Report on the TREC-3 Experiment: A Learning Scheme in a Vector Space Model

Jacques Savoy, Melchior Ndarugendamwo, Dana Vrajitoru

Faculté de droit et des sciences économiques
Université de Neuchâtel
Pierre-à-Mazel 7
CH - 2000 Neuchâtel (Switzerland)

Summary

This paper describes and evaluates a retrieval scheme, or more precisely an additional retrieval mechanism based on interdocument relationships, that can be integrated in almost all existing retrieval schemes (e.g., Boolean, hybrid Boolean, vector-processing or probabilistic models). The intent of our approach consists of inferring knowledge about document contents based on the relevance assessments of past queries. Through a learning process, our scheme establishes relevance links between documents found relevant for the same request. Based on this information and a list of retrieved records for the current request, the proposed mechanism tries to improve the ranking of the retrieved items in a sequence most likely to satisfy user intent. The underlying hypothesis of this mechanism states that future requests addressed to the system should have some degree of similarity with previous queries, or that the retrieval apparatus will process requests for which it has already found a partial, appropriate answer in the past.

Participation: Category: B Query: ad-hoc, fully automatic

Introduction

To find pertinent information from a large text collection, most retrieval models represent both documents and requests by a set of weighted keywords. To extract relevant records from this collection, the retrieval function computes a similarity value or estimates a probability of relevance based on both document and query surrogates.

When applying such a scheme, the system considers documents as separate entities. To relax this assumption, some studies have proposed various techniques and have reported evaluations describing the importance of interdocument relationships (e.g., [Kwok 88], [Turtle 91]). Our main research objective is also to analyze and assess interdocument relationships as a useful source of document contents evidence. In this vein, we have already investigated the relative importance of explicit (e.g., bibliographic reference), implicit (bibliographic coupling, co-citation) and computed links (nearest neighbor) between documents [Savoy 94a]. In this study, we are concerned with the means by which the system may have derived other relationships between documents based on past queries and their relevance assessments.

This paper is made up of two sections. The first describes our learning scheme and presents some related works. The second section shows and explains results obtained using the Wall Street Journal corpus, a subset of the TIPSTER-DARPA collection, and discusses some problems related to traditional evaluation methodology.

1. Learning Scheme

Evaluation of current retrieval models has shown that their retrieval effectiveness is far from perfect and one of the principal explanations of this lack of effectiveness is related to the ambiguity of natural language. This problem has two facets: on the one hand, the same idea or concept may be expressed by various forms [Furnas 87], and, on the other hand, the same word may have more than one meaning, even in a specialized corpus [Krovetz 92].

In order to resolve this difficulty, Blair [90] suggests that a retrieval model must have better document contents representation:

> "The central problem of Information Retrieval is how to represent documents for retrieval. The most intricate or carefully designed retrieval algorithm cannot compensate for inappropriately represented documents. ... The central task of Information Retrieval research is to understand how documents should be represented for effective retrieval.

This is primarily a problem of language and meaning." [Blair 90, vii]

This may lead to a perception of the retrieval system as an adaptive process, allowing better communication between the searcher and the indexer (or the author(s)) [Blair 90].

From a practical point of view, one feasible approach to the design of such a learning scheme consists of taking into account the knowledge obtained from past queries, or more precisely, from their relevance judgments, in order to enhance system's retrieval effectiveness over time. We also believe that documents found relevant for a given request do share similar concepts [Savoy 94b]. Thus, past queries and their relevance assessments may be a useful source of information about the meaning of documents and may be helpful in ranking the retrieved records in a sequence that more closely reflects the user's intent.

The first subsection describes the main principles underlying our learning scheme. The second presents the design of our adaptive model and the guidelines for its implementation. The third subsection shows statistics related to query similarities and the fourth one describes the main features of related researches.

1.1. Motivation

The aim of a learning scheme is to provide the system with the ability to record its successes and failures, and thus infer knowledge useful for increasing its performance over time. To define such a mechanism, we have to specify the underlying hypotheses, determine how the system learns and how it stores and uses the knowledge provided by previous experiments.

Our learning scheme is based on the following principles:

a) Documents known to be relevant to the same query tend to contain similar concepts and must deal with similar subjects;
b) No conclusions can be drawn about documents found nonrelevant for a given request.

On the one hand, our learning scheme is based exclusively on successes, i.e., on the presence of couples of retrieved and relevant documents. On the other hand, our procedure does not take into account the shared presence of retrieved and nonrelevant items. Nonrelevant records retrieved by the system are those documents that have at least one common keyword with the request. However, such keyword matching does not always imply word sense matching:

"Word sense mismatches are far more likely to appear in nonrelevant documents than in those that are relevant." [Krovetz 92, p. 139].

