Evaluation of Learning Schemes Used in Information Retrieval

Jacques Savoy, Dana Vrajitoru

Institut interfacultaire d'informatique
Université de Neuchâtel
Pierre-à-Mazel 7
CH - 2000 Neuchâtel (Switzerland)

e-mail: {Jacques.Savoy, Dana.Vrajitoru}@seco.unine.ch

Abstract

Searching within the context of information retrieval may be viewed as a communication process between the users and the indexers (or the authors). It is known that in expressing the same concept or idea, different people tend to use different words or phrases, and also that the meaning of words attached to document surrogates tends to change over time. To overcome these phenomena, various learning schemes have been designed so as to automatically infer knowledge about document content from the relevance assessments of past queries. Thus, in contrast to most retrieval models that represent the semantic content of documents as static entities, these adaptive search models might change the descriptions of documents through an inductive learning scheme.

The evaluation of such dynamic document space strategies may be based on retrospective tests within which the same set of queries is applied to train and test the system. Based on cross-validation principles, this paper suggests a more "honest" evaluation methodology that demonstrates a radically different perspective relative to retrieval effectiveness measures used by current information retrieval learning schemes. Moreover, this paper demonstrates that the size of the test-collection might also lead to misleading conclusions about the search performance of a retrieval scheme.

Keywords

Learning schemes, document space modification, genetic algorithm, retrospective evaluation, cross-validation and leaving-one-out methodology.

Introduction

Evaluation of current retrieval models has shown that their retrieval effectiveness is far from perfect. One of the principal explanations of their lack of effectiveness is related to the ambiguity of natural language. This problem involves two facets: first, the same idea or concept may be expressed by various forms (Furnas et al. 1987) or even an infinity number of expressions (Blair 1990), and, second, the same word may have more than one meaning, even in a specialized corpus (Krovetz & Croft 1992).

In order to reduce theses difficulties, Blair (1990) suggests that a retrieval model must have better document content 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 1990, vii)

This may lead to a perception of the retrieval system as an adaptive process, allowing better communication between the searcher and the indexer (Blair 1990, Chapter 4). From a practical point of view, one feasible approach is to design learning schemes that take into account the knowledge obtained from past queries, or more precisely, from their relevance judgments, in order to enhance the retrieval effectiveness over time (Ide & Salton 1971; Brauen 1971; Friedman et al. 1971). We also believe that documents found relevant for a given request do share similar concepts (Savoy 1994). Past queries and their relevance assessments may thus provide useful information about the meaning of documents and must be used to modify document surrogates. Other learning paradigms have recently been proposed for improving information retrieval, including neural network, genetic programming, memory-based reasoning or symbolic learning (Michie et al. 1994; Chen 1995; Schütze et al. 1995).

The main purpose of this study is to compare how various learning schemes function with larger test-collections (e.g., the CACM corpus containing 3,204 documents) in contrast to those used in previous related reports, where the evaluation was based on toy-sized examples (around 20 documents) or small test-collections (around 400 documents).

> "In our preliminary investigation, all three machine learning algorithms performed satisfactorily for small document samples, but the effect of the sample size needs to be examined more carefully." (Chen 1995, p. 212)

As a second goal, we will demonstrate that in information retrieval, retrospective evaluation is an inappropriate evaluation method for learning schemes.

This paper is then made up of three sections. The first describes various learning schemes and presents the corresponding retrospective evaluations showing very attractive retrieval performances. The second section explains an alternate and more "honest" evaluation methodology based on the cross-validation paradigm that induces a different view of the retrieval effectiveness of these learning strategies. The third section presents a query-by-query analysis and explains in more details the learning effects over some document descriptions.

1. Learning Scheme

The aim of an adaptive 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 experience.

In information retrieval, we believe that a learning scheme, able to automatically acquire knowledge from examples, must be 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 non-relevant for a given request.

On the one hand, a learning scheme is based exclusively on successes, i.e., on the presence of retrieved and relevant documents. On the other hand, such an adaptive strategy does not take into account the shared presence of retrieved and non-relevant items. Non-relevant records retrieved by the system are those documents that have at least one common keyword with the request. Such keyword matching does not however always imply word sense matching:

> "Word sense mismatches are far more likely to appear in nonrelevant documents than in those that are relevant." (Krovetz & Croft 1992, p. 139).

We believe 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 nearest city. This assumption is confirmed by relevance feedback studies demonstrating that positive relevance information depicts more valuable information than does negative (Salton & Buckley 1990). The study by Aalbersberg (1992) shows that taking negative relevance information into account may increase the overall performance by around 1.5%, a non-significant improvement. However, the relevance feedback approach considers only relevant data provided by the current request, and thus it can be viewed as a transient learning scheme.

We thus believe that "if a set of items is relevant for a given request, they will tend to be relevant again for other queries" and the learning scheme must help the system to retrieve them. Thus, as a result of the learning stage, documents will have more similar surrogates because they have been found relevant to the same request(s). Moreover, since some documents can be found relevant for different requests, we suggest not modifying the weights of indexing terms present only in the surrogate and not in the request.

The first subsection describes some theoretical justifications for considering learning schemes in information research. The second describes a vector space model to be used as a baseline solution for objectively comparing various learning strategies. The third describes and evaluates some adaptive schemes which are based only on positive relevance feedback, while learning schemes using both positive and negative relevance information are presented in the fourth section. The fifth describes an application of the genetic programming approach as a learning scheme in information retrieval. The sixth describes the main features of related researches, and the last section depicts and evaluates the relevance feedback approach as a method for transient leaning.

1.1. Motivations for Considering an Inductive Learning Scheme

Considering document space as a dynamic component is motivated by several reasons. First, the manual (or automatic) indexing theory is based on the following two implicit assumptions:

"1. That a subject description has a clearly definable meaning; and,

2. That a document has a single subject which can be represented by a small number of correctly assigned subject descriptions." (Blair 1990, p. 169)

but these tacit hypothesis seem to be erroneous (Blair 1990). Assigning appropriate indexing terms must facilitate the communication between searchers trying to describe the information they need and indexers attempting to describe the content of documents. However, as demonstrated by Furnas et al. (1987), the probability that two persons select the same term to mean the same thing is less than 0.20, and studies on the consistency in indexing (Zunde & Dexter 1969; Markey 1984; Cleverdon 1984) or on formulating requests confirm this finding (Iivonen 1995). Given that the retrieval task is a communication process, the system must have the opportunity to adapt itself, which means permanently modifying document representatives in order to enhance further searches. More precisely, this learning strategy might include the opportunity to add search terms in order to increase the document surrogates and thus to capture the domain knowledge of the searchers and information specialists.

Second, when documents representatives are based on words or phrases, they may be vulnerable to scientific and technological obsolescence. Changes in scientific terminology may lead to the emergence of new terms or phrases and a currently good document surrogate may be inappropriate five or ten years later.

Third, allowing indexing weights to fluctuate encourages the system not to depend on an automatic indexing assumption (which can be inaccurate in some circumstances). Assuming that most important descriptors have a high within-document frequency (the tf component) and / or a low with-document frequency (the idf component) may be questionable for some indexing decisions.

Fourth, the search system might permanently modify pertinent document representatives in attempting to associate them more closely, only to satisfy a group of users submitting a set of roughly similar requests. Thus, after analyzing some of these queries, the system may alter the document surrogates for retrieving new related requests more easily. Such an assumption is valid if the relevance assessments between users are consistent and relatively stable (Harman 1995b). For Harter (1996) however, variations in relevance judgments that reflect psychological, situational and subjective conditions seem to be the norm rather than the exception.

Fifth, a static document space can be compared to the concept of a traditional library within which books are assigned to a specific and permanent location. By contrast, a dynamic document space allows the system to reorder related records closer to each other in response to subjective user opinion, expressed through various search keywords or phrases.

