Report on the TREC-8 Experiment: Searching on the Web and in Distributed Collections

Jacques Savoy, Justin Picard

Institut interfacultaire d'informatique
Université de Neuchâtel (Switzerland)
e-mail: {Jacques.Savoy, Justin.Picard}@seco.unine.ch
Web page: http://www-seco.unine.ch/info

Summary

The Internet paradigm permits information searches to be made across wide-area networks where information is contained in web pages and/or whole document collections such as digital libraries. These new distributed information environments reveal new and challenging problems for the IR community. Consequently, in this TREC experiment we investigated two questions related to information searches on the web or in digital libraries: (1) an analysis of the impact of hyperlinks in improving retrieval performance, and (2) a study of techniques useful in selecting more appropriate text databases (database selection problem encountered when faced with multiple collections), including an evaluation of certain merging strategies effective in producing, single, ranked lists to be presented to the user (database merging problem).

Introduction

There is an increasing interest in hypertext systems, digital libraries and in effective web searching [Bernes-Lee 94]. Due to the huge number of pages and links, browsing cannot be viewed as an adequate searching process, even with the introduction of tables of contents or other classified lists (e.g., Yahoo!) [Alschuler 89]. As a result, effective query-based mechanisms for accessing information will always be needed [Halasz 88]. The search engines currently available on the web [Leighton 99], [Gordon 99] are hardly able to cover all available information [Lawrence 99] and they are characterized by many drawbacks [Hawking 99a]. Moreover, in order to enhance their retrieval effectiveness, most of them ignore hypertext links. Recent works in IR on the web seem to acknowledge that hyperlink structures can be very valuable in locating information [Marchiori 97], [Kleinberg 98], [Brin 98], [Chakrabarti 99], [Bharat 98]; and according to Chakrabarti et al. [99]:

> "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 99, p. 550-551]

With small variations, similar hypotheses are also cited by other authors [Kleinberg 98], [Bharat 98]. Our previous studies on citation schemes [Savoy 94], [Savoy 96a], [Savoy 97], [Picard 98] tend to suggest however that citation information might improve average precision, but only on the order of 5% to 8% when used with good retrieval schemes.

The first chapter of this paper verifies whether or not hyperlinks improve retrieval effectiveness. In the second chapter, we describe experiments on the ad hoc track. In this case, we acknowledge that currently it is becoming more and more difficult to store and manage the growing document collections within a single computer. Recent advances in network technology do however allow us to disseminate information sources by partitioning a single huge corpus (or distributing heterogeneous collections) over a local-area network (Intranet). Most retrieval mechanisms currently proposed however are based on conventional IR models [Salton 89], and where a centralized document collection is assumed.

To access these distributed collections, our IR model sends a request to several separate and selected text databases (each having its own search engine) on the one hand, and on the other, it implies merging of the resultant output lists in the form of an "optimal" single list to be presented to the user. Thus, our approach must address the following problems [Dreilinger 97]:

- selecting the appropriate set of information servers to which the query will be sent (collection selection problem);
- converting the information need into a format readable by the selected search engines (e.g., based on the Z90.50 protocol for inter-system retrieval [Kahle 93], or on STARTS model [Gravano 97]);
- selecting and sorting the result lists obtained by different information servers to form a unique result list (database merging problem).

Chapter two of this paper reflects our interest in addressing the first and last problems, both of which may be viewed as serious. To evaluate our hypothesis, we used the SMART system as a test bed for implementing the various vector-processing weighting schemes along with the OKAPI probabilistic model [Robertson 95]. This year our experiments were conducted on an Intel Pentium III/450 (cache: 1MB, memory: 256 MB, disk: RAID0 EIDE with 2 x 27 GB) and all of our experiments are fully automated.

1. Small Web Track

Our participation with the web track addresses the following question: do hyperlinks provide useful evidence in enhancing a search engine's retrieval?

Some statistics describing the web collection are listed in Table 1 and other characteristics are described in [Hawking 99b]. Of note is that this corpus possesses 1,171,795 hyperlinks leading to an average of 4.73 hyperlinks per page (used primarily for navigational purpose across the web site). Relative to the Web which is currently estimated to contain about 800 million web pages [Lawrence 99], our test collection might be viewed as being relatively small. There is consequently the risk that a large portion of the hyperlinks between pages having different URLs (defined as the IP number) will be unusable, because the destination node will very likely be outside of the collection. According to our computations, there were 2,797 hyperlinks to pages on different hosts, representing 0.24% of the total. Moreover, most of these links were grouped in clusters (e.g., one or a few web pages from one site point to one or a few web pages from another site).