Our prior feeling is that negative relevance feedback information does not really represent useful information. By analogy, if you are lost in a desert, a negative relevance feedback only tells you that "you are on the wrong path" and does not provide "efficient" hints as to the path leading to the for i = 1, 2, ... nearest city. This fact is confirmed by relevance feedback studies which have demonstrated that positive relevance information depicts more valuable information than negative one [Salton 90]. However, this approach considers only relevance data given for the current request and a direct comparison with this scheme is therefore not suitable.

1.2. Implementation of our Learning Scheme

In order to represent the information given by the previous experiments or requests, we have designed a special interdocument relationship called a *relevance link*. This link type connects two documents found relevant for a given query. Associated with each link, a *relevance value* specifies how many times both the linked documents are found relevant.

To account for the information provided by the learning stage, our retrieval scheme works in two phases. In the first, the retrieval status value (RSV) of each document is computed according to a well-known retrieval scheme. To achieve this step, one can use the p-norm model, a vector-processing scheme or a probabilistic retrieval strategy. In the second stage, the ranking of retrieved documents is modified according to the presence of relevance links according to the following equation.

$$RSV(D_i) = RSV_{init}(D_i) + \sum_{k=1}^{s} \alpha_{ik} \cdot RSV_{init}(D_k) \quad \text{for } i = 1, 2, \ldots, m \qquad (1)$$

in which α_{ik} reflects the strength of the relationship between Documents i and k and s the number of neighbors of Document i. At the initial stage, the retrieval status value of a document depends only on the similarity between its surrogate and the query ($RSV_{init}(D_i)$, computed according to the Vector Space Model presented later in this paper). The value α_{ik} can be either a constant or a function of the relevance value of the link connecting Documents i and k.

To illustrate the way our retrieval proposal works, Figure 1 depicts some relevance links with their respective relevance values. In the first step of our retrieval process, a vector-processing scheme attributes a retrieval status of 0.8 to Document 11. According to Formula 1, this weight is propagated through links to Documents 3, 7 and 10. If we define the strength of the link between Nodes 11 and 7 as 0.3, Document 7 will increase its retrieval status value by 0.24.

In order to improve the efficiency of our retrieval scheme and to guarantee a reasonable processing time, we modify the retrieval status value not for all retrieved records, but we select the first m best-ranked documents after the initial stage to activate the relevance links (the constant m in Equation 1).

We believe that relevance links indicate semantic relationships between documents and may be valuable in the searching process. Although Blair [90] considers such a scheme to be a useful pedagogical tool, he questions its retrieval effectiveness:

> "Bush [45] recognized early ... how inquirers could benefit from the "traces" left by searches conducted by informed inquirers. While this is an important notion, realistically each inquirer's searches are unique enough that a record of previous searches might only be marginally useful for finding specific information." [Blair 90, p. 181]

The main underlying hypothesis of our retrieval model is that coming requests have some relationships with previous ones. On the contrary, if future queries are totally dissimilar with past queries, our scheme will have little hope of improving and may possibly decrease the retrieval effectiveness of the response.

1.3. Similarity Between Queries

In order to provide an indication of the degree of similarity between requests in three test-collections, we have computed some statistics, depicted in Table 1. This table shows that the CISI collection included more pertinent records per query than the CACM corpus (perhaps "too many relevant documents per query" [Fox 83, p. 7]). The Wall Street Journal collection included in the TIPSTER-DARPA collection reveals a similar pattern.

The second part of Table 1 illustrates the computed similarity between requests according to their relevance assessments. For this computation, we used the Dice's simple coefficient [van Rijsbergen 79, p. 39]. Requests from the CISI test-collection reveal a higher degree of similarity between them than for those of the CACM or in the WSJ. However, the mean similarity between queries is rather low in two cases (CISI: 0.04; CACM: 0.0182) and very low for the subset of the TIPSTER collection (WSJ: 0.00228).

Moreover, the estimated standard error is relatively high indicating that the empirical distribution of the similarity values is mostly in the range 0.0 to 0.1. The second part of Table 1 confirms this fact. For the CISI corpus, 88.4% (526 over a total of 595) of the similarity values are less or equal to 0.1 while for the CACM and WSJ these numbers are 99.45% and 99.5% respectively.

In our evaluations, we have built two sets of relevance links, namely the RF set containing all relevance links and the RF1 set including all relevance links having a relevance value strictly greater than one. For example, in Figure 1, the RF1 set contains only one relevance link (between Document 3 and 7). Table 2 presents the statistics associated with both sets.

From this data, one can see that for the CISI collection, the set RF contains 66,067 links from which 63,889 (around 97%) have a relevance value of one (CACM: 96.7%). The WSJ corpus depicts a more extreme case (1,707,152 over 1,738,429 links in total or 98.2%).