Finally, we could conclude that dissimilar requests may express the same information need and that similar queries may represent different user information needs. On the basis of search terms agreement, such a finding seems to condemn a learning approach. Iivonen (1995) however, has demonstrated that even if the level of agreement among researchers in the selection of search keywords is very low (around 31%), the concept-consistency level among searchers is relatively high (around 87%), and particularly high among experienced users (see also (Saracevic & Kantor 1988)).

> "Consistency in the selection of search terms and consistency in the selection of search concepts are clearly two different phenomena." (Iivonen 1995, p. 150).

Furthermore, dissimilar documents can be relevant for the same request, and the underlying action of an inductive learning strategy consists of redescribing co-relevant document surrogates in order to increase their similarity over time (Gordon 1991).

1.2. Baseline Solution

In this study, we have chosen the vector space model which uses the cosine similarity measure as the leading search strategy (Salton 1971). We have adopted this search model because it is related to various studies dealing with inductive learning. To automatically build a vector for each document and query, the system eliminates common words appearing in the text and removes the suffix of each remaining single-term, based on Porter's algorithm. Finally, to assign a weight w_{ij} which measures the relative importance of each single-term t_j in a document D_i, the system applies the following formula:

$$w_{ij} = ntf_{ij} \cdot nidf_j \qquad \text{where } ntf_{ij} = \frac{tf_{ij}}{\max tf_i} \qquad \text{and } nidf_j = \log\left[\frac{n}{df_j}\right] / \log(n) \tag{1}$$

where tf_{ij} is the frequency of the term t_j in the document D_i, n the number of documents D_i in the collection, and df_j the number of documents in which t_j occurs. The outcome of this procedure is the normalization of each weight w_{ij} between 0 and 1. In the same vein, the notation w_{qj} indicates the weight of term t_j in the current query.

To evaluate a retrieval strategy, the system computes the average precision at eleven standard recall values (van Rijsbergen 1979, Chapter 7). Moreover, to decide whether a 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 & Bates 1977, p. A25). This decision rule tacitly assumes that it is easier to provide significant enhancement to a retrieval scheme having poor retrieval effectiveness than to a search strategy having relatively high retrieval performance. In this spirit and based on previous studies (Salton & Buckley 1988), we believe it is important to evaluate and compare search strategies using a baseline set providing satisfactory retrieval effectiveness, which is the case in this paper.

As demonstrated by other studies (Tague-Sutcliffe & Blustein 1994), we believe that average precision for eleven standard recall values represents a good retrieval effectiveness measure. However, the ad hoc decision rule can be replaced by more a sophisticated statistical approach when the corresponding decision is viewed as critical or very important (Savoy 1997).

As a final remark, it is interesting to note that when building a test-collection for evaluating search algorithms, we try (consciously or not) to write queries for which the relevance judgments are as dissimilar as possible. Such a practice mirrors the designer's wishes that underlying requests must represent a broad range of the different concepts contained in the corpus. When designing an adaptive retrieval mechanism, we formulated a contradictory hypothesis which should hold in commercial retrieval services or, at least, we believe so.

1.3. Positive Document Space Modification

In order to permanently store positive relevance feedback information, various researchers have suggested the modification of document surrogates. In this vein, Model 1 of Ide & Salton (1971) proposes altering the representative of all relevant documents D_i in order to obtain a "better" document surrogate D'_i, according to the following formula :

$$D'_i = \begin{cases} (1 - \alpha) \cdot D_i + \alpha \cdot Q & \text{if } D_i \text{ is relevant} \\ D_i & \text{otherwise} \end{cases} \tag{2}$$

in which the unknown constant α is defined such that $0 • \alpha • 1$. The main basic principle of this formulation is to assign search terms not contained in a document surrogate with a weight of $\alpha \cdot w_{qj}$. If a keyword appears both in a relevant record and in the request, its weight is modified by $(1 - \alpha) \cdot w_{ij} + \alpha \cdot w_{qj}$.

From our point of view, a query is an imprecise and incomplete description of the user's information need. We thus suggest not reducing the weight of indexing terms that appear only in the relevant records. Furthermore, when the system modifies the indexing term weights for a relevant document, we believe that this change should be greater for relevant documents appearing in the bottom of the retrieved list than for those near the top. To achieve this purpose, we denote by D_{top} the first record returned by the system (relevant or not), and by $\text{SIM}(D_{top}, Q)$ its similarity with the current request Q. We therefore suggest modifying the relevant document vector as follows:

$$w'_{ij} = \begin{cases} w_{ij} + \alpha \cdot w_{qj} + \beta \cdot (\text{SIM}(D_{top}, Q) - \text{SIM}(D_i, Q)) & \text{if } D_i \text{ is relevant} \\ w_{ij} & \text{otherwise} \end{cases} \tag{3}$$

In a related work, Brauen (1971) also suggests transforming only relevant document surrogates. Moreover, contrary to previous approaches, a distinction is made between the weighting scheme and the selection procedure. When implementing such an inductive 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 : the system must add the "synonym" t_j to the corresponding document surrogate D_i with a corresponding weight β;
2) a concept t_j is present in both the request 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.

These considerations can be formalized as follows:

$$w'_{ij} = \begin{cases} \beta & \text{if } w_{qj} > 0 \text{ and } w_{ij} = 0 \\ w_{ij} + \gamma \cdot (1 - w_{ij}) & \text{if } w_{qj} > 0 \text{ and } w_{ij} > 0 \\ \delta \cdot w_{ij} & \text{if } w_{qj} = 0 \text{ and } w_{ij} > 0 \end{cases} \quad (4)$$

in which the following inequalities must be respected $0 < \beta \bullet 1$, $0 \bullet \gamma \bullet 1$ and $0 < \delta \bullet 1$.

The retrieval evaluations of the previously described learning schemes are based on a relatively small test-collection, more precisely on the CRANFIELD corpus (425 documents, 155 queries). To evaluate these learning schemes on larger corpora, we have used the CACM collection (3,204 articles from the journal "Communications of the ACM", and 50 queries), and the CISI collection containing 35 queries and 1,460 documents extracted from the information science literature (Fox 1983). The results depicted in the following table represent the best values obtained after trying a set of different values for each parameter.

	Precision (% change)	
Model \ Collection	CACM 50 queries	CISI 35 queries
Vector-Space Model cosine, D = ntf · nidf, Q = ntf · nidf	32.56	20.12
Ide & Salton (1971), Model 1, Eq. 2 CACM: $\alpha = 0.6$; CISI: $\alpha = 0.4$	95.43 (+193.03)	90.92 (+352.00)
Eq. 3 CACM: α=0.6, β=0.5; CISI: α=0.6, β=0.5	98.85 (+203.55)	89.68 (+345.79)
Brauen (1971), Eq. 4, $\delta = 0.9$ CACM: $\gamma = \beta = 0.4$; CISI: $\gamma = \beta = 0.6$	97.90 (+200.67)	90.06 (+347.68)

Table 1: Retrospective evaluation of positive document modification

From Table 1, all learning strategies provide very attractive performance results, significantly enhancing retrieval effectiveness compared to the baseline solution. We cannot select the best learning scheme from these results because when a given strategy performs well for one test-collection, its performance is inferior for the second corpus.