In order to proceed with our evaluation, we used the non-interpolated average precision at eleven recall values, based on 1,000 retrieved items per request. To determine whether or not a given search strategy is better than another, we need a decision rule. The following rule of thumb could provide serve as such a rule: a difference of at least 5% in average precision is generally considered significant and a 10% difference is considered material [Sparck Jones 77, p. A25].

From the original WWW pages, we retained the following logical sections: <title>, <h1>, <center>, <big> and for delimiting document boundaries: <docno>. Thus, the most common tags <P> (or <p>, together with </P>, </p>) have been removed. Text delimited by the tags <DOCHDR>, </DOCHDR> were also removed. A classical stemming procedure was applied and stopwords were removed.

1.1. Pseudo-Relevance Feedback

It is recognized that pseudo-relevance feedback (blind expansion) is a useful technique for enhancing retrieval effectiveness. Thus, we have evaluated the OKAPI search model with and without query expansion to verify whether or not this technique might improve retrieval performance when faced with different query formulations (such technique is known to be time-consuming). In this study, we have adopted Rocchio's approach [Buckley 96] with $\alpha = 0.75$, $\beta = 0.75$ and the system is allowed to add 17 search terms to the original query during feedback which are extracted from the 30-best ranked documents. The resulting retrieval effectiveness is depicted in Table 2a.

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). However Table 2b shows that retrieval time is significantly increased with procedure.

Size (in MB)	2,000 MB
# of web pages extracted from 969 URLs	247,491
# of distinct indexing terms in the collection	1,850,979
# of distinct index terms / web page	
mean	218.25
standard error	326.42
median	125
maximum	22722
minimum	1
# of indexing terms / web page	
mean	554.295
standard error	1402.86
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)	8,868
from Topic #351 to Topic #400	6,589
from Topic #401 to Topic #450	2,279

Table 1: Small Web Collection Statistics

	Precision (% change)		
Model \ Query	Title	Title & Desc	Title, Desc & Narr
doc = OKAPI, query = NPN	23.49	27.39	30.34
with query expansion	29.55 (+25.80%)	31.36 (+14.49%)	30.74 (+1.32%)

Table 2a: Average Precision of Blind Query Expansion

	Search Time in sec. (% change)		
search time (original) / request	0.3033	0.5279	0.8185
search time (expand) / request	4.570 (+1406%)	4.748 (+999%)	5.138 (+527%)

Table 2b: Search Time per Request (in sec.)

1.2. Hyperlinks

Based on our previous studies on citation schemes [Savoy 96a], [Savoy 97], [Picard 98], we have taken hyperlinks into account to hopefully improve retrieval effectiveness. The common point of spreading activation techniques [Cohen 87] used in our previous works [Savoy 96a], [Savoy 97] and the probabilistic argumentation systems (PAS) [Picard 98] used here is to consider links as a way of improving the initial ranking of documents.

Instead of directly trying to use the hyperlinks for retrieval, we believe it is better to understand how they relate to the relevance of a document, and to estimate to what degree this relationship holds (Section 1.2.1). Then we will apply the spreading activation technique and PAS to integrate these links into the retrieval process (Section 1.2.2). Finally we will draw some conclusions on the potentiality offered by links for retrieval on the web, in regard of the experimental results obtained (Section 1.2.3).

1.2.1. Relationships between Hyperlinks and Relevance

The hypothesis underlying our experiments is that hyperlinks contain some information about relevance. Before starting experiments, it is therefore certainly 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 better fit the particular situation at hand.

Our main idea in using hyperlinks is to consider tthat they may propagate some score or probability. But when should a link propagate information to other documents? Clearly if the document is not relevant, this 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 this probability, the greater the link's information about relevance. It would then be interesting to estimate this probability using a training set, in order to get an idea on what can (and cannot) be expected from links. For this purpose we used Topics #351 to #400.

A possible technique for estimating this link probability is the following. For each relevant document, we compute the fraction of linked documents that are themselves relevant, then we compute the average of this fraction on all queries (Algorithm 1). An objection to this method is that some documents are linked to more than one relevant document, and will have a higher probability of being relevant. To avoid an upward biased estimate, we exclude these documents from 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 vary by one or even two orders of magnitude. In order to keep a few queries from dominating the computation, we take Algorithm 2 but compute the median instead of the mean (Algorithm 3). The resulting probability estimates are given in Table 3.