Finally, it is interesting to note that when building a test-collection, we are trying (consciously or not) to write queries for which the relevance judgments are as dissimilar as possible, a phenomena reflected in the above statistics. Such a practice mirrors the designer's wishes that the underlying requests must cover different concepts contained in the corpus. When we designed our additional retrieval mechanism, we formulated a contradictory hypothesis which should hold in commercial retrieval services or, at least, we hope so.

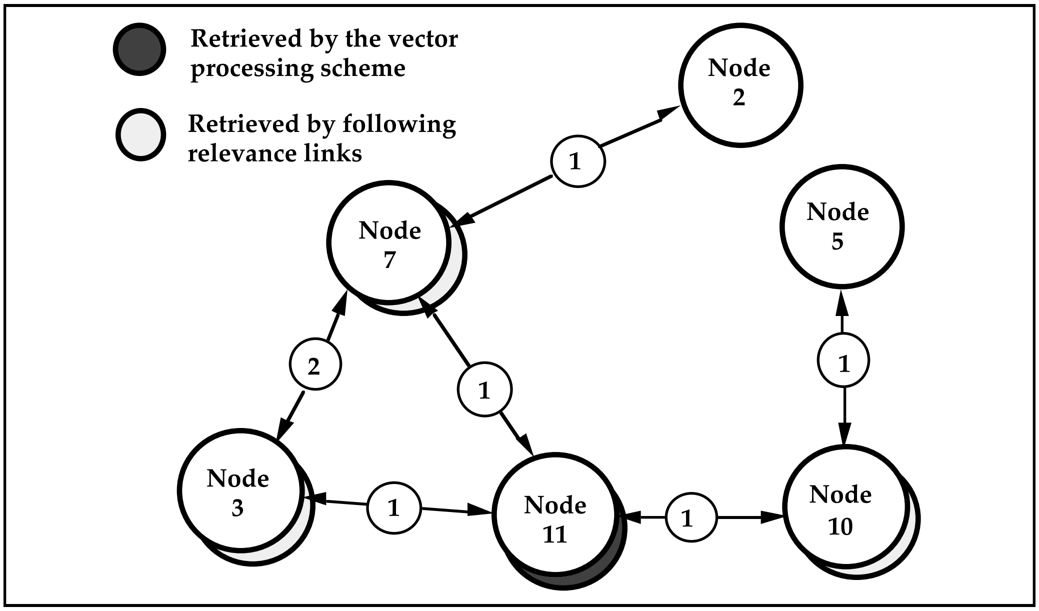

Figure 1: Retrieval of Information Using Relevance Links

Statistics \ Collection	CACM 50 queries	CISI 35 queries	WSJ 150 queries
# documents	3,204	1,460	173,252
# relevant documents	792	1,742	17,069
# distinct relevant documents	554	925	14,280
Mean rel. doc per request	15.84	49.77	113.79
Estimated standard error	12.59	39.96	104.56
Min. # relevant documents	1 (#q: 33)	1 (#q: 6)	2 (#q: 121)
Max. # relevant documents	51 (#q: 25)	144 (#q: 20)	591 (#q: 56)
Similarity between queries (Dice)			
Mean similarity measure	0.0182	0.04	0.00228
Estimated standard error	0.0756	0.0671	0.0151
# 0.00 < SIM ≤ 0.05	1,117	438	11,057
# 0.05 < SIM ≤ 0.1	40	88	66
# 0.10 < SIM ≤ 0.15	27	38	29
# 0.15 < SIM ≤ 0.2	6	17	10
# 0.20 < SIM ≤ 0.25	5	2	6
# 0.25 < SIM ≤ 0.3	3	4	2
# 0.30 < SIM ≤ 0.35	8	2	2
# 0.35 < SIM ≤ 0.4	5	1	0
# 0.40 < SIM ≤ 0.45	2	2	2
# 0.45 < SIM ≤ 0.50	1	0	0
# 0.50 < SIM ≤ 0.55	4	1	1
# 0.55 < SIM ≤ 0.60	4	2	0
# 0.60 < SIM ≤ 0.65	1	0	0
# 0.65 < SIM ≤ 0.70	1	0	0
# 0.70 < SIM ≤ 0.75	0	0	0
# 0.75 < SIM ≤ 0.80	0	0	0
# 0.80 < SIM ≤ 0.85	1	0	0
# 0.85 < SIM ≤ 1.00	0	0	0

Table 1: Relevance Information Characteristics

Statistics \ Collection	CACM 50 queries	CISI 35 queries	WSJ 150 queries
Relevance Link Value RF Set			
# links	8,876	66,067	1,738,429
$\alpha = 1$	8,265	63,889	1,707,152
$\alpha = 2$	495	2,044	30,337
$\alpha = 3$	90	121	907
$\alpha = 4$	23	13	31
$\alpha = 5$	3	0	2
$\alpha = 6$ and more	0	0	0
Mean α value	1.085	1.035	1.019
Estimated standard error	0.343	0.197	0.440
Relevance Link Value RF1 Set			
# links	611	2,178	31,277
$\alpha = 2$	495	2,044	30,377
$\alpha = 3$	90	121	907
$\alpha = 4$	23	13	31
$\alpha = 5$	3	0	2
$\alpha = 6$ and more	0	0	0
Mean α value	2.237	2.067	2.031
Estimated standard error	0.535	0.274	0.180