1.4. Document Space Modification

Model 2 from Ide & Salton (1971) suggests modifying the surrogates associated with both relevant and nonrelevant records. This modification is intended to move nonrelevant records further away from relevant ones. For this model, the following procedure is applied:

1) If, for all relevant documents D_i, and for all nonrelevant documents D_k, the inequality $\mathrm{SIM}(D_i, Q) > \mathrm{SIM}(D_k, Q) + \theta$ is respected, then no adjustment is made;
2) Otherwise, if there exists a nonrelevant document D_k not yet adjusted, and for which $\mathrm{SIM}(D_k, Q) + \theta \bullet \mathrm{SIM}(D_i, Q)$ holds, then

$$\begin{aligned} D'_i &= (1 - \alpha) \cdot D_i + \alpha \cdot Q \\ D'_k &= (1 - \alpha) \cdot D_k - \alpha \cdot Q \end{aligned} \tag{5}$$

3) If no nonrelevant document D_k exists which has not been previously adjusted, modifications on all relevant documents D_i are still performed according to Equation 5.

Friedman et al. (1971) describe 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 the user's judgments about document content. This proposition is based on three principles. The first specifies that the modification of indexing term weights can occur only for the "good" keywords or for those appearing more frequently in the relevant or in the nonrelevant documents. As a second principle, Friedman et al. (1971) suggest deriving the new indexing term weight w'_{ij} in proportion to the existing weight w_{ij}. Thirdly, these 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 set of nonrelevant records (a more detailed explanation can be found in the Appendix 1). This learning strategy can be expressed using the following formula:

$$w'_{ij} = \begin{cases} w_{ij} \cdot (1 + T_j) & \text{if } D_i \text{ is relevant} \\ 0 & \text{if } D_i \text{ is nonrelevant} \end{cases} \tag{6}$$

which implies a radical change for nonrelevant documents.

Based on the ADI collection (82 documents, 11 queries) used for Friedman's experiment, and the CRANFIELD corpus (425 documents, 155 queries) for the study done by Ide & Salton (1971), retrieval evaluations from these learning schemes have resulted in improved retrieval effectiveness. Our results, based on the CACM and CISI collections are depicted in Table 2, demonstrate also significant enhancement over the baseline.

Model \ Collection	Precision (% change) CACM 50 queries	Precision (% change) CISI 35 queries
Vector-Space Model cosine, D = ntf · nidf, Q = ntf · nidf	32.56	20.12
Ide & Salton (1971), Model 2, Eq. 5 CACM: α=0.3, θ=0; CISI: α=0.1, θ=0 CACM: α=0.3, θ=0.25; CISI: α=0.1, θ=0.15	 86.27 (+164.20) 86.87 (+166.76)	 72.97 (+262.27) 72.98 (+262.67)
Friedman et al. (1971), Eq. 6 CACM: $\delta = 0.05$, $\alpha_1 = \alpha_2 = 1.2$; CISI: $\delta = 0.01$, $\alpha_1 = \alpha_2 = 0.2$	 93.31 (+186.58) 	 46.74 (+132.32)
Genetic algorithm based on (Gordon 1988) 4 descriptors per doc., 10 generations	 32.78 (+0.68)	 19.83 (-1.44)

Table 2: Retrospective evaluation of document space modification schemes

1.5. Genetic Algorithm

From a different perspective, Gordon (1988) suggests a learning scheme based on a genetic algorithm (Goldberg 1989; Koza 1992) to enhance retrieval effectiveness (see also (Blair 1990; Chen 1995)). In this approach, each document is described by various surrogates obtained using various binary indexing policies (e.g., based on document titles and/or abstracts, 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 other ones; and (2) cross-over between pairs of surrogates which generate variety among descriptors. A mutation operation or the random modification of some descriptor characteristics is also sometimes used (Yang et al. 1993) but this operation may be considered as having marginal importance (Goldberg 1989). Moreover, the selection, cross-over and mutation operations are not clearly defined, and various algorithms have been suggested (Goldberg 1989; Koza 1992). Finally, precise result from a given genetic-based system depends on the number of variants, the number of generations, and the underlying parameters of the reproduction, cross-over and mutation schemes.

To evaluate Gordon's scheme on a scale larger than in previous studies (e.g., 18 documents in (Gordon 1988)), we have used various logical sections of a document to build four competing descriptors. The construction of the first descriptor is obtained using all logical sections like all other evaluations reported in this study. Additional surrogates have been created using only the title as the second surrogate, the CR (Computing Review Abstract) as the third, and finally the keywords given by the author(s) as the last descriptor. Contrary to Gordon's study based on a binary indexing strategy, our evaluation is grounded on a weighted indexing scheme. The retrospective evaluation of this scheme depicted in Table 2 does not demonstrate any enhancement over the baseline.

1.6. Other Related Researches in Permanent Document Modification

Of course, other learning strategies have already been proposed and evaluated, and most of them are directly related to the probabilistic retrieval model (Kwok 1990). Some current probabilistic retrieval models (Gey 1994; Fuhr & Pfeifer 1994) do not consider directly the terms or phrases existing in document and query surrogates, but their features do infer a probability of relevance based on logistic regression. Such features can be the absolute and relative within document or within request frequency, the inverse document frequency, the document length, the number of terms a record and a query have in common, etc. Experimental results have shown attractive retrieval effectiveness. In such retrieval schemes, a permanent modification means that the system must from time to time recompute the corresponding coefficients of the logistic regression.

Based on another paradigm, the neural network model is also suggested as a method for permanently modifying document representation (e.g., in the AIR system (Belew 1989)), and especially within the probabilistic model (Kwok 1995).

In order to represent the information given by the previous experiments or requests, Savoy (1994) has designed a special interdocument relationship called the relevance link. This link type connects two documents found relevant for a given query. To account for these document relationship, the suggested retrieval scheme is based on the spreading activation principle. In a related vein, Gordon (1991) shows that relevance assessments can be a valid source of information to form clusters of co-relevant documents. In this study, a genetic algorithm is used to adaptively redescribe documents in order to obtain increasingly similar descriptions.

1.7. Relevance Feedback

Because this approach makes only a transient use of relevance information, we will call this scheme a transient inductive learning scheme (Rocchio 1971), (Salton & Buckley 1990). In this study, as a baseline for further comparisons, we have evaluated the "Ide dec-hi" relevance feedback model (Ide 1971) within which a new query noted Q_{k+1} is based on the previous query Q_k plus the weight of all indexing terms of all found and relevant documents D_i, for i = 1, 2, ... rel, minus the weight of the indexing terms of the top most irrelevant document ($D_{topnonrel}$).

$$Q_{k+1} = Q_k + \sum_{i=1}^{rel} D_i - D_{topnonrel} \tag{7}$$

Such a relevance feedback scheme seems to perform slightly better than other suggested strategies (Salton & Buckley 1990). The retrospective evaluation of this algorithm is depicted in Table 3a illustrating a very attractive retrieval performance.

	Precision (% change)	
Model \ Collection	CACM 50 queries	CISI 35 queries
Vector-Space Model cosine, D = ntf · nidf, Q = ntf · nidf	32.56	20.12
Relevance given for the first 5 top documents	43.50 (+33.44)	26.12 (+29.84)
10 top documents	47.95 (+47.25)	28.60 (+42.16)
15 top documents	51.23 (+57.31)	30.88 (+53.53)
20 top documents	54.00 (+65.84)	33.06 (+64.37)
30 top documents	58.77 (+80.50)	36.35 (+80.67)

Table 3a: Retrospective evaluation of relevance feedback

However, when comparing search strategies with and without relevance feedback, it is known that the comparison must be made without considering the retrieved documents (relevant or not) for which a relevance assessment has been obtained (Chang et al. 1971). Thus, in the results depicted in Table 3b, all records previously seen by the user are removed from the collection, leading to a retrieval effectiveness measure based on a residual collection. Furthermore, queries with all relevance records in the top 30 retrieved items are removed from the evaluation. The data in Table 3b confirms the enhancement found in Table 3a and seems to indicate that the retrospective evaluation may be a valid evaluation methodology. However, as shown in the second section, this first impression is erroneous.