From data depicted in Table 3, one can find that depending on the algorithm used, the estimate may vary greatly. The experiments presented in the next subsection make direct use of this probability, and work better for the smallest estimates found with Algorithm 3. This finding strongly suggests that this value is a better estimate of the link probability. It is lower than equivalent estimates found with the CACM collection (based on bibliographic references rather than hyperlinks).

Other experiments, which are not displayed here evaluated the impact on a document's probability of relevance, given that it is linked or not to one of the five best ranked documents. It seems that being linked to one of the five best ranked documents does not affect the probability of relevance for the 25 best-ranked documents, and increases it slightly for higher ranks. This result tends to confirm that hyperlinks should have a small impact on retrieval effectiveness.

1.2.2. Experiments

For the two techniques, we only considered only links from/to the 50 best-ranked documents. We took the initial rank and score of each document, and computed a retrieval status value (spreading activation) or a degree of support (PAS), after the integration of link information. Documents were then reranked according to this new score/degree of support.

We first experimented with the simple technique of spreading activation. In that method, the degree of match of a document and a query, as initially computed by the IR system (denoted s(D)), 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 of a certain type (incoming or outgoing). In that case, the final retrieval status value (RSV) of a document D linked to n documents is:

Estimation Method	Incoming Links	Outgoing Links
Algorithm 1	0.145	0.106
Algorithm 2	0.066	0.090
Algorithm 3	0.062	0.051

Table 3: Probability Estimates of Links for Our Three Algorithms

$$RSV(D) = s(D) + \lambda \cdot \sum_{i=1}^{n} s(D_i)$$

We experimented with several values of the propagation factor λ. Even for the smallest values of λ, a deterioration in retrieval effectiveness resulted, and this deterioration increased monotonically for increasing parameter values. This tends to show that simple and intuitive techniques, which 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.

In a second set of experiments, we used probabilistic argumentation systems (PAS) [Picard 98]. For this study, we used a simplified version of our approach where a document's degree of support (and thus its rank) 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(link), and then assessed according to Section 1.7.1. To compute the probability of relevance of a document given its rank p(D | rank), we fitted a logistic regression [Bookstein 92] to its rank for the set of training Topics #351 to #400.

The individual contribution of a document D_i is then $p(D_i \mid rank) \cdot p(link)$, instead of $s(D_i) \cdot \lambda$ used with the spreading activation technique. In the case where a document had more than one source of evidence indicating relevance, the spreading activation technique summed the individual contributions. In the PAS technique, the initial rank of a document and the contribution of each linked document were considered as different sources of evidence. A source of evidence e_i has a certain probability $p(e_i)$ to being valid, and the degree of support (DSP) of a document is computed as the probability that at least one of the source of evidence is valid.

$$dsp(D) = 1 - \prod_{i=1}^{n} (1 - (p(e_i))$$

Experiments using all incoming or outgoing links did not demonstrate any improvement. We then decided to include only the most important sources of evidence: the initial rank of the document D, the best incoming document D_{in} and the best outgoing document D_{out} .

$$\begin{aligned} dsp(D) = 1 - (1 - (p(D|rank)) \cdot \\ (1 - p(D_{in}|rank) \cdot p(link_{in})) \cdot \\ (1 - p(D_{out}|rank) \cdot p(link_{out})) \end{aligned}$$

For the values of $p(link_{in})$ and $p(link_{out})$ computed with Algorithm 3, we obtained improvements of between 1% to 1.5% for Topics #351 to #400. Other values of these probabilities did not yield higher retrieval effectiveness. The results obtained on Topics #401 to #450 are given in Table 4. However, hyperlinks may be valuable for other purposes; for example, citation information have been used to define co-citation clusters for better visualizing the relationships between disciplines, fields, specialties, and individuals papers [Small 99].

1.2.3. Official Web Runs

Our official run (UniNEW2Ct, content-only) resulted in an average precision of 31.50, 41 times above the median and for the two queries (#424, #434), it displays the best results. The related official run (UniNEW2Link, content & links) shows a small but not significant degradation in average performance.

Official Run Name	Average Precision	# ≥ Median	# Best
UniNEWCt	27.39	34	0
UniNEWLink	27.47 (+0.29%)	44	3
UniNEW2Ct	31.50	41	2
UniNEW2Link	31.37 (-0.41%)	44	9

Table 4: Summary of Our Official Runs for the Web Track

1.3. Summary of Results

The various experiments carried out within the web track showed that:

- Hyperlinks do not result in any significant improvement (at least as implemented in this study). Link information seems to be marginally useful for retrieval on the web, especially when the retrieval system produces relatively high retrieval effectiveness;
- Pseudo-relevant feedback techniques (blind query expansions) result in significant improvement but they increase search times (by a factor of ten in our implementation);

2. Distributed Collections

To evaluate the retrieval effectiveness of our distributed IR model, we formed four separate sub-collections according to the source of the available documents. Table 5 summarizes various statistics about these four sub-collections and depicts general statistics of the collection named TREC8.