Table 2: Relevance Links Statistics

1.4. Related Research

In order to include permanently relevance feedback information, various researchers have proposed modification of the document surrogates in a vector-processing scheme. In this vein, Friedman et al. [71] present a framework within which the index term weight w_{ij} assigned to term t_j in document representative d_i can be modified to reflect user's judgments about document content. This proposition is grounded on three principles. The first one specifies that the modification of indexing term weights can occur only for the "good" keywords or for those appearing more frequently in the relevant documents. As a second principle, Friedman et al. [71] suggest deriving the new indexing term weight w'_{ij} in proportion to the existing weight w_{ij}. Thirdly, the authors suggest that the indexing weight modification must be based on the importance of term t_j in: (1) the current request, (2) the set of relevant documents, and (3) the sample of nonrelevant records.

In a related work, Brauen [71] also suggests transforming the document surrogates. In this approach, the system modifies only the document representatives of relevant records (relevance document modification). When implementing such a learning scheme, the system had to consider three cases:

1) a concept t_j is present in the request and absent from the relevant document d_i; thus, the system must add the "synonym" t_j to the corresponding document surrogate;
2) a concept t_j is present in both the request q_k and the relevant document d_i; the indexing term weight w_{ij} must be reinforced;
3) a concept t_j is present only in the relevant document d_i; the indexing term weight w_{ij} must be reduced.

From a different perspective, Gordon [88] suggests a learning scheme based on a genetic algorithm to enhance the retrieval effectiveness. In this approach, each document is described by various surrogates obtained using various binary indexing policies (e.g., based on document abstract, on titles, using full-text or derived from a manual indexing process). The retrieval system considers more than one description for each record and the competition between them will eliminate inappropriate surrogates while retaining more accurate ones. An iterative process affects document surrogates by including or removing index terms based on: (1) the reproduction of descriptions according to their average matching score in which a better representation has a higher chance to survive than others surrogates; and (2) cross-over between pairs of surrogates to generate new descriptors more appropriate to the retrieval of the corresponding document. The retrieval evaluations of the previously described learning schemes are based on relatively small test-collections: ADI collection (82 documents, 11 queries) for Friedman's experiment, CRANFIELD corpus (424 documents, 155 queries) in Brauen's paper, and Gordon's scheme (18 documents).

Of course, other learning strategies have already been proposed and evaluated, and most of them are directly related to the probabilistic retrieval model [Kwok 90]. Current probabilistic retrieval models [Cooper 92], [Gey 94], [Fuhr 91], [Fuhr 94] consider statistical clues present in the texts of document and queries to infer a probability of relevance. Following [Gey 94], these are hints of the absolute and relative term frequency in the document and in the request, the inverse document frequency and the relative term frequency in the collection. Experimental results have shown attractive retrieval effectiveness. Moreover, Gey [94] has shown that one can compute the value of various parameters according to a given test-collection and report them for other test-collections.

When comparing these probabilistic models with our learning scheme, one can see that they do not operate at the same level of granularity. By analogy with physics, the probabilistic retrieval models lay stress on the components of a document; they operate on an atomic level, whereas our approach, considering words as ambiguous entities, works at a molecular level.

2. Evaluation

To evaluate our proposed strategy and in order to be able to manage a large collection, we have worked with the SMART system [Salton 71]. This vector-processing scheme retrieves, for each request, an ordered list of retrieved records forming the input of our retrieval scheme. To implement our learning model, we have written the needed programs in Smalltalk-80 (an interpreted object-oriented language) and communication between these two systems is achieved by a common file.

2.1. Evaluation of the Vector Space Model

To represent each document and each query by a set of weighted keywords, we have used the SMART indexing system. To select the more appropriate weighting scheme for this operation, we have conducted a set of experiments based on different weighting formulas.

Firstly, to assign an indexing weight w_{ij} reflecting the importance of each single-term t_j, $j = 1, 2, \ldots, t$, in a document d_i, we may use the following equation:

$$\text{NNN:}\quad w_{ij} = tf_{ij} \tag{2}$$

where tf_{ij} depicts the frequency of the term t_j in the document d_i (or in the request).

