ltc	23.38	23.89 (+2.18%)	24.00 (+2.65%)	23.87 (+2.10%)
doc = atn, query = ntc	25.67	26.04 (+1.44%)	26.06 (+1.52%)	26.01 (+1.32%)
doc = ntc, query = ntc	13.85	13.87 (+0.14%)	13.87 (+0.14%)	13.85 (0.00%)
doc = ltc, query = ltc	13.67	13.79 (+0.88%)	13.79 (+0.88%)	13.87 (+0.88%)
doc = lnc, query = ltc	10.72	10.79 (+0.65%)	10.88 (+1.49%)	10.86 (+1.31%)
doc = lnc, query = lnc	7.24	7.41 (+2.35%)	7.34 (+1.38%)	7.35 (+1.52%)
doc = anc, query = ltc	8.23	8.42 (+2.31%)	8.69 (+5.59%)	8.46 (+2.79%)
doc = nnn, query = nnn	7.12	7.16 (+0.56%)	7.15 (+0.42%)	7.05 (−0.98%)
doc = bnn, query = bnn	9.56	9.96 (+4.18%)	10.04 (+5.02%)	9.90 (+3.56%)

activation experiments, using all incoming or outgoing links did not demonstrate any improvement, except in some cases. We then decided to include only the most important sources of evidence, the same way as for spreading activation. For example, we considered the initial rank of document D_i, the best incoming document D_{in} and the best outgoing document D_{out}. Taking $p(\text{link}_{\text{in}})$ and $p(\text{link}_{\text{out}})$ as the probability of incoming and outgoing links, document D_i has the following degree of support:

$$\text{DSP}(D_i) = 1 - (1 - p(D_i \mid \text{rank})) \cdot [1 - p(D_{\text{in}} \mid \text{rank}) \cdot p(\text{link}_{\text{in}})] \cdot [1 - p(D_{\text{out}} \mid \text{rank}) \cdot p(\text{link}_{\text{out}})].$$

Table 13 shows the results obtained using only the best incoming document, the best outgoing document, and both.

The experiments shown here presented pessimistic results regarding the potential use of hyperlinks for information retrieval. In general, the increase in precision obtained after integration of the hyperlinks was very small (not more than a few percent), except for the "doc = bnn, query = bnn" search model (up to 7.32% in Table 12). Although in this collection there were only a few links between pages with different URLs, this is not so different from the web: Bray (1996) reports that 80% of web pages have no outgoing link to a different URL. Moreover, considering that $p(\text{link})$ is around 0.05 for incoming and outgoing links, this means that on average only one document in twenty linked to a relevant document will itself be relevant. This poor overall retrieval effectiveness is confirmed by Gordon and Pathak's study (1999). If nothing else is known about a document, this can be useful information. However, it is questionable whether this information may significantly help a good retrieval system such as Okapi, where on average, already 45% of the best-ten ranked documents are relevant.

3.4.3. Computing hub and authority scores

We applied a weighted version of Kleinberg's algorithm as suggested by Bharat and Henzinger (1998), where the hub and authority scores of a document propagated to linked documents are weighted by the initial score of the document, such that documents more likely to be relevant have

Table 14
Average precision obtained after application of a weighted form of Kleinberg's algorithm

Model	Average precision	
	Baseline	With links
doc = Okapi, query = npn	26.68	8.74
doc = atn, query = ntc	25.67	9.25
doc = ntc, query = ntc	13.85	6.82

a stronger effect on their neighbors. The updating formulas for the hub and authority scores $H^c(D_j)$ and $A^c(D_j)$ of document D_i after c iterations are:

$$A^{c+1}(D_i) = \sum_{D_j=\text{parent}(D_i)} H^c(D_j), \quad H^{c+1}(D_i) = \sum_{D_j=\text{child}(D_i)} A^c(D_j)$$

which is computed for 200 best-ranked documents retrieved by a classical search model, together with their children and parents. The hub and authority scores were updated for five iterations (while the ranking did not change after this point), and a normalization procedure was applied after each step. Table 14 shows the precision obtained for three of the most important weighting schemes: the Okapi probabilistic model, "doc = atn, query = ntc" which is the best vector-space weighting scheme, and "doc = ntc, query = ntc" which is widely known as the tf–idf weighting formula.