In this study with our distributed IR model, we assumed that each search engine used the same indexing scheme and the same retrieval procedure. Such a distributed context reflects a local area network more closely than does the Internet where different search engines may collaborate to search for information [Le Calvé 99]. Our approach may be more closely identified by the following characteristics. In the first stage and based on the current query, our IR model must select the more appropriate set of sub-collections on which the search will be done (Section 2.2, see also [Callan 95], [Xu 98], [Fuhr 99], [Hawking 99a]). Based on this selection procedure, the query will be sent to the selected text databases and depending on the results, the system will merge them into a single result list to be presented to the user (Section 2.3).

Before describing the collection selection and the collection fusion approaches, Section 2.1 identifies retrieval effectiveness measures achieved by various search models with the whole collection and with each of our four sub-collections. These results from this evaluation are useful in our context, since our investigations are not limited to a single search model. Rather, they may be used with different search strategies, leading hopefully to a more general conclusion.

2.1. Environment

In order to obtain a rough idea regarding the retrieval effectiveness of our sub-collections compared to that of the whole TREC8 collection, we conducted different experiments using various weighting schemes, the vector-processing model (denoted using SMART parlance, see Appendix 1) and the OKAPI probabilistic model. To adjust the underlying parameters of the OKAPI search model, we used advl = 750, b = 0.9, k_1 = 2. For the LNU weighting scheme, we set the parameters to: slope = 0.2 and pivot = 150.

The results depicted in Table 6 show that the retrieval effectiveness of each sub-collections was higher than that of the whole collection, but it must be remembered that the number of queries and the number of relevant documents were not the same across all sub-collections. We do think however that this information indicates that a good selection procedure may enhance the retrieval effectiveness compared to the average precision achieved from using the whole collection.

Collection	FT	FR	FBIS	LA Times	TREC8
Size (in MB)	564 MB	395 MB	470 MB	475 MB	1,904 MB
# of documents	210,158	55,630	130,471	131,896	528,155
# of distinct index terms / document					
mean	124.4	131.16	141.56	158.46	136.84
standard error	93.26	127.95	125.04	124.11	114.54
median	101	128	107	122	108
maximum	3,050	23,517	5,677	5,040	23,515
minimum	6	2	6	10	2
# of indexing terms / document					
mean	195.62	320.11	267.2	262.86	240.89
standard error	172.66	1,128.3	598.82	248.5	501.35
median	151	289	168	184	171
maximum	13,761	211,944	61,300	16,100	211,934
minimum	6	2	10	10	2
# of distinct indexing terms	375,499	196,220	502,099	337,492	1,008,463
min idf	$0.092 \cdot 10^{-4}$	$0.1845 \cdot 10^{-4}$	$0.0805 \cdot 10^{-4}$	$0.3794 \cdot 10^{-4}$	$6905.49 \cdot 10^{-4}$
max df	210,156	55,629	130,470	131,891	264,765
time to build the inverted file	32:01	12:55	24:33	33:06	
from Topics #301 to #450					
# of relevant documents	4,903	844	4,410	3,535	13,692
# of queries	144	69	116	143	150
from Topics #301 to #400					
# of relevant documents	3,233	638	2,743	2,350	8,964
# of queries	95	50	60	98	100
from Topics #401 to #450					
# of relevant documents	1,670	206	1,667	1,185	4,728
# of queries	49	19	43	45	50

Table 5: Statistics on TREC8 Collections

	Precision				
Collection Model	FT 49 queries 1,670 rel.	FR 19 queries 206 rel.	FBIS 43 queries 1,667 rel.	LA TIMES 45 queries 1,185 rel.	TREC8 50 queries 4,728 rel.
OKAPI - NPN	40.00	38.27	33.75	31.11	29.65
LNU - LTC	34.17	25.64	25.50	26.94	24.57
ATN - NTC	33.96	35.56	30.65	27.90	26.25
NTC - NTC	18.63	17.35	13.92	15.79	13.09
LTC - LTC	23.60	30.85	22.59	21.15	17.49
LNC - LTC	25.28	23.75	20.32	24.67	19.40
LNC - LNC	18.26	11.24	12.42	21.86	12.05
ANC - LTC	24.39	26.20	20.53	23.04	17.51
NNN - NNN	6.97	3.25	2.71	6.90	1.61
BNN - BNN	9.00	5.74	5.04	3.65	3.12