To normalize each indexing weight between 0 and 1, we may consider the cosine normalization which is:

$$\text{LNC:}\quad w_{ij} = \frac{\log(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t} [\log(tf_{ik}) + 1]^2}} \tag{3}$$

Finally, we may also take account of the distribution of each indexing term in the collection by giving a higher weight to sparse words and a lower importance to more frequent terms (idf component).

$$\text{LTC:}\quad w_{ij} = \frac{[\log(tf_{ij})+1] \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ([\log(tf_{ik})+1] \cdot idf_k)^2}},$$

$$\text{with}\quad idf_j = \log\left[\frac{n}{df_j}\right] \tag{4}$$

in which n represents the number of documents d_i in the collection, df_j the number of documents in which t_j occurs, and idf_j the inverse document frequency.

The retrieval effectiveness of various combinations of these weighting formulas are reported in Table 3. Since latter evaluation outcomes are computed according to the ten standard recall values, Table 3 depicts results obtained using 10 recall-precision points. Finally, to decide whether one search strategy is better than another, the following rule of thumb is used: a difference of at least 5% in average precision is generally considered significant and, a 10% difference is considered very significant [Sparck Jones 77, p. A25].

For an unknown reason, the best weighting scheme seems to include the idf component only to weight the keywords included in requests and not during the indexing of documents (doc = LNC, query = LTC). The presence of spelling errors can be a partial explanation of such an unexpected result. In the WSJ corpus, low-frequency words are often no longer English. Since the idf scheme assigns extremely high weights to those misspelled terms, the normalization procedure given by Equation 4 also attributes a high value to those terms. The documents containing such terms cannot be retrieved because they will have a relatively small retrieval status value.

In the following results, the weighting scheme "doc = LNC, query = LTC" has been used in the first stage of our retrieval system and forms the baseline of our comparisons. The evaluation under the label "UNINE1" reflects this weighting scheme for queries from #151 to #200.

2.2. Retrospective Evaluation

In order to evaluate a learning strategy, we may provide the learning system with all the available information (all the requests with their relevance assessments in our case). The retrieval effectiveness obtained under such circumstances is called a retrospective test or the apparent performance measure. The resulting average precision represents an upper bound of the performance of the underlying model. From a practical point of view, this measure is computed according to Equation 5,

$$P_{app} = \frac{1}{r} \cdot \sum_{k=1}^{r} AP_k(Q) \quad (5)$$

in which AP_k denotes the average precision at ten standard recall value for the k^{th} query, considering that the learning scheme is fitted using the entire query sample Q (having a size denoted by r or 150 in this paper).

Such retrospective evaluation returns retrieval effectiveness values that are too optimistic (biased high), reflecting an unrealistic situation. For example, in Table 4a, the learning scheme using all the relevance links (RF set) returns performance results that are *too good to be true*.

	Precision (% change)
Model (# of queries) \ Collection	WSJ
Vector Space Model (150 queries) using 10 recall-precision points	
doc = NNN, query = NNN	8.24
doc = LNC, query = LNC	24.81 (+201.1%)
doc = LTC, query = LNC	26.16 (+217.5%)
doc = LTC, query = LTC	28.93 (+251.1%)
doc = LNC, query = LTC	31.94 (+287.6%)

Table 3: Evaluation of the Vector Processing Scheme Done by SMART

	Precision (% change)
Model \ Collection	WSJ
Vector-processing (r=150 queries) (doc = LNC, query = LTC, 1,000 doc.)	31.9
Full Relevance Feedback	
P_{app} (α = 0.1, m: 10)	73.1 (+129.6%)
P_{app} (α = 0.15, m: 10)	77.8 (+144.4%)
P_{app} (α = 0.2, m: 10)	79.9 (+151.0%)
P_{app} (α = 0.3, m: 10)	81.7 (+156.4%)
P_{app} (α = 0.5, m: 10)	82.9 (+160.1%)
P_{app} (α = 0.9, m: 10)	83.8 (+163.2%)
Full Relevance Feedback (α = 0.9)	
P_{app} (m: 5)	82.8
P_{app} (m: 10)	83.8 (+1.2%)
P_{app} (m: 20)	82.9 (+0.3%)
P_{app} (m: 30)	83.2 (+0.5%)
P_{app} (m: 50)	83.2 (+0.5%)
P_{app} (m: 100)	81.5 (-1.6%)

Table 4a: Evaluation of Vector-Space Model with Full Relevance Feedback (RF Set)

	Precision (% change)
Model \ Collection	WSJ
Vector-processing (r=150 queries) (doc = LNC, query = LTC, 1,000 doc.)	31.9
Full Relevance Feedback	
P_{app} (α = 0.1, m: 10)	35.8 (+12.4%)
P_{app} (α = 0.15, m: 10)	36.5 (+14.5%)
P_{app} (α = 0.2, m: 10)	36.9 (+15.9%)
P_{app} (α = 0.3, m: 10)	37.4 (+17.5%)
P_{app} (α = 0.5, m: 10)	37.7 (+18.4%)
P_{app} (α = 0.9, m: 10)	37.8 (+18.5%)
Full Relevance Feedback (α = 0.9)	
P_{app} (m: 5)	36.3
P_{app} (m: 10)	37.8 (+4.1%)
P_{app} (m: 20)	38.1 (+5.0%)
P_{app} (m: 30)	38.6 (+6.3%)
P_{app} (m: 50)	38.3 (+5.5%)
P_{app} (m: 100)	37.1 (+2.2%)