4. Conclusion

Convinced that isolated retrieval effectiveness evaluations are not very useful, we have carried out different experiments based on various search strategies within the web test collection. These experiments show that:

- indexing strategies based on a simple list of keywords ("doc = bnn, query = bnn") result in better retrieval performance than those simply taking occurrence frequencies into account ("doc = nnn, query = nnn", Table 2);
- for these corpora, traditional indexing strategies based on within-document term frequency, document-wide term frequency and normalization ("doc = ntc", or "doc = ltc") provide significantly lower average precision than an indexing approach based only on term frequency and collection frequency ("doc = atn", Table 2);
- IR models performing well on TREC-8 corpus also seem to perform well on the web corpus (Table 10);
- adding search keywords to the query may significantly enhance average precision (Table 3);
- adding weights to terms appearing in the Title or H1 logical sections has no significant effect on average precision (Table 6);
- limiting indexing to the Title or Title and H1 sections results in poor retrieval effectiveness (Table 7);

- ignoring a stemming procedure (the Lovins' approach in our case) significantly degrades average precision for most retrieval strategies (Table 8);
- using a stoplist may significantly enhance retrieval effectiveness (Table 8);
- blind query expansions significantly improve average precision but introduces additional and non-negligible response times (Table 9);
- taking hyperlinks into account may improve the average precision, but the variation is not significant when using the spreading activation or PAS approach based on good retrieval schemes (Tables 12 and 13).

Our evaluation methodology is based on the ability of indexing and searching algorithms to find individual pages that may not always represent the best starting points for browsing. In this case, it might be more appropriate to define retrieval mechanisms that search for good sites or hubs, as those described in Kleinberg's paper (1998). Hyperlinks seem however to provide a very useful information for extracting patterns representing various cyber communities or sets of authoritative pages relative to broadly represented topics (Chakrabarti, Van den Berg & Dom, 1999; Small, 1999; Wasserman & Faust, 1997). Finally, we recognize that most web users want a high precision and are ready to accept a lower recall value when the computer system answers quickly and when using a search engine that can be viewed as intelligent.

Acknowledgements

The authors would like to thank D. Hawking for building web test collections and C. Buckley from SabIR for giving us the opportunity to use the SMART system, without which this study could not have been conducted. This research was supported by the SNF (Swiss National Science Foundation) under grant 20-50'578.97 and 20'55'771.98).

Appendix A. Weighting schemes

To assign an indexing weight w_{ij} that reflects the importance of each single-term T_j in a web page D_i, we may take three different factors into account. They are represented by the three code letters respectively:

