PRECISION WEIGHTING — AN EFFECTIVE AUTOMATIC INDEXING METHOD

C.T. Yu* and G. Salton+

TR 75-232

April 1975

*Department of Computing Science, University of Alberta, Edmonton, Alberta.

+Department of Computer Science, Cornell University, Ithaca, N.Y. 14853.

This study was supported in part by the Canadian Research Council and in part by the National Science Foundation under grant GJ 43505.

Precision Weighting — An Effective Automatic Indexing Method

C.T. Yu* and G. Salton+

Abstract

A great many automatic indexing methods have been implemented and evaluated over the last few years, and automatic procedures comparable in effectiveness to conventional manual ones are now easy to generate. Two drawbacks of the available automatic indexing methods are the absence of reliable linguistic inputs during the indexing process, and the lack of formal, analytical proofs concerning the effectiveness of the proposed methods.

The precision weighting procedure described in the present study uses relevance criteria to weight the terms occurring in user queries as a function of the balance between relevant and nonrelevant documents in which these terms occur; this approximates a semantic know-how of term importance. Formal mathematical proofs are given under well defined conditions of the effectiveness of the method.

1. Introduction

The basic components of an automatic reference retrieval system are now well understood. Documents and user queries are generally represented by vectors of terms (descriptors, keywords, concepts, etc). Typically a given

*Department of Computing Science, University of Alberta, Edmonton, Alberta.

+Department of Computer Science, Cornell University, Ithaca, N.Y. 14853

This study was supported in part by the Canadian Research Council and in part by the National Science Foundation under grant GJ 43505.

document D_i [query Q_k] might be represented as a vector $(d_{i1}, d_{i2}, \ldots, d_{in})$ $[(q_{k1}, q_{k2}, \ldots, q_{kn})]$, where d_{ij} $[q_{kj}]$ represents the weight of term j in D_i $[Q_k]$.

For each document D_i and query Q_k, a similarity function $S(D_i, Q_k)$ can be computed to represent the closeness between the query and the corresponding document. For retrieval purposes, the similarity coefficients between the documents and a given user query can then be utilized to arrange the documents in decreasing order of the similarity functions, followed by the retrieval of those documents whose query-document similarity exceeds a given threshold T.

Given a ranked list of documents in decreasing query-document similarity order, and a knowledge of the relevance, or nonrelevance, of each document with respect to the query, it is possible to compute recall and precision values for various levels of the retrieval threshold T. Recall and precision are defined as the proportion of relevant items retrieved, and the proportion of retrieved items that are relevant, respectively, and a good retrieval system is one which for most user queries produces acceptable values of precision at given levels of recall. By processing the same user query set in several different modes against a given collection, and computing recall and precision values averaged over the set of queries in each case, it is possible to ascertain the relative effectiveness of each processing mode.

Consider now the problem of generating the document and query vectors, that is of choosing appropriate terms and term weights, capable of representing document and query content. A large number of automatic indexing strategies are available for this purpose, among which the following appear most attractive: [1-5]

a) good index terms occur in a given document collection neither too frequently, nor too rarely;

b) terms which occur in too many documents and are therefore likely to produce inadequate retrieval precision should be combined with other appropriate terms to form term phrases;

c) terms which occur in too few documents and are therefore likely to produce inadequate recall should be incorporated into thesaurus classes, and the thesaurus classes should be assigned for content identification instead of the individual terms;

d) the best terms are those which occur with high frequency in certain documents (high term frequency), but whose overall frequency across the documents of a collection is fairly low (low document frequency); this suggests a term weighting function which varies directly with term frequency and inversely with document frequency.

It is not difficult to show performance improvements when one or more of the foregoing indexing devices are incorporated into an actual retrieval process. However, the evidence concerning the effectiveness of a given system is normally based wholly on experimental evidence, reflected in recall and precision measurements; such measurements may show a high average performance, even when the results are poor for certain queries, or for certain recall levels.

This suggests two principal drawbacks of the current work in automatic indexing:

a) the semantic, or linguistic aspect of the role of individual terms or concepts in query or document texts is given up in favor of formal characteristics, such as their frequency distributions, or their location in the body of a text;

b) the approach toward measuring retrieval effectiveness is experimental in nature, and no attempt is made to produce mathematical proofs of the superiority or inferiority of one method over another.

In the precision weighting method to be described in the remainder of this study both of these objections are remedied to some extent. The linguistic aspect is introduced by distinguishing among the text words on the basis not only of frequency characteristics, but also of the document type in which the terms occur, that is, whether a term occurs primarily in documents identified as relevant to a given user query, or whether on the contrary the term prevails among the nonrelevant. A precision weight attached to each query term is then used as a partial indication of the linguistic characterization of the terms.

Given such a precision weighting system, and an assumption concerning the distribution of the vocabulary across the documents of a collection, formal proofs are then provided that at every level of recall, the precision weighting system may be expected to be superior to a system in which the terms in the query and document vectors are unweighted.