Table 4b: Evaluation of Vector-Space Model with Full Relevance Feedback (RF1 Set)

When the apparent performance measure is too optimistic, it is generally an indication that the underlying learning scheme is over-fitted, too narrow for the given data, and cannot forget the details. What we really expect from a learning model is its capability to generalize given information, to retain the main features of the given information and to find useful relationships between data. In our point of view, the learning knowledge derived from RF set is over-fitted and this fact will be confirmed when considering the following subsection. Thus, the performance obtained using the RF1 set seems to depict a more realistic situation (see Table 4b).

2.3. Predictive Evaluation

If the evaluation results under full relevance feedback are usually misleading, more accurate or more "honest" evaluation estimate must be discussed. The basic principle underlying such an evaluation methodology is the following: the performance of a retrieval system must be based on requests other than those given to the learning scheme. Since in each test-collection the number of available queries is relatively small, evaluation must use all the available requests to adjust its parameter settings, on the one hand, and, on the other, all the available queries must be used to measure the performance of the proposed retrieval scheme. This latter fact may contribute to an objective comparison with a system ignoring learning.

To take account of these criteria, the hold-out method suggests splitting the queries sample into two disjoint parts: one subsample will be applied in the learning stage and the other will be used during the evaluation process. This division must be carried out randomly, without looking at the requests themselves. However, not all queries can be exploited both in the learning scheme and during the evaluation.

To overcome this drawback, multiple train-and-test experiments or random subsampling approaches can be considered within which all queries are used for testing, and almost all requests for training [Stone 74]. More precisely, the leaving-one-out approach, a special case of the cross-validation method, represents a solution which works as follows. The query sample Q of size r is divided into r sets. In the k^{th} set, one can find all requests except the k^{th} one. The model is fitted according to r-1 requests and an evaluation measure is computed according to the k^{th} query (not included in the learning sample). The above procedure is repeated for k = 1, 2, ... r and we combine the r prediction values to obtain an average precision measure (see Equation 6),

$$P_{lv} = \frac{1}{r} \cdot \sum_{k=1}^{r} AP_k(Q_{-k}) \qquad (6)$$

in which AP_k denotes the average precision at ten standard recall values for the k^{th} query under the condition that the learning scheme is fitted using the query set Q minus this k^{th} request. Such an evaluation strategy results in a real predictive measure because the system does not have any information about the current request during both the learning and the retrieval stages.

Table 5a depicts the retrieval evaluation obtained using the RF set. From these data, one can conclude that taking account of all relevance links does not improve the retrieval effectiveness when the value of the parameter α is less than or equal to 0.1. Setting this parameter to a higher value significantly decreases the retrieval performance. When considering the impact of the parameter m, one can see that the best value seems to be five.

	Precision (% change)
Model \ Collection	WSJ
Vector-processing (r=150 queries) (doc = LNC, query = LTC, 1,000 doc.)	31.9
Leaving-one-out	
P_{lv} (α = 0.1, m: 10)	30.9 (-2.9%)
P_{lv} (α = 0.15, m: 10)	29.9 (-6.3%)
P_{lv} (α = 0.2, m: 10)	28.8 (-9.6%)
P_{lv} (α = 0.3, m: 10)	26.9 (-15.4%)
P_{lv} (α = 0.5, m: 10)	24.5 (-23.1%)
P_{lv} (α = 0.9, m: 10)	21.0 (-34.2%)
Leaving-one-out (α = 0.1)	
P_{lv} (m: 5)	31.6
P_{lv} (m: 10)	30.9 (-2.2%)
P_{lv} (m: 20)	29.2 (-7.6%)
P_{lv} (m: 30)	28.0 (-11.4%)
P_{lv} (m: 50)	25.8 (-18.4%)
P_{lv} (m: 100)	22.4 (-29.1%)

Table 5a: Evaluation of Vector-Space Model Using the Leaving-one-out Method (RF Set)

	Precision (% change)
Model \ Collection	WSJ
Vector-processing (r=150 queries) (doc = LNC, query = LTC, 1,000 doc.)	31.9
Leaving-one-out	
P_{lv} (α = 0.1, m: 10)	31.8 (+0.0%)
P_{lv} (α = 0.15, m: 10)	31.7 (-0.5%)
P_{lv} (α = 0.2, m: 10)	31.6 (-0.8%)
P_{lv} (α = 0.3, m: 10)	31.4 (-1.3%)
P_{lv} (α = 0.5, m: 10)	31.2 (-2.1%)
P_{lv} (α = 0.9, m: 10)	30.7 (-3.5%)
Leaving-one-out (α = 0.1)	
P_{lv} (m: 5)	31.9
P_{lv} (m: 10)	31.8 (-0.3%)
P_{lv} (m: 20)	31.7 (-0.6%)
P_{lv} (m: 30)	31.5 (-1.2%)
P_{lv} (m: 50)	31.3 (-1.9%)
P_{lv} (m: 100)	30.4 (-4.7%)