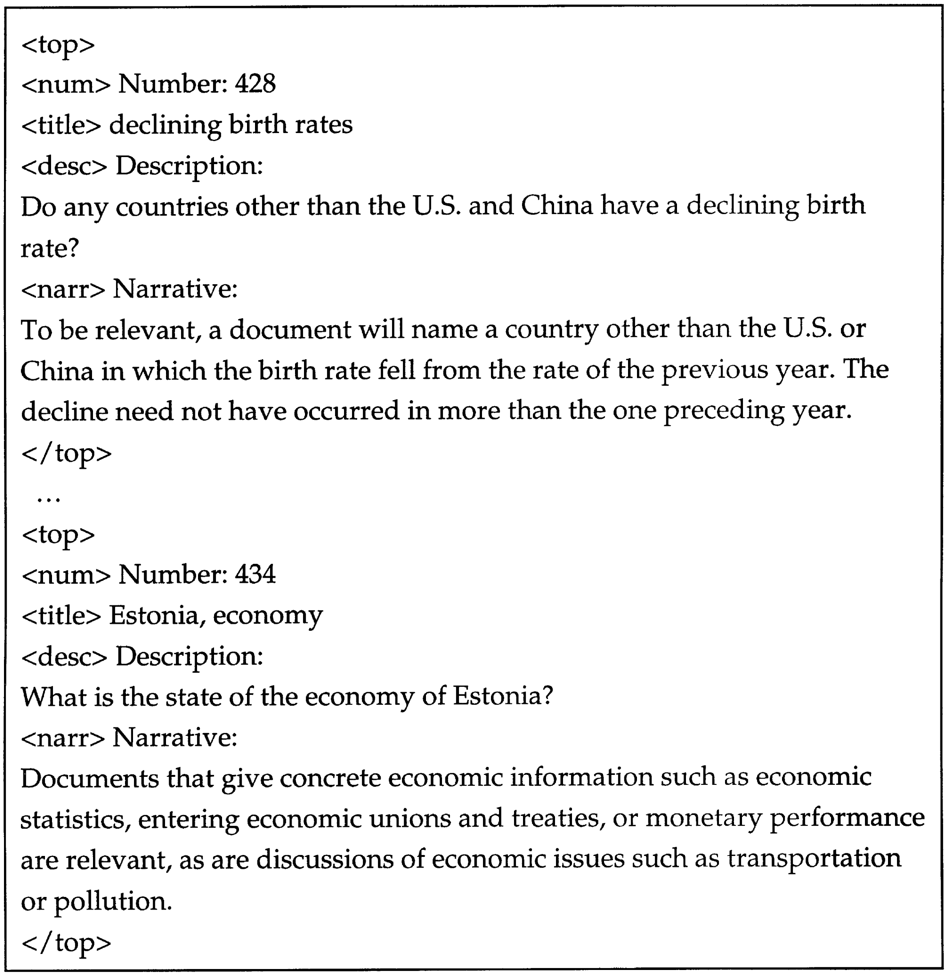
- the within-document term frequency, noted tf_{ij} (first letter);
- the collection-wide term frequency, noted df_j (second letter);
- the normalization scheme (third letter).

In Table 15, the document length (the number of indexing terms) of D_i is noted by nt_i, the constant advl is fixed at 900, the constant b at 0.9, the constant k_1 at 2, the constant pivot at 125 and the constant slope at 0.1. Finally, the Okapi weighting scheme for document corresponds to:

$$w_{ij} = (k_1 + 1) \cdot [\mathrm{tf}_{ij}/(K + \mathrm{tf}_{ij})] \quad \text{with } K = k_1 \cdot [(1 - b) + b \cdot (l_i/\mathrm{avdl})]$$

Table 15
Weighting schemes

n	new_tf = tf_{ij} (occurrence frequency of T_j in the document D_i)
b	new_tf = binary weight (0 or 1)
a	new_tf = $0.5 + 0.5 \cdot (\mathrm{tf}_{ij}/\max \mathrm{tf} \text{ in } D_i)$
l	new_tf = $\ln(\mathrm{tf}_{ij}) + 1.0$
L	new_tf = $[\ln(\mathrm{tf}_{ij}) + 1.0]/[1.0 + \ln(\text{mean } (\mathrm{tf} \text{ in } D_i))]$
n	new_wt = new_tf (no conversion is to be done)
t	new_wt = new_tf $\cdot \ln[N/\mathrm{df}_j]$
p	new_wt = new_tf $\cdot \ln[(\mathrm{N} - \mathrm{df}_j)/\mathrm{df}_j]$
n	w_{ij} = new_wt (no conversion is to be done)
c	divide each new_wt by sqrt (sum of (new_wts squared)) to get w_{ij}
u	w_{ij} = new_wt/$[(1 - c) \cdot \mathrm{mean}(nt) + c \cdot nt_i]$



<top>
<num> Number: 428
<title> declining birth rates
<desc> Description:
Do any countries other than the U.S. and China have a declining birth rate?
<narr> Narrative:
To be relevant, a document will name a country other than the U.S. or China in which the birth rate fell from the rate of the previous year. The decline need not have occurred in more than the one preceding year.
</top>
...
<top>
<num> Number: 434
<title> Estonia, economy
<desc> Description:
What is the state of the economy of Estonia?
<narr> Narrative:
Documents that give concrete economic information such as economic statistics, entering economic unions and treaties, or monetary performance are relevant, as are discussions of economic issues such as transportation or pollution.
</top>

Fig. 3. Query examples.

within which K represents the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean noted by advl.