Table 6: Average Precision of Isolated Collections (Query = Title, Desc & Narr)

2.2. Selection Strategy

As a first attempt to define a selection procedure, we wanted a strategy that, based on the current request, might produce a binary outcome, specifying whether or not the underlying sub-collection contained pertinent document(s) or not. Our challenge was to define an automatic procedure that would answer to the question "Does this collection (with its search engine) provide a satisfactory answer (at least one relevant document) to this question?". Therefore, the expected answer was not an integer value specifying the number of records to be retrieved from the underlying sub-collection but a binary outcome. With such a procedure, the computer could be aware of the limits of its knowledge, knowing when it does not know.

In this study, we wanted to verify whether or not past requests might be useful sources of evidence for such selection purposes. To achieve this, we defined a selection procedure based on the k-nearest neighbors (k-NN) technique that works as follows (see also [Voorhees 95], [Voorhees 96], [Savoy 96b]).

For each new topic Q, the system found the k nearest neighbors in the set of all existing requests Q_j, j = 1, 2, ..., m (m = 149, k = 3, cosine measure). The three-best ranked past requests were retrieved and the system determined whether or not, for those three requests, the underlying sub-collection contained any pertinent records. Based on the majority rule, the system might decide whether or not to conduct a search into the underlying sub-collection.

During the testing stage of our system (based on Topics #301 to #400), we noticed that the FT sub-collection contained pertinent information for 95 queries out of a total of 100, while the LA sub-collection had relevant documents for 98 queries. Therefore, we decided, for each new request (Topics #401 to #450), to search in both the FT and LA sub-collections without considering our selection procedure. On the other hand, based on the training requests (Topics #301 to #400), the FR collection may produce relevant information for 50 queries and the FBIS sub-collection for 60. Therefore, we apply our selection procedure only for these two sub-collections.

The complete evaluation of our decision rule is given in Tables 7. First, in Table 7a, the decision taken by the system is represented in the rows while the true state of Nature is depicted in the columns. For example, the number "8" indicates that 8 times the system decided to retrieve information from the FR sub-collection and these decisions were correct (true positive). Of course, our selection procedure produces also errors, e.g., for the FR collection, it decided four times to conduct a search while this corpus did not hold any relevant information (false positive).

As an overall correctness indicator, we would compute the accuracy of the decision rule by dividing the number of correct answers (true positive + true negative) by the number of cases. Other evaluation measures are depicted in Table 7b. From these results, it can be seen that the k-nearest neighbors (k-NN) technique does not result in a satisfactory overall performance. Our selection rule is not very sensitive and often fails to conduct a search when it is appropriate.

Our selection procedure is thus far from perfect and the retrieval performance it achieves is also affected by its poor decision-making performance, as shown in the last column of Table 8 (merging according to the raw-score strategy, see Section 2.3). Indicated in the second column of this table is the average precision achieved when all the documents formed a single huge collection (baseline). Depicted in the third column is the average performance we might expect when, for all requests, we decided to search in all the sub-collections and merged the four result lists based on the raw-score merging strategy (see Section 2.3). Under the heading "Optimal Selection" are listed the average precision obtained using an error-free (perfect) selection procedure, ignoring sub-collections having no relevant information for a given query (merging done by the raw-score scheme).

FR	true state			FBIS	true state	
prediction	do retrieve	no retrieve		prediction	do retrieve	no retrieve
do retrieve	8	4		do retrieve	13	2
no retrieve	11	27		no retrieve	30	5
total	19	31		total	43	7

Table 7a: Evaluation of Our Selection Procedure

Measure \ Collection	FR	FBIS
Accuracy (# correct decisions / # cases)	35 / 50 = 0.7	18 / 50 = 0.36
Sensitivity (# true positive / # positive cases)	8 / 19 = 0.42	13 / 43 = 0.302
Specificity (# true negative / # negative cases)	27 / 31 = 0.871	5 / 7 = 0.714