Table 5b: Evaluation of Vector-Space Model Using the Leaving-one-out Method (RF1 Set)

When only considering relevance links having a relevance value greater than 1 (RF1 set), the retrieval performance cannot increase significantly as shown in Table 5b. When testing the system with various values for the parameters α and m, we cannot find a significant change in the retrieval effectiveness over the baseline ignoring learning. We also have to try to take account for 2,000 retrieved records instead of 1,000, but this alternative does not present any significant change over the results depicted in Table 5b.

From these results, it seems clear that the queries included in the WSJ collection do not have any pertinent relationship between them, or at least, such relationships are not detected by our learning model. This fact confirms our prior feeling as stated in Section 1.3.

The retrieval results submitted to the conference board under the label "UNINE2" are obtained using $\alpha = 0.05$ and $m = 5$ representing a conservative setting. This parameter setting reflects our prior opinion that the relevance judgments of queries #151 to #200 will not have a high degree of similarity with older requests.

2.4. Analysis of Official Results

The official results are based on queries #151 through #200. The results obtained under the label "UNINE1" represents the baseline or the first stage or our retrieval strategy (vector-processing scheme with index term weight = LNC, search term weight = LTC). The performance under the column "UNINE2" is obtained using our additional retrieval scheme with $\alpha = 0.05$ and $m = 5$.

From Table 6, we cannot conclude that our additional retrieval strategy represents significant change over the vector-processing scheme. Our approach retrieves the same relevant records as the vector space model but ranks them in a more suitable sequence, especially for medium or high recall values.

To define the setting for the UNINE2 experiment, we are faced, by analogy, with the following dilemma:

Solution 1: you may win $500.
Solution 2: you obtain a lottery ticket for which the probability of winning $1,000 is 0.5 and the probability of winning $0 is 0.5.

In both approaches, the expected win is the same ($500), however, Solution 2 can be considered risky. Since we have a loathing for risk, we have chosen Solution 1 in our parameters setting.

Conclusion

This paper suggests a learning algorithm based on interdocument relationships established according to relevance assessments obtained from previous requests. The underlying hypothesis of this scheme states that the relevance judgments of future queries will have a high degree of similarity with the relevance assessments of previous requests. To take account of this information, we propose an additional additive scheme within which relevance links are considered to increase the similarity between documents and query, and thus modify the ranking of retrieved documents.

Based on the WSJ collection, a retrospective test shows very attractive retrieval performance but the leaving-one-out method, representing a more realistic predictive measure, does not confirm this previous evaluation. We can conclude that the results of a retrospective test must be interpreted with caution. Since the queries included in a test-collection are written such that they cover different topics contained in the corpus, they do not have a high degree of similarity between them. This fact contradicts the underlying hypothesis of our learning scheme and may be a plausible explanation for the absence of any significant retrieval enhancement. However, even in such circumstances, our retrieval scheme does not significantly decrease retrieval effectiveness over a baseline ignoring learning.

If, traditionally, learning schemes are used mainly with probabilistic retrieval models, our solution is advantageous by being integrated with various Boolean models (p-norm, fuzzy set extension, hybrid Boolean strategies) or with the vector-processing scheme.

In this study, we never take relevance judgments such as relevance feedback into account to reformulate the initial query [Salton 90]. Although we do not reject this attractive proposition, our objective is to evaluate the effectiveness of the initial search; therefore, relevance feedback can be used after this first search to enhance the retrieval effectiveness.

Acknowledgments

This research was supported by the SNFSR (Swiss National Foundation for Scientific Research) under grant 21-37'345.93 and under grant SNFSR SPP 5NE3-33498. The authors would also like to thank Mr. J. Cavadini, former Minister of Education of the Canton of Neuchâtel, for his support in purchasing the needed hardware without which this experiment could not be possible. The authors also thank M. Choquette from University of Montreal for his help in using the SMART system.