Appendix B. Query examples

See Fig. 3.

Appendix C. TREC-8 test collection

See Table 16.

Table 16
TREC-8 test collection statistics

Size (in MB)	1904 MB
# of documents	528,155
# of distinct indexing terms	1,008,463
# of queries	100
# of distinct index terms/document	
Mean	136.84
Standard error	114.54
Median	108
Maximum	23,515
Minimum	2
# of indexing terms/document	
Mean	240.89
Standard error	501.35
Median	171
Maximum	211,934
Minimum	2
# of relevant documents (100 queries)	9402
Mean	94.02
Standard error	82.247
Median	68.5
Maximum (Query #354)	361
Minimum (Query #430)	6

References

Alschuler, L. (1989). Hand-crafted hypertext – Lessons from the ACM experiment. In E. Barrett, *The society of text, hypertext, hypermedia, and the social construction of information* (pp. 343–361). Cambridge, MA: MIT Press.

Bailey, P., Craswell, N., & Hawking, D. (2000). Dark matter on the web. In *Poster-Proceedings of WWW9* (pp. 60–61). Amsterdam: Elsevier.

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of ACM-SIGIR'98* (pp. 104–111), New York: ACM.

Bray, T. (1996). Measuring the web. In *Proceedings of WWW5*. Amsterdam: Elsevier.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW7* (pp. 107–117). Amsterdam: Elsevier.

Bookstein, A., O'Neil, E., Dillon, M., & Stephens, D. (1992). Applications of loglinear models for informetric phenomena. *Information Processing & Management*, *28*(1), 75–88.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART. In *Proceedings of TREC'4* (pp. 25–48). NIST Publication #500-236.

Chakrabarti, S., Van den Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. In *Proceedings of WWW8* (pp. 545–562). Amsterdam: Elsevier.

Cleverdon, C. W., Mills, J., & Keen, M. (1966). Factors determining the performance of indexing systems. Cranfield, UK: ASLIB Cranfield Research Project.

Cohen, P. R., & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, *23*(4), 255–268.

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

Croft, W. C., & Thompson, R. H. (1987). I3R: a new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, *38*(6), 389–404.

Dean, J., & Henzinger, M. R. (1999). Finding related pages in the world wide web. In *Proceedings of WWW8* (pp. 389–401). Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Fox, E. A., Nunn, G. L., & Lee, W. C. (1988). Coefficients for combining concept classes in a collection. In *Proceedings of ACM-SIGIR'88* (pp. 291–307). New York: ACM.

Frei, H. P., & Stieger, D. (1995). The use of semantic links in hypertext information retrieval. *Information Processing & Management*, *31*, 1–13.

Frisse, M. E. (1988). Searching for information in a hypertext medical handbook. *Communications of the ACM*, *31*(7), 880–886.

Furnas, G., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971.

Garfield, E. (1983). *Citation indexing: its theory and application in science, technology and humanities* (2nd ed.). Philadelphia: ISI Press.

Gordon, M., & Pathak, P. (1999). Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing & Management*, *35*(2), 141–180.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Hawking, D., Craswell, N., Thistlewaite, P., & Harman, D. (1999a). Results and challenges in web search evaluation. In *Proceedings of WWW8* (pp. 243–252). Amsterdam: Elsevier.

Hawking, D., Voorhees, E., Bailey, P., & Craswell, N. (1999b). Overview of TREC-8 web track. In *Proceedings of TREC'8*. NIST Publication.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of ACM-SIGIR'93* (pp. 329–338). New York: ACM.