Table 7b: Various Evaluation Measures of Our Selection Rule

	Precision (% change)			
Strategy	Single Collection	No Selection	Optimal Selection	Our Selection Approach
OKAPI-NPN	29.65	27.39 (-7.62%)	29.31 (-1.15%)	22.64 (-23.64%)
LNU - LTC	24.57	23.75 (-3.33%)	24.55 (-0.08%)	19.25 (-21.65%)
ATN - NTC	26.25	24.64 (-6.13%)	26.18 (-0.27%)	20.51 (-21.87%)
NTC - NTC	13.09	12.89 (-1.53%)	13.59 (+3.82%)	11.56 (-11.69%)
LTC - LTC	17.49	16.26 (-7.03%)	17.49 (0.00%)	13.61 (-22.18%)
LNC - LTC	19.40	19.00 (-2.06%)	19.81 (+2.11%)	15.45 (-20.36%)
LNC - LNC	12.05	12.31 (+2.16%)	13.05 (+8.30%)	10.13 (-15.93%)
ANC - LTC	17.51	17.47 (-0.23%)	18.32 (+4.63%)	13.88 (-20.73%)
NNN - NNN	1.61	1.60 (-0.62%)	2.62 (+62.73%)	3.31 (+105.6%)
BNN - BNN	3.12	3.15 (+0.96%)	3.74 (+19.98%)	2.22 (-28.85%)

Table 8: Average Precision of Various Selection Strategies and Merging Done by the Raw-Score Strategy (Query = Title, Desc & Narr)

2.3. Collection Merging

Recent works have suggested that some solutions to the merging of separate answer lists may be obtained from distributed information services. As a first approach, we might assume that each database contains approximately the same number of pertinent items and that the distribution of the relevant documents is the same across the servers' answers. Based only on the ranking of retrieved records, we might interleave the results in a round-robin fashion. According to previous studies [Voorhees 95], [Callan 95], the retrieval effectiveness of such interleaving schemes is around 40% below the performance achieved by a single retrieval scheme working, with a single huge collection representing the entire set of documents. The third column of Table 9 confirms this finding but to a lesser extent (around -27%).

In order to take account of the score achieved by the retrieved document, we might formulate the hypothesis that each information server applies the same or a very similar search strategy and that the similarity values are therefore directly comparable [Kwok 95], [Moffat 95]. Such a strategy, called raw-score merging, produces a final list, sorted by the retrieval status value computed by each separate search engine. However, as demonstrated by Dumais [94], collection-dependent statistics in document or query weights may vary widely among sub-collections; and therefore, this phenomenon may invalidate the raw-score merging hypothesis. The fourth column of Table 9 indicates the retrieval effectiveness of such merging approach, showing a relatively interesting performance in our case (degradation of around -2.5%). Thus, the raw-score merging seems to be a simple and valid approach when a huge collection is distributed across a local-area network and operated within the same retrieval scheme.

As a third merging strategy, we may normalize each sub-collection's retrieval status value (RSV) by dividing it by each result list's maximum RSV. The fifth column of Table 9 shows its average precision, representing surprisingly poor retrieval effectiveness (average reduction of -25%).

Finally, we suggest using the logistic regression approach to resolve merging problems that have shown interesting performance levels when merging heterogeneous result lists produced by different search models where only ranks of the retrieved items are available as a key for merging [Le Calvé 99]. In the current case, the explanatory variables are the logarithm of the rank of the retrieved item together with its score. The average precision achieved by this method shown in the last column of Table 9 is similar to the raw-score merging strategy.

2.4. Official Ad Hoc Runs

Our first official run (UniNET8St, ad hoc, automatic, short queries) resulted in an average precision of 29.06, 38 times greater than the median and for two queries (#403, #416), it revealed the best results. Our second official run (UniNET8Lg, ad hoc, automatic, long queries) resulted in an average precision of 31.38, 40 times greater than the median and for four queries (#416, #429, #431, #438), it revealed the best results. Both results were obtained using the OKAPI retrieval scheme with blind query expansion ($\alpha = 0.75$, $\beta = 0.75$) and the system was allowed to add 50 search terms to the original query during feedback, with added terms extracted from the 5-best ranked documents.

2.5. Conclusion

When dealing with distributed collections across a local area network and using the same retrieval model for all these sub-collections, our experiments show that:

- Selection procedure, based on k-NN technique, does not seem to be worthwhile approach;
- Based on various search strategies, it seems that the raw-score approach might be a valid first attempt for merging result lists provided by the same retrieval model.

Acknowledgments

The authors would like to thank 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 SNSF (Swiss National Science Foundation) under grant 20-50'578.97 and 20'55'771.98). The authors would also like to thank Yves Rasolofo for his help computing the logistic regression coefficient estimates.