Of course, variations on the previously described relevance feedback scheme have been suggested. For example, Robertson (1990) suggests adopting different procedures for the selection of new search terms and for the weighting of search keywords, see also (Buckley & Salton 1995). For Harman (1992), the number of terms to be included in the expanded request can depend on the average document size of the underlying collection, and Allen (1995) seems to confirm this finding. For Aalbergsberg (1992), the number of documents shown to the user must not be an arbitrary number (e.g., 20 or 30 as before), but limited to one at a time, and such a strategy simplifies the design of the man-machine interface.

	Precision (% change)	
Model \ Collection	CACM 46 queries	CISI 35 queries
Without the first 5 top documents	19.34	17.69
Relevance for the first 5 top doc.	25.95 (+34.19)	22.31 (+26.16)
Without the first 10 top documents	16.31	16.70
Relevance for the first 10 top doc.	23.24 (+48.62)	22.71 (+36.02)
Without the first 15 top documents	15.57	16.34
Relevance for the first 15 top doc.	23.83 (+53.08)	23.24 (+42.28)
Without the first 20 top documents	13.48	14.62
Relevance for the first 20 top doc.	22.10 (+64.01)	22.83 (+56.14)
Without the first 30 top documents	11.70	13.07
Relevance for the first 30 top doc.	21.87 (+86.95)	23.73 (+81.62)

Table 3b: Evaluation of relevance feedback (residual collection)

In a related study, Yang et al. (1993) suggest using a genetic algorithm in order to automatically modify the weights assigned to search keywords. In this approach, the document representatives are binary weighted and are never altered by the learning scheme. To generate query variants, the proposed system duplicates the initial search terms but weights them randomly (normally distributed around the initial weights). During the learning stage, the search term weights of query vector variants are modified through (1) the reproduction of the best request descriptions; (2) the interchange of weights between competing query vectors (cross-over), and (3) the incorporation of new weights by mutation. Thus, in this inductive learning strategy no new search keywords are introduced and no terms are deleted. The improvement factor of this learning technique is unknown because no comparison is reported by Yang et al. (1993).

2. Evaluation Methodology

Retrospective evaluation returns retrieval effectiveness values that are too optimistic (biased high), reflecting an unrealistic situation. However, some theoretical justifications have been proposed to validate such an evaluation methodology (e.g., Kwok 1990). The results obtained under these circumstances represent an upper bound of the retrieval effectiveness. This can be achieved when a large number of experiments have been done and the content of the document is static.

In our point of view, this approach may be viewed as being of academic interest only, and the retrieval performance obtained under such conditions is *too good to be true*. For example, are the results of Table 1 really valid as a predictive performance? The comparison of the retrospective vs. residual evaluation in Tables 3 seems to indicate that the retrospective test is not really bad, maybe just a little too optimistic. However, can this example justify the application of the retrospective evaluation?

Moreover, when the retrospective 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 that it has the capability to generalize given information, to retain the main features of the given information and to find useful relationships between data.

2.1. Predictive Evaluation

If the evaluation results under retrospective tests are usually misleading, a 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. In both cases, 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 where a system ignores learning.

To take account of these criteria, the train-and-test (or hold-out) method suggests splitting the query 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. For example, Ide & Salton (1971) have used 125 queries for training purposes and the remaining 30 requests to test their learning strategy. Not all queries however can be exploited both in the learning scheme and during the evaluation, and a direct comparison with results obtained with the entire query set is not possible.

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 1974; Geisser 1975; Stone 1977). More precisely, the leaving-one-out approach, a special case of the cross-validation method, represents a solution which works as follows. The request sample R of size r is divided into r sets, each containing r-1 queries. In the k^{th} set, one can find all the 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 as described in the following equation:

$$AP_{lv} = \frac{1}{r} \cdot \sum_{k=1}^{r} AP_k \tag{8}$$

in which AP_k denotes the average precision at eleven standard recall values for the Q_k query under the condition that the learning scheme is fitted using the query set R 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 the learning stage. Moreover, the entire (small) request sample is used both in the evaluation and in the training stage.

2.2. Leaving-one-out Evaluation

From the results depicted in Table 1, we may conclude that all positive document modification strategies represent very significant improvement over search model that ignores learning. Moreover, with the performances reported, it was not clear which learning scheme represents, for both test-collections, the best schemes. When applying the leaving-one-out approach, the results shown in Table 4a provide another outlook regarding these learning strategies.

First, none of these learning schemes represents a significant enhancement when using both test-collections simultaneously. This conclusion is really different than the conclusion which can be drawn from the retrospective evaluation ("use a whatever-you-want learning scheme").

Regarding the document space modification strategies which include both positive and negative feedback information, we have found, in our retrospective evaluation, that these approaches present very attractive results. The leaving-one-out evaluation shown in Table 4b only partly confirms this finding. In fact, Table 4b shows that Ide & Salton's Model 2 presents a significant enhancement for both test-collections based however on different parameter settings.

	Precision (% change)	
Model \ Collection	CACM 50 queries	CISI 35 queries
Vector-Space Model cosine, D = ntf · nidf, Q = ntf · nidf	32.56	20.12
Ide & Salton (1971), Model 1, Eq. 2		
CACM: $\alpha = 0.3$; CISI: $\alpha = 0.08$	35.40 (+8.72)	20.83 (+3.55)
CACM: $\alpha = 0.4$; CISI: $\alpha = 0.1$	35.29 (+8.38)	20.92 (+3.99)
Eq. 3		
CACM: α=0.3, β=0.2; CISI: α=0.06, β=0.03	39.38 (+20.95)	20.91 (+3.91)
CACM: α=0.4, β=0.2; CISI: α=0.08, β=0.02	38.74 (+18.99)	20.79 (+3.31)
Brauen (1971), Eq. 4		
CACM: γ=β=0.15, δ=0.9; CISI: γ=β=0.05, δ=0.9	35.46 (+8.89)	20.36 (+1.20)
CACM: ; γ=β=0.1, δ=1.0; CISI: γ=β=0.05, δ=1.0	36.65 (+12.57)	19.51 (-3.02)

Table 4a: Leaving-one-out evaluation of positive document modification

Since the genetic algorithm suggested by Gordon (1988) does not present any improvement over the baseline when considering retrospective evaluation (see Table 2), the leaving-one-out approach was not tried for this scheme.

Model \ Collection	Precision (% change) CACM 50 queries	CISI 35 queries
Vector-Space Model cosine, D = ntf · nidf, Q = ntf · nidf	32.56	20.12
Ide & Salton (1971), Model 2, Eq. 5 CACM: α=0.3, θ=0; CISI: α=0.1, θ=0	35.56 (+9.19)	22.32 (+11.93)
queries presented in natural order CACM: α=0.3, θ=0.25; CISI: α=0.1, θ=0.15	36.10 (+10.84)	22.29 (+10.82)
queries given in a random order CACM: α=0.3, θ=0.25; CISI: α=0.1, θ=0.15	36.15 (+11.03)	22.33 (+10.98)
Friedman et al. (1971), Eq. 6 CACM: δ = 0.05, $\alpha_1 = \alpha_2$ = 1.2;	38.45 (+18.09)	
CISI: δ = 0.01, $\alpha_1 = \alpha_2$ = 0.2		20.86 (+3.70)

Table 4b: Leaving-one-out evaluation of document space modification schemes

As a last question, we want to know whether query order may play a role during the learning process. To verify this hypothesis, we have carried out a second experiment on Ide & Salton's Model 2, within which the requests are presented to the learning procedure in a random order. The results shown in Table 4b demonstrate clearly that query order does not significantly modify the retrieval performance (CACM: 36.10 vs. 36.15 (+0.14%); CISI: 22.29 vs. 22.33 (+0.18%)).