Statistics \ Specification	UNINE1	UNINE2
Retrieved:	50000	50000
Relevant:	3913	3913
Relevant and retrieved:	3191	3191
Interpolated Recall - Precision		
at 0.00	80.25	81.07 (+1.0)
at 0.10	62.97	62.72 (-0.4)
at 0.20	54.23	53.89 (-0.6)
at 0.30	43.92	44.28 (+0.8)
at 0.40	36.81	37.99 (+3.2)
at 0.50	30.52	32.51 (+6.5)
at 0.60	25.13	26.88 (+6.9)
at 0.70	19.14	20.82 (+8.8)
at 0.80	13.40	15.20 (+13.4)
at 0.90	7.05	8.45 (+19.9)
at 1.00	1.23	1.29 (+4.9)
non-interpolated average precision	31.90	32.79 (+2.8)
Precision:		
at 5 docs:	52.0	52.0 (0.0)
at 10 docs:	47.4	49.0 (+3.4)
at 15 docs:	45.6	45.87 (+0.6)
at 20 docs:	43.2	43.6 (+0.9)
at 30 docs:	40.4	40.67 (+0.7)
at 100 docs:	28.06	28.36 (+1.1)
at 200 docs:	20.43	20.63 (+1.0)
at 500 docs:	11.1	11.16 (+0.5)
at 1000 docs:	6.38	6.38 (0.0)
R-Precision (precision after R docs ret), Exact:	34.42	35.02 (+1.7)

Table 6: Official Evaluation of Vector-Space Model (UNINE1) vs. Including Relevance Links (UNINE2)

References

[Blair 90] D. C. Blair: Language and Representation in Information Retrieval. Elsevier, Amsterdam (Holland), 1990.

[Brauen 71] T. Brauen: Document Vector Modifications. in The SMART Retrieval System - Experiments in Automatic Document Processing, G. Salton (Ed.), Prentice-Hall Inc., Englewood Cliffs, NJ, 1971, 456-484.

[Bush 45] V. Bush: As we may Think. Atlantic Monthly, 176(1), 1945, 101-108.

[Cooper 92] W. Cooper, F. C. Grey, D. P. Gabney: Probabilistic Retrieval Based on Staged Logistic Regression Proceedings ACM-SIGIR'92, Copenhague (DK), June 1992, 198-210.

[Fox 83] E. A. Fox: Characterization of Two Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts. Cornell University, Department of Computer Science, Technical Report TR 83-561, September 1983.

[Friedman 71] S. R. Friedman, J. A. Maceyak, S. F. Weiss: A Relevance Feedback System Based on Document Transformation. in The SMART Retrieval System - Experiments in Automatic Document Processing, G. Salton (Ed.), Prentice-Hall Inc., Englewood Cliffs, NJ, 1971, 447-455.

[Fuhr 91] N. Fuhr, C. Buckley: A Probabilistic Learning Approach for Document Indexing. ACM Transactions on Information Systems, 9(3), 1991, 223-248.

[Fuhr 94] N. Fuhr, U. Pfeifer: Probabilistic Information Retrieval as a Combination of Abstraction, Inductive Learning, and Probabilistic Assumptions. ACM Transactions on Information Systems, 12(1), 1994, 92-115.

[Furnas 87] G. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais: The Vocabulary Problem in Human-System Communication. Communications of the ACM, 30(11), 1987, 964-971.

[Gey 94] F. C. Grey: Inferring Probability of Relevance Using the Method of Logistic Regression Proceedings ACM-SIGIR'94, Dublin (IR), July 1994, 222-231.

[Gordon 88] M. Gordon: Probabilistic and Genetic Algorithms for Document Retrieval. Communications of the ACM, 31(10), 1988, 1208-1218.

[Krovetz 92] R. Krovetz, W. B. Croft: Lexical Ambiguity and Information Retrieval. ACM-Transactions on Information Systems, 10(2), 1992, 115-141.

[Kwok 88] K. L. Kwok: On the Use of Bibliographically Related Titles for the Enhancement of Document Representations. Information Processing & Management, 24(2), 1988, 123-131.

[Kwok 90] K. L. Kwok: Experiments with a Component Theory of Probabilistic Information Retrieval Based on Single Terms as Document Components. ACM Transactions on Information Systems, 8(4), 1990, 363-386.

[van Rijsbergen 79] C. J. van Rijsbergen: Information Retrieval. Butterworths, 2nd edition, London (UK), 1979.

[Salton 71] G. Salton (Ed.): The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall Inc., Englewood Cliffs (New Jersey), 1971.

[Salton 90] G. Salton, C. Buckley: Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science, 41(4), 1990, 288-297.

[Savoy 94a] J. Savoy: Ranking Schemes in Hybrid Boolean Systems: A New Approach. ACM Transactions on Information Systems, 1994, accepted with revisions.

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

[Sparck Jones 77] K. Sparck Jones, R. G. Bates: Research on Automatic Indexing 1974-1976. Technical Report, Computer Laboratory, University of Cambridge (England), 1977.

[Stone 74] M. Stone: Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, Series B, 36(2), 1974, 111-147.

[Turtle 91] H. Turtle, W. B. Croft: Evaluation of an Inference Network-Based Retrieval Model. ACM Transactions on Information Systems, 9(3), 1991, 187-222.