	Precision (% change)				
Strategy Model	Single Collection	Round -Robin	Raw-Score Merging	Normalized Score	Logistic Regression
OKAPI-NPN	29.65	21.61 (-27.12%)	27.39 (-7.62%)	22.66 (-23.58%)	26.83 (-9.51%)
LNU - LTC	24.57	17.72 (-27.88%)	23.75 (-3.33%)	17.35 (-29.38%)	23.86 (-2.89%)
ATN - NTC	26.25	19.07 (-27.35%)	24.64 (-6.13%)	19.74 (-24.80%)	23.29 (-11.28%)
NTC - NTC	13.09	9.25 (-29.33%)	12.89 (-1.53%)	9.59 (-26.74%)	12.64 (-3.44%)
LTC - LTC	17.49	13.12 (-24.99%)	16.26 (-7.03%)	12.96 (-25.90%)	16.67 (-4.69%)
LNC - LTC	19.40	13.69 (-29.43%)	19.00 (-2.06%)	14.11 (-27.27%)	18.82 (-2.99%)
LNC - LNC	12.05	9.40 (-21.99%)	12.31 (+2.16%)	8.71 (-27.72%)	12.75 (+5.81%)
ANC - LTC	17.51	13.40 (-23.47%)	17.47 (-0.23%)	13.21 (-24.56%)	17.52 (+0.06%)
NNN - NNN	1.61	2.76 (+71.43%)	1.60 (-0.62%)	0.77 (-52.2%)	3.54 (+119.88%)
BNN - BNN	3.12	2.71 (-13.14%)	3.15 (+0.96%)	2.34 (-25.0%)	2.78 (-10.90%)

Table 9: Average Precision of Various Merging Strategies (Query = Title, Desc & Narr)

Official Run Name	Average Precision	# ≥ Median	# Best
UniNET8St	29.06	38	2
UniNET8Lg	31.38 (+7.98%)	40	4

Table 10: Summary of our Official Ad Hoc Runs

References

[Alschuler 89] L. Alschuler : Hand-Crafted Hypertext - Lessons from the ACM Experiment. In E. Barrett (Ed.), The Society of Text, Hypertext, Hypermedia, and the Social Construction of Information, (pp. 343-361), The MIT Press, Cambridge (MA), 1989.

[Bernes-Lee 94] T. Bernes-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, A. Secret: The World-Wide Web. Communications of the ACM, 37(8), 1994, 76-82.

[Bharat 98] K. Bharat, M. Henzinger: Improved Algorithms for Topic Distillation in Hyperlinked Environments. Proceedings of ACM-SIGIR'98, Melbourne (Australia), August 1998, 104-111.

[Bookstein 92] A. Bookstein, E. O'Neil, M. Dillon, D. Stephens: Applications of Loglinear Models for Informetric Phenomena. Information Processing & Management, 28(1), 1992, 75-88.

[Brin 98] S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of WWW8, Brisbane (Australia), April 1998, 107-117. http://google.stanford.edu.

[Buckley 96] C. Buckley, A. Singhal, M. Mitra, G. Salton: New Retrieval Approaches using SMART. Proceedings of the TREC'4, Gaithersburg (MD), NIST publication 500-236, 1996, 25-48.

[Callan 95] J. P. Callan, Z. Lu, W. B. Croft: Searching Distributed Collections with Inference Networks. Proceedings of the ACM-SIGIR'95, Seattle (WA), 1995, 21-28.

[Chakrabarti 99] S. Chakrabarti, M. Van den Berg, B. Dom: Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. Proceedings of WWW8, Toronto (ON), May 1999, 545-562.

[Cohen 87] P. R. Cohen, R. Kjeldsen: Information Retrieval by Constrained Spreading Activation in Semantic Networks. Information Processing & Management, 23(4), 1987, 255-268.

[Dreilinger 97] D. Dreilinger, A. E. Howe: Experiences with Selecting Search Engines using Metasearch. ACM Transactions on Information Systems, 15(3), 1977, 195-222.

[Dumais 94] S. T. Dumais: Latent Semantic Indexing (LSI) and TREC-2. Proceedings of TREC'2, Gaithersburg (MD), NIST Publication #500-215, 1994, 105-115.

[Fuhr 99] N. Fuhr: A Decision-Theoretic Approach to Database Selection in Networked IR. ACM Transactions on Information Systems, 1999, to appear.

[Gordon 99] M. Gordon, P. Pathak: Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines. Information Processing & Management, 35(2), 1999, 141-180.