3. A Closer Look to the Effects of a Learning Scheme

This section presents a more detailed description of the effects of a learning scheme. As an exhaustive study is not possible, we have decided to report on an analysis of only one learning scheme. The search strategy selected is Model 2 suggested by Ide & Salton (1971) which results in significant improvement for both test-collections (see Table 4b).

However, it is known that average precision, like all statistics generalizations, may hide performance irregularities among requests when comparing two retrieval schemes. Thus, a given retrieval strategy may perform very well for some requests but poorly for others (Harman 1995a). The first subsection shows how the mean precision ignores some important differences between queries on the one hand, and on the other, within a given request when looking at the rank of documents.

The second subsection analyses the modification of document surrogates over time. Since we know that a document is relevant for some requests and non-relevant for others, our analysis will include both types of requests.

3.1. Query-by-Query Analysis

We know that the average precision may hide performance irregularities among requests when comparing two retrieval strategies. Thus, in order to obtain a general understanding of this variability, Table 5 presents the distribution of the queries according to the fact that the retrieval effectiveness has been significantly improved (more than 5%), decreases significantly (less than 5%), or has statistically the same performance (± 5%). It is interesting to note that even with a learning scheme having an average enhancement around 10%, we may find some performance degradation for 20% of the requests.

	CACM 50 queries	CISI 35 queries
Number of queries presenting a … significantly positive effect	26 (52%)	22 (62.9%)
significantly negative effect	10 (20%)	7 (20%)
statistically equivalent performance	14 (28%)	6 (11.1%)

Table 5: Distribution of retrieval effectiveness modification among queries

		#relevant	Precision	
	Query	docum.	before	after
CACM #17	Optimization of intermediate and machine code	13	18.22	44.97 (+146.8%)
CACM #10	Parallel languages; languages for parallel computation	28	62.57	42.23 (-32.5%)
CISI #4	Image recognition and any other methods of automatically transforming printed text into computer-ready form	10	12.38	25.04 (+102.3%)
CISI #13	What criteria have been developed for the objective evaluation of information retrieval and dissemination systems	8	35.99	26.50 (-26.4%)

Table 6: Contents and characteristics of selected queries

In order to obtain more detailed information, it may be interesting to look closely at some requests that represent either positive or negative performance improvements. Within a single request, it may be useful to verify whether or not, the learning scheme will produce the same effect on all relevant documents. To investigate this question, Table 6 shows the selected queries and their characteristics.

Figure 1 and 2 are examples of pictures worth a thousand words, depicting for two requests, the position of each relevant item in the resulting list (or more precisely, the natural logarithm of the rank) both before and after the learning procedure. In these figures, the relevant records are sorted according to their identification number. A white circle represents the document position before the learning procedure while a black square indicates the rank obtained after the learning process. For example, in Figure 1, following the learning stage, we see that the description of Document 1613 progresses favorably, promoting this document higher up the result list. By contrast, when looking at another request as shown in Figure 2, the same document exhibits an opposite pattern; Document #1613 obtains a better position before learning than after.

Thus, Figures 1 and 2 illustrate that the learning effects are different over a given set of relevant documents. Moreover, within a given query, the learning effect is beneficial for some relevant items, but not for all. Based on these figures, it may be difficult to predict whether or not the learning scheme will enhance the retrieval performance for the corresponding query. For example, Figure 1 depicts an overall enhancement while Figure 2 shows a performance degradation.

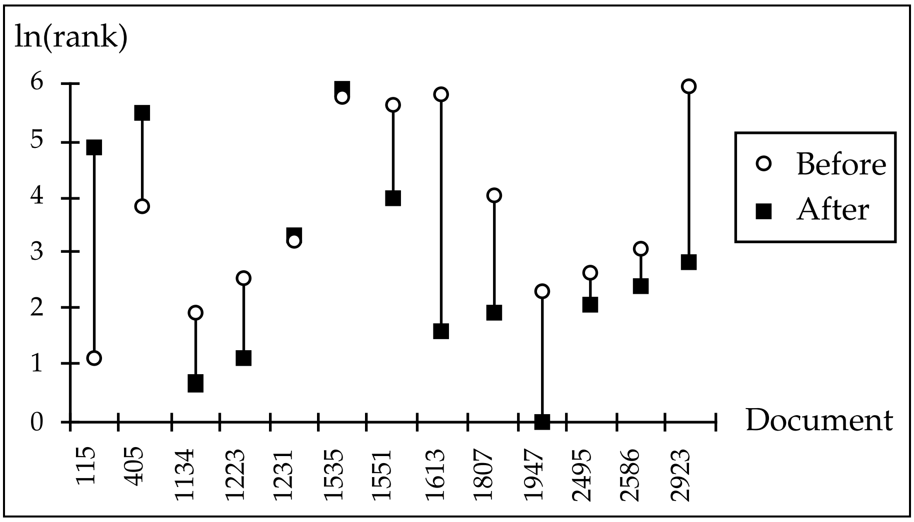

Figure 1: Effects on relevant document rankings of Query #17 from the CACM collection

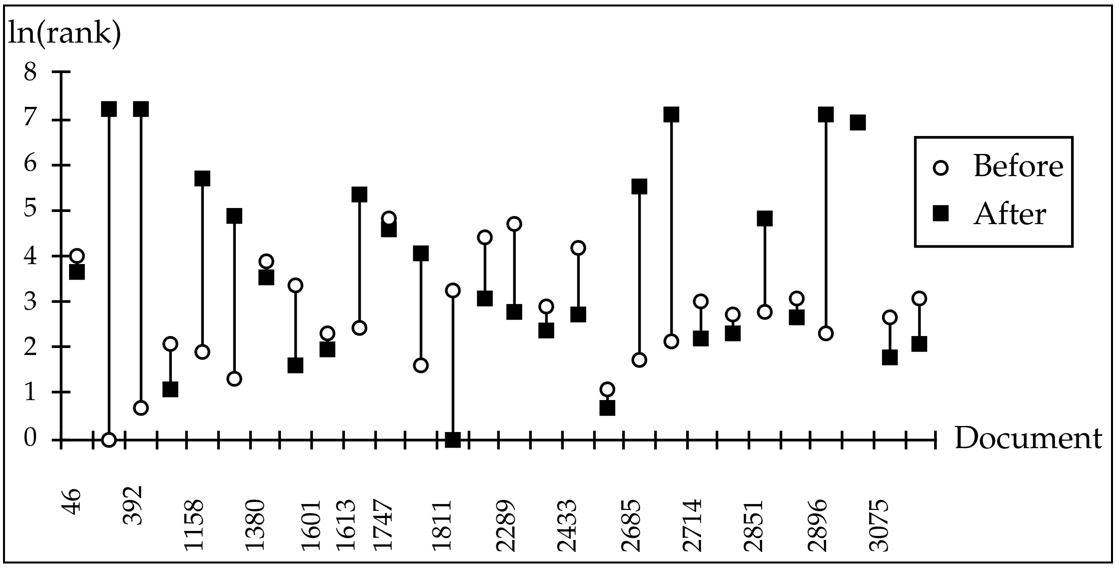

Figure 2: Effects on relevant document rankings of Query #10 from the CACM collection

It is not really useful, however, to study each particular relevant item. After all, if a relevant record at a given rank is replaced by another relevant record, the mean precision will not change. What the user really wants is relevant information and not a particular relevant document.

To obtain better a understanding of the performance modifications, we had to organize our data according to a different point of view. In fact, as long as all relevant documents are interchangeable, it is not important to look at the position achieved by each particular relevant record. Rather, we are essentially concerned with the positioning represented by the relevant items.

Based on this new perspective, we have compared the ranks of the relevant and retrieved documents before and after the learning process. Figures 3 to 6 show these ranking modifications for our four selected queries.