Jansen, B. J., 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.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of ninth ACM-SIAM Symposium on Discrete Algorithms* (pp. 668–677). New York: ACM.

Kwok, K. L. (1988). On the use of bibliographically related titles for the enhancement of document representations. *Information Processing & Management*, *24*, 123–131.

Lawrence, S., & Giles, L. C. (1999). Accessibility of information on the web. *Nature*, *400*(6740), 107–110.

Le Calvé, A., & Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing & Management*, *36*(3), 341–359.

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. (1997). *Practical digital libraries: books, bytes, and bucks*. San Francisco: Morgan Kaufmann.

Lesk, M. (1998). "Real world" searching panel at SIGIR 97. *ACM-SIGIR Forum,* 32 (1), 1–4.

Liu, M. (1993). The complexities of citation practice: a review of citation studies. *Journal of Documentation*, *49*(4), 370–408.

Lovins, J. B. (1982). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*(1), 22–31.

Marchiori, M. (1997). The quest for correct information on the web: hyper search engines. In *Proceedings of WWW6*. Amsterdam: Elsevier.

Michelson, D., Amreich, M., Grissom, G., & Ide, E. (1971). An experiment in the use of bibliographic data as a source of relevance feedback in information retrieval. In G. Salton, *The SMART retrieval system – Experiments in automatic document processing* (pp. 430–443). Englewood Cliffs, NJ: Prentice-Hall.

Picard, J. (1998). Modeling and combining evidence provided by document relationships using probability argumentation systems. In *Proceedings of ACM-SIGIR'98* (pp. 182–189). New York: ACM.

Raghavan, P., Broder, A., Henzinger, M. R., Manber, U., & Pinkerton, B. (1999). Finding anything in the billion page web: are algorithms the key. Panel abstract. In *Proceedings of WWW8* (pp. 682–683). Amsterdam: Elsevier.

Robertson, S. E., Maron, M. E., & Cooper, W. S. (1982). Probability of relevance: a unification of two competing models for document retrieval. *Information and Technology: Research & Development*, *1*, 1–21.

Robertson, S. E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, *36*(1), 95–108.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Salton, G. (1989). *Automatic text processing: the transformation analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Saracevic, T. (1975). Relevance: a review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*, 321–343.

Savoy, J. (1992). Bayesian inference networks and spreading activation in hypertext systems. *Information Processing & Management*, *28*(3), 389–406.

Savoy, J. (1994). A learning scheme for information retrieval in hypertext. *Information Processing & Management*, *30*(4), 515–533.

Savoy, J. (1996). An extended vector-processing scheme for searching information in hypertext system. *Information Processing & Management*, *32*(2), 155–170.

Savoy, J. (1997a). Ranking schemes in hybrid Boolean systems: a new approach. *Journal of the American Society for Information Science*, *48*(3), 235–253.

Savoy, J. (1997b). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, *33*(4), 495–512.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, *50*(10), 944–952.

Schwartz, C. (1998). Web search engines. *Journal of the American Society for Information Science*, *49*(11), 973–982.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*(9), 799–813.

Sparck Jones, K., & Bates, R.G. (1977). Research on automatic indexing 1974–1976, Technical Report. University of Cambridge, UK: Computer Laboratory.

Spink, A., & Saracevic, T. (1997). Interactive information retrieval: sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science*, *48*(8), 741–761.

Tague-Sutcliffe, J., & Blustein, J. (1995). A statistical analysis of the TREC-3 data. In *Proceedings of TREC'3* (pp. 385–398). NIST Publication #500-225.

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, *9*, 187–222.

Voorhees, E. M., & Harman, D. (2000). Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing & Management*, *36*(1), 3–35.

Wasserman, S., & Faust, K. (1997). *Social network analysis: methods and applications*. Cambridge: Cambridge University Press.

Weinstock, M. (1971). Citation indexes. In *Encyclopedia of library and information science* (pp. 16–40). New York: Marcel Dekker.