[Gravano 97] L. Gravano, K. Chang, H. García-Molina, C. Lagoze, A. Paepcke: STARTS - Stanford Protocol Proposal for Internet Retrieval and Search. Computer Systems Laboratory, Stanford University, Stanford (CA).

[Halasz 88] F. G. Halasz: Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems. Communications of the ACM, 31(7), 1988, 836-852.

[Hawking 99a] D. Hawking, P. Thistlewaite: Methods for Information Server Selection. ACM Transactions on Information Systems, 17(1), 1999, 40-76.

[Hawking 99b] D. Hawking, N. Craswell, P. Thistlewaite, D. Harman: Results and Challenges in Web Search Evaluation. Proceedings WWW8, Toronto (ON), 1999, 243-252.

[Kahle 93] B. Kahle, H. Morris, J. Goldman, T. Erickson, J. Curran: Interfaces for Distributed Systems of Information Servers. Journal of the American Society for Information Science, 44(8), 1993, 453-485.

[Kleinberg 98] J. Kleinberg: Authoritative Sources in a Hyperlinked Environment. Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, January 1998, 668-677.

[Kwok 95] K. L. Kwok, L. Grunfeld, D. D. Lewis: TREC-3 Ad-hoc, Routing Retrieval and Thresholding Experiments using PIRCS. Proceedings of TREC'3, Gaithersburg (MD), NIST Publication #500-225, 1995, 247-255.

[Lawrence 99] S. Lawrence, C. Lee Giles: Accessibility of Information on the Web. Nature 400 (6740), 8th July 1999, 107-110.

[Le Calvé 99] A. Le Calvé, J. Savoy: Database Merging Strategy based on Logistic Regression. Information Processing & Management, 1999, to appear.

[Leighton 99] H. V. Leighton, J. Srivastava: First 20 Precision among World Wide Web Search Services (Search Engines). Journal of the American Society for Information Science, 50(10), 1999, 870-881.

[Marchiori 97] M. Marchiori: The Quest for Correct Information on the Web: Hyper Search Engines. Proceedings of WWW6, Santa Clara (CA), April 1997.

[Moffat 95] A. Moffat, J. Zobel: Information Retrieval Systems for Large Document Collections. Proceedings of TREC'3, Gaithersburg (MD), NIST Publication #500-225, 1995, 85-93.

[Picard 98] J. Picard: Modeling and Combining Evidence Provided by Document Relationships using Probability Argumentation Systems. Proceedings of ACM-SIGIR'98, Melbourne (Australia), 1998, 182-189.

[Robertson 95] S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu: Large Test Collection Experiments on an Operational, Interactive System: OKAPI at TREC. Information Processing & Management, 31(3), 1995, 345-360.

[Salton 89] G. Salton: Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading (MA), 1989.

[Savoy 94] J. Savoy: A Learning Scheme for Information Retrieval in Hypertext. Information Processing & Management, 30(4), 1994, 515-533.

[Savoy 96a] J. Savoy: Citation Schemes in Hypertext Information Retrieval. In Information Retrieval and Hypertext, M. Agosti, A. Smeaton (Eds), Kluwer, Amsterdam (NL), 1996, 99-120.

[Savoy 96b] J. Savoy, M. Ndarugendamwo, D. Vrajitoru: Report on the TREC-4 Experiment: Combining Probabilistic and Vector-Space Schemes. Proceedings TREC'4, NIST publication 500-236, Gaithersburg (MD), October 1996, 537-547.

[Savoy 97] J. Savoy: Ranking Schemes in Hybrid Boolean Systems: A New Approach. Journal of the American Society for Information Science, 48(3), 1997, 235-253.

[Small 99] H. Small: Visualizing Science by Citation Mapping. Journal of the American Society for Information Science, 50(9), 1999, 799-813.

[Sparck Jones 77] K. Sparck Jones, R. G. Bates: Research on Automatic Indexing 1974-1976. Technical Report, Computer Laboratory, University of Cambridge, UK.

[Voorhees 95] E. M. Voorhees, N. K. Gupta, B. Johnson-Laird: Learning Collection Fusion Strategies. Proceedings of the ACM-SIGIR'95, Seattle (WA), 1995, 172-179.

[Voorhees 96] E. M. Voorhees: Siemens TREC-4 Report: Further Experiments with Database Merging. Proceedings TREC'4, NIST publication 500-236, Gaithersburg (MD), 1996, 121-130.

[Xu 98] J. Xu, J. P. Callan: Effective Retrieval with Distributed Collections. Proceedings of the ACM-SIGIR'98, Melbourne (Australia), 1998, 112-120.