For example, the first point on the left in Figure 3 represents the natural logarithm for the ranking of the first relevant retrieved document. This point corresponds to a recall value of 1/13 (because this query contains 13 relevant items, as indicated in Table 6). Knowing the ranking or the ranking modification, we may compute an approximate precision value associated with this recall point. Moreover, we may obtain a general impression of the improvement produced by the learning scheme.

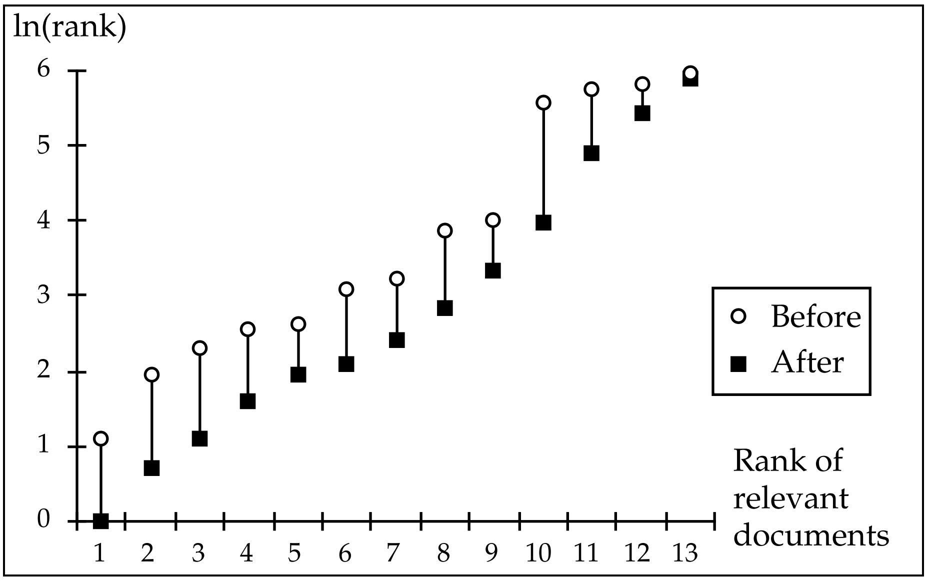

Figure 3: Effects on relevant document rankings of Query #17 from the CACM collection

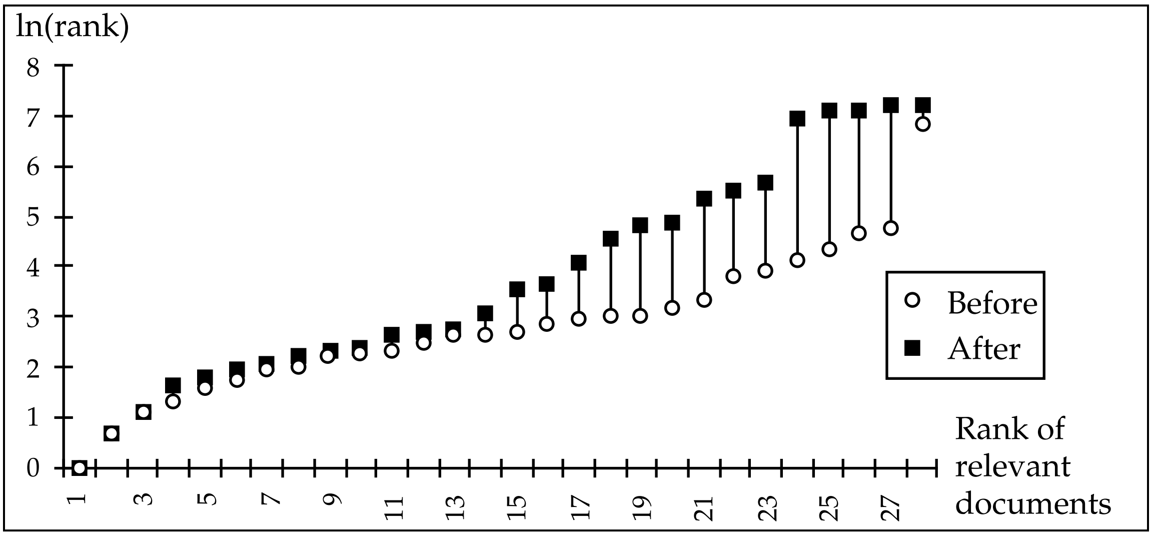

Figure 4: Effects on relevant document rankings of Query #10 from the CACM collection

Figures 3 and 4 show the systematic effect of all relevant documents, respectively positive in Figure 3 and negative in Figure 4. As another example, Figure 5 depicts an overall retrieval enhancement because the relevant documents appear higher in the resulting list. However, this figure also displays some counter-examples when looking at each particular document. Finally, Figure 6 presents the unfavorable case; most of the relevant items obtained a higher rank after the learning stage.

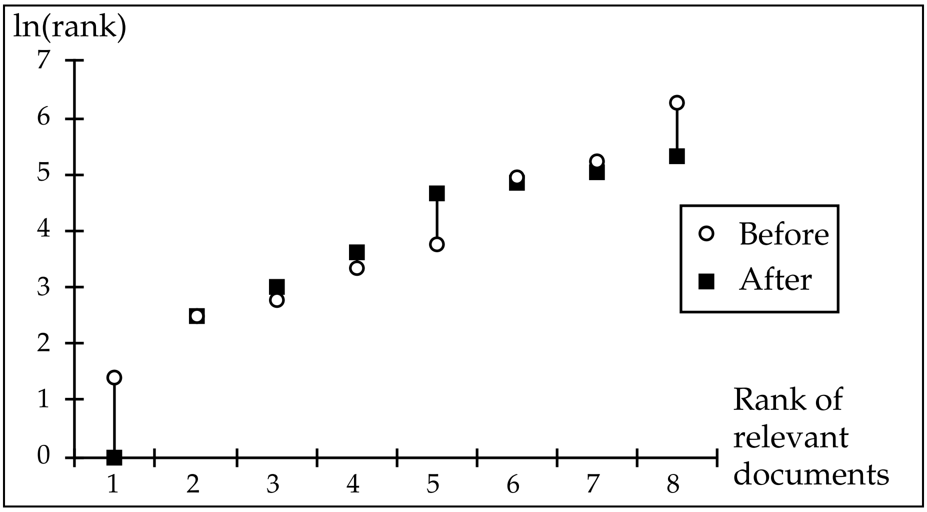

Figure 5: Effects on relevant document rankings of Query #4 from the CISI collection

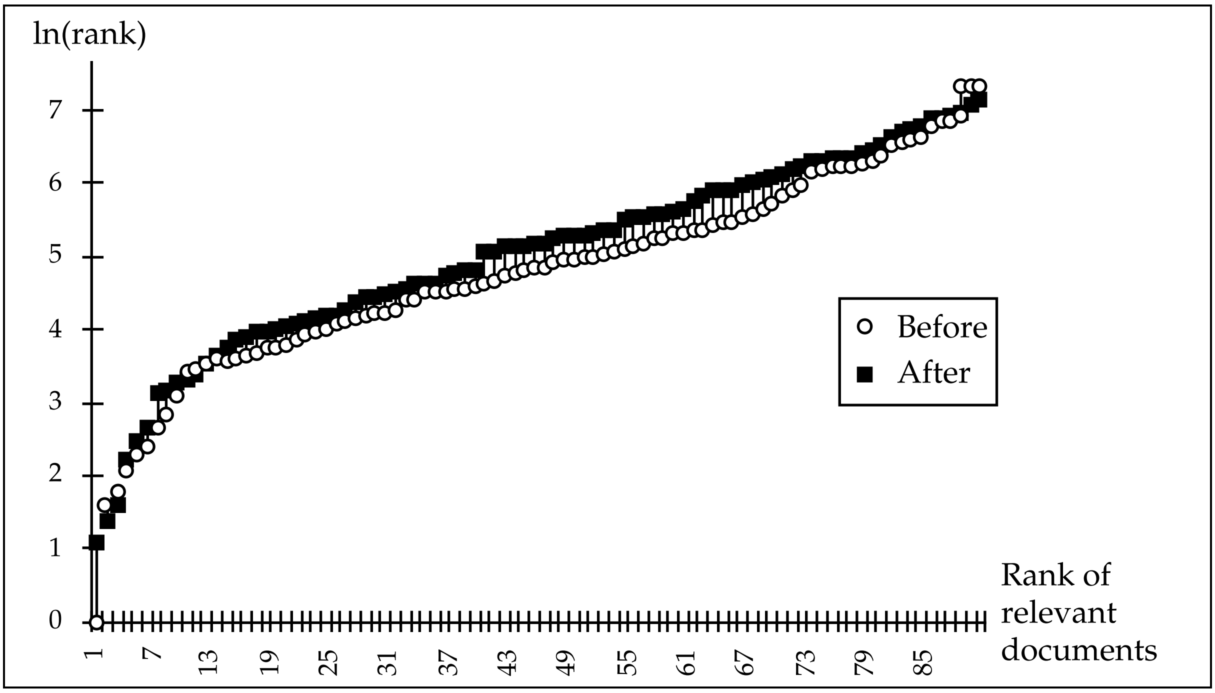

Figure 6: Effects on relevant document rankings of Query #13 from the CISI collection

3.2. Document-by-Document Analysis

Looking at each individual query represents only the first part of our analysis. Documents surrogates change over time in response to queries where they are relevant or not. As a detailed report of all document surrogates is clearly impossible, we have selected four documents extracted from both test-collections (see Table 7).

Collection	Title
CACM #1613	One-Pass Compilation of Arithmetic Expressions for a Parallel Processor
CACM #1795	Optimal Code for Serial and Parallel Computation.
CISI #709	The Chemical Abstracts Service Chemical Registry System. I General Design
CISI #375	Encyclopedia of Information Systems and Services

Table 7: Title of selected documents

These selected records were retrieved by various queries; they are relevant for some requests and non-relevant for other. Following the pattern of the previous subsection, for each collection we have chosen a document showing a modification that suits our expectations, and also a counter-example.

We have tested these four documents against all queries for which they are relevant (statistics depicted in Table 8a) and requests for which they are not relevant (see Table 8b). In each case, we have computed the number of used queries, the mean similarity and mean rank both before and after the learning process.

As shown in Table 8, Document #1613 presents a generally satisfactory transformation; its mean similarity improves when this record is pertinent (Table 8a) leading to a better ranking (96.8 vs. 45.8). However, the learning procedure also tends to improve the mean similarity when this record is non-relevant (Table 8b). This effect however is not really important; the mean rank progresses from 848.7 to 412.8 but these positions are too low in the resulting list to have any real effect on the average precision.

	CACM				CISI			
	#1613		#1795		#709		#375	
number of queries	4	5	3	2	8	8	9	9
	before	after	before	after	before	after	before	after
mean similarity	17.7	19.3	40.1	8.3	5.3	13.0	6.5	5.1
mean rank	96.8	45.8	4.3	268.0	256.1	73.1	146.1	227.6

Table 8a: Statistics on selected documents before and after learning when these records are relevant

If Document #709 of the CISI collection follows the same pattern, both Documents #1795 and #375 exhibit a contrary evolution. The learning process has, on means at least, a negative effect on these document surrogates. For example, when Document #1795 is relevant, the mean similarity decreases leading to a mean rank before learning of 4.3 to a mean rank of 268.0 after learning (Table 8a).

	CACM				CISI			
	#1613		#1795		#709		#375	
number of queries	34	30	17	25	20	17	22	26
	before	after	before	after	before	after	before	after
mean similarity	4.6	5.3	11.0	3.4	2.6	3.9	6.8	7.7
mean rank	848.7	412.8	304.2	865.7	430.1	310.7	197.0	157.0

Table 8b: Statistics on selected documents before and after learning when these records are non-relevant

To conclude this analysis, we have taken a particular and closer look at Document #1613 from the CACM collection, for the following reasons. We already knew that this record was relevant for both Queries #10 and #17 (see Figures 1 to 4), and the content of these requests is depicted in Table 6. Through the learning scheme, Document surrogate #1613 presents a positive modification as a result of Query #17 (see Figure 1) but also a negative evolution when looking at Request #10 (see Figure 2). In order to explain the underlying reasons of this phenomenon, we have reported the required information in Table 9.

Table 9 exhibits the document surrogate modification resulting from various search terms extracted from two requests. Found in the first column are all keywords appearing in Query #10 and #17 from the CACM collection. Since a particular search term may occur in other requests, all numbers written in the same row refer to query numbers having the same search term (e.g., the word «language», included in Query #10, appears also in Requests #3, #18, #36, #39, #40, #11 and #59). However, the document analyzed, #1613 in this case, is not necessarily relevant for all queries. Thus in the second column, we have indicated query numbers for which Document #1613 is also relevant. The third column indicates request numbers for which Document #1613 is retrieved but not relevant. The last column depicts query numbers for which Document #1613 is not retrieved and not relevant.

Looking back at Figure 1, we may see that the description for Document 1613 evolves in a positive way and the data depicted in Table 9 may explain this development. Three keywords («code», «intermediate», «machine») appear only in queries having Document 1613 as a relevant record (Queries #17, #18 and #3). For the search keyword «optimization», the corresponding indexing weight must be reduced because this term occurs mainly in requests for which Document #1613 is judged non-relevant (Queries #13, #14, #45 and #48).

search term of Query #10	also relevant for queries	non-relevant for queries	not retrieved and non-relevant for queries
computation		1, 21, 22, 23, 25, 42, 43, 44, 48, 63	60
language	3, 18	36, 39, 40	11, 59
parallel	18	19, 62, 63	
search term of query #17			
code	18		13
intermediate	3		
machine	18		
optimization		13, 14, 45, 48	15

Table 9: Analysis of Document #1613 from the CACM collection across various queries

The situation is different when looking at Query #10. First, the appropriate modification for the search term «computation» seems clear: the corresponding indexing weight must be reduced because this term appears in 10 other requests, all these requests having Document #1613 as a non-relevant item. Since the same procedure is applied to indexing terms «language» and «parallel», and when considering Query #10, the rank following learning for Document #1613 cannot be higher. In fact, for all search terms, the corresponding indexing weight in this document surrogate has been decreased.

Conclusion

This paper describes various learning schemes used within the vector space model. Such approaches can be viewed as a solution to the communication process between indexers and inquirers who are using dramatically different terminology. Various inductive learning algorithms have been suggested in the past and we have also proposed a learning scheme modifying only the relevant records. Setting the appropriate value for various parameters included in these approaches does not seem however to be collection independent and the specification of these values can be problematic for a new text database.

A retrospective evaluation of these inductive learning strategies leads to very significant improvement in retrieval effectiveness. We suggest however that such evaluation methodology be reconsidered because it uses the same set of requests both during the training of the learning system and during its evaluation.

As an alternate and more honest evaluation methodology, this paper suggests using the cross-validation, or more precisely, the leaving-one-out methodology. On this basis, various learning schemes depict negative results which, in our point of view, do not mean that a learning scheme is inappropriate in information retrieval, but rather that current strategies, at least as implemented in this study, do not permit significant enhancement to retrieval effectiveness. Moreover, this paper demonstrates clearly that encouraging results obtained with a toy-size collection cannot be generalized even for medium size corpora like the CACM or CISI test-collections. Thus, we do not recommend drawing any conclusions about retrieval performance on the basis of toy-sized examples.

Acknowledgments

This research was supported in part by the SNSF (Swiss National Science Foundation) under grants 21-37’345.93 and 20-43’217.95.

Appendix 1: Computation Used by the Friedman et al. (1971) Learning Scheme

For each indexing term t_j, compute:

$$R_j = \frac{\sum_{i=1}^{rel} w_{ij}}{rel} \qquad \text{and } N_j = \frac{\sum_{i=1}^{nonrel} w_{ij}}{nonrel}$$

$$Dis_j = R_j - N_j$$

in which w_{ij} indicates the importance of the indexing term t_j in the document D_i, w_{qj} the importance of the search term t_j in the current request, rel the number of relevant documents, nonrel the number of nonrelevant items, and in which Dis_j measures the discrimination value associated with term t_j.

If $Dis_j > \delta$, the corresponding term t_j can be viewed as a good discriminator for relevant documents, and if $Dis_j < \delta$, the underlying term can be considered as a good discriminator for nonrelevant records. The system may thus select the appropriate terms t_j for which $|Dis_j| > \delta$, and for those terms compute:

$$F_j^1 = \frac{w_{qj}}{\sum_{k=1}^{t} w_{qk}} \qquad F_j^2 = \frac{\sum_{i=1}^{rel} w_{ij}}{\sum_{i=1}^{rel} \sum_{k=1}^{t} w_{ik}} \qquad \text{and } F_j^3 = \frac{\sum_{i=1}^{nonrel} w_{ij}}{\sum_{i=1}^{nonrel} \sum_{k=1}^{t} w_{ik}}$$

within which, for each remaining term t_j, F_j^1 measures its relative importance in the request, F_j^2 its importance in the relevant documents, and F_j^3 its importance in the nonrelevant records.

Finally, for each selected term t_j, compute T_j as follows:

$$T_j = \begin{cases} \alpha_1 \cdot F_j^1 + \alpha_2 \cdot F_j^2 & \text{if } Dis_j > \delta \text{ or if } w_{qj} > 0 \\ -\alpha_2 \cdot F_j^3 & \text{if } Dis_j < \delta \text{ and } w_{qj} = 0 \end{cases}$$

References

Allen, J. (1995). *Relevance feedback with too much data.* Proceedings ACM-SIGIR'95, Seattle, WA, 337-343.

Aalbergsberg, I. J. (1992). *Incremental relevance feedback.* Proceedings ACM-SIGIR'92, Copenhagen, DK, June 1992, 11-22.

Belew, R. K. (1989). *Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents*. Proceedings of the ACM-SIGIR'89, Cambridge, MA, 11-20.

Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam, NL: Elsevier.

Brauen, T. (1971). Document vector modifications. In G. Salton (Ed.), *The SMART retrieval system - experiments in automatic document processing* (pp. 456-484). Englewood Cliffs, NJ: Prentice-Hall.

Buckley, C. & Salton, G. (1995). *Optimization of relevance feedback weights.* Proceedings ACM-SIGIR'95, Seattle, WA, 351-357.

Chang, Y. K., Cirillo, C. & Razon, J. (1971). Evaluation of feedback retrieval using modifying freezing, residual collection, and test and control groups. In G. Salton (Ed.), *The SMART retrieval system - experiments in automatic document processing* (pp. 355-370). Englewood Cliffs, NJ: Prentice-Hall.

Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*. 46(3), 194-216.

Cleverdon, C. W. (1984). Optimizing convenient on-line access to bibliographic databases. *Information Service & Use*. 4, 37-47.

Fox, E. A. (1983). *Characterization of two experimental collections in computer and information science containing textual and bibliographic concepts*. Technical Report TR 83-561, Department of Computer Science, Cornell University.

Friedman, S. R., Maceyak, J. A. & Weiss, S. F. (1971). A relevance feedback system based on document transformation. In G. Salton (Ed.), *The SMART retrieval system - experiments in automatic document processing* (pp. 447-455). Englewood Cliffs, NJ: Prentice-Hall.

Fuhr, N. & Pfeifer, U. (1994). Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions. *ACM Transactions on Information Systems*, 12(1), 92-115.

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.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*. 70(2), 320-328.

Gey, F. C. (1994). *Inferring probability of relevance using the method of logistic regression*. Proceedings of the ACM-SIGIR'94. Dublin, IR, 222-231.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.

Gordon, M. (1988). Probabilistic and genetic algorithms for document retrieval. *Communications of the ACM*. 31(10), 1208-1218.

Gordon, M. (1991). User-based document clustering by redescribing subject descriptions with a genetic algorithm. *Journal of the American Society for Information Science*. 42(5), 311-322.

Harman, D. (1992). *Relevance feedback revised*. Proceedings ACM-SIGIR'92, Copenhagen, DK, 1-10.

Harman, D. (1995a). Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, 31(3), 271-289.

Harman, D. (1995b). *Overview of the third text retrieval conference (TREC-3)*. Proceedings TREC'3, NIST Gaithersburg, MD, 1-19.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*. 47(1), 37-49.

Ide, E. (1971). New experiments in relevance feedback. In G. Salton (Ed.), *The SMART retrieval system - experiments in automatic document processing* (pp. 337-354). Englewood Cliffs, NJ: Prentice-Hall..

Ide, E. & Salton, G. (1971). Interactive search strategies and dynamic file organization in information retrieval. In G. Salton (Ed.), *The SMART retrieval system - experiments in automatic document processing* (pp. 373-393). Englewood Cliffs, NJ: Prentice-Hall.

Iivonen, M. (1995). *Searchers and searchers: Differences between the most and least consistent searchers.* Proceedings ACM-SIGIR'95, Seattle, WA, 149-157.

Koza, J. (1992). *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA: The MIT Press.

Krovetz, R. & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM-Transactions on Information Systems*. 10(2), 115-141.

Kwok, K. L. (1990). Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*. 8(4), 363-386.

Kwok, K. L. (1995). A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*. 13(3), 324-353.

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research*. 6, 155-177.

Michie, D., Spiegelhalter, D. J. & Taylor, C. C. (Ed.) (1994). *Machine learning, neural and statistical classification*. New-York, NY: Ellis Horwood.

van Rijsbergen, C. J. (1979). *Information retrieval*. London, UK: Butterworths.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*. 46(4), 359-364.

Rocchio, J. J. Jr.: Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system - experiments in automatic document processing* (pp. 313-323). Englewood Cliffs, NJ: Prentice-Hall..

Salton, G. (Ed.) (1971). *The SMART retrieval system - Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management*. 24(5), 513-523.

Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*. 41(4), 288-297.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III Searchers, searches, and overlap. *Journal of the American Society for Information Science*. 39(3), 197-216.

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

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

Schütze, H., Hull, D. A. & Pedersen, J. O., (1995). *A comparison of classifiers and document representations for the routing problem.* Proceedings ACM-SIGIR'95, Seattle, WA, 229-237.

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*. 36(2), 111-147.

Stone, M. (1977). Cross-validatory: a review. *Mathematische Operationsforschung und Statistik, Series Statistics.* 9, 127-139.

Tague-Sutcliffe, J. & Blustein, J. (1994, November). *A statistical analysis of the TREC-3 data*. Proceedings of the 3rd Text REtrieval Conference TREC'3, Gaithersburg, MD, 385-398.

Yang, J.-J., Korfhage, R. R. & Rasmussen, E. (1993). *Query improvement in information retrieval using genetic algorithms*. Proceedings TREC'1, NIST Gaithersburg, MD, 31-58.

Zunde, P. & Dexter, M. E. (1968). Indexing Consistency and Quality. *American Documentation*. 20(3), 259-267.