2. The Precision Weight Method

Before embarking on the mathematical development, it may be useful briefly to outline the proof procedures and the assumptions leading to the results.

Query and document vectors are assumed to be binary, that is, d_{ij} $[q_{ij}]$ equals 1 whenever term j is present in document [query] i, and is 0 otherwise.

The similarity function s between queries and documents is assumed to be

$$s(D_i, Q_k) = \sum_{j=1}^{n} d_{ij}\, q_{kj}.$$

For binary vectors, s represents the number of matching terms between the query and document vectors, respectively.

The evaluation of the effectiveness of a particular method of term assignment is based on a comparison of the retrieval precision at given levels of the recall. Consider a specified recall level γ, and let $|R|$ be the total number of relevant items for a given query. Then the precision P_γ at recall level γ may be defined as

$$P_\gamma = \frac{\gamma|R|}{\begin{array}{l}\text{Total number of documents to be}\\ \text{retrieved in order to obtain } \gamma|R|\\ \text{relevant ones}\end{array}}$$

The computation of P_γ makes it necessary to identify the number of irrelevant documents that must be retrieved for each increase of 1 in the number of relevant documents obtained. This in turn requires the following assumption to be made regarding the occurrences and composition of the relevant and irrelevant documents in the collection:

Assumption 1: For each query, the corresponding query terms are assumed to be randomly distributed across the set of relevant documents R, and across the set of nonrelevant documents I. That is, the probability of occurrence of a given query term j_k has the same value for all relevant documents in R; similarly, the value is the same for all nonrelevant documents in I (although the two probabilities may differ among themselves).

More formally, consider query Q_j with terms {1, 2, ..., m}. Let r_{j_k} (h_{j_k}) be the number of relevant documents (nonrelevant documents) containing the kth term of Q_j, respectively. It is then assumed that the probability of a relevant (nonrelevant) document containing term j_k is equal to $r_{j_k}/|R|$ ($h_{j_k}/|I|$), where $|R|$ ($|I|$) are the number of relevant (number of nonrelevant) documents in the collection.

Under this assumption, it is easy to show (see Appendix 1) that the expected number of relevant documents containing exactly the set of query terms $\{j_1, j_2, \ldots, j_i\}$ is

$$\frac{\left(\prod_{k=1}^{i} r_{j_k}\right)\left(\prod_{\ell \varepsilon T_1} (|R| - r_\ell)\right)}{|R|^{m-1}}, \tag{1}$$

where T_1 is the complete set of terms in query Q_j {1, 2, ..., m} less the terms occurring in the initial product, that is $T_1 = \{1, 2, \ldots, m\} - \{j_1, j_2, \ldots,$

Similarly, the expected number of nonrelevant documents containing exactly the terms is

$$\frac{\left(\prod_{k=1}^{i} h_{j_k}\right)\left(\prod_{\ell \varepsilon T_1} (|I| - h_\ell)\right)}{|I|^{m-1}}. \tag{2}$$

It is shown in the next section how expressions (1) and (2) can be used to compute the precision of the retrieval for certain levels of recall, that is, following the retrieval of a fixed number of documents relevant to a given query.

Consider now the precision weight system. For each term j_k in each query Q_j, the term precision $pr(j_k)$ is defined as

$$pr(j_k) = \left(\frac{r_{j_k}}{|R| - r_{j_k}}\right) \Bigg/ \left(\frac{h_{j_k}}{|I| - h_{j_k}}\right) . \qquad (3)$$

Obviously the function of expression (3) assigns high values to query terms prevalent in the relevant items and rare in the nonrelevant, and vice-versa for those prevalent mainly in the nonrelevant items.

Given the term precision $pr(j_k)$, a term weight p_{jk} can now be assigned to each query term j_k such that

$$p_{jk} > p_{j\ell} \Longleftrightarrow pr(j_k) \geq pr(j_\ell). \qquad (4)$$

Using term weights of the type introduced in (4), it is possible to construct from each original query Q_j a new query $Q_j{}^*$ by using the weighted terms instead of the original ones, that is

$$Q_j{}^* = (q_{j1} \cdot p_1, q_{j2} \cdot p_2, \ldots, q_{jm} \cdot p_m).$$

It can be shown (see Appendix 2) that an assignment of term weights exists conforming to inequality (4) with the following properties: given two documents D_i and D_k exhibiting respectively u and v matching terms with Q_j, then

$$s(D_i, Q_j{}^*) > s(D_k, Q_j{}^*) \qquad (5)$$

whenever i) $u > v$;

or ii) $u = v$, and D_i contains a query term not also in D_k that exceeds in weight any query term in D_k that is not also in D_i.

The second condition implies that when two documents exhibit the same number of matching query terms, it is sufficient to consider those unique query terms that occur in one of the two documents, but not in both. The higher query-document similarity will then be assigned to that document which contains the highest weighted query term among the unique ones.

More precisely, consider the case for i matching query terms out of m, $1 \leq i \leq m$. There are C_i^m different subsets of m terms each containing exactly i elements. If the increasing numeric order of the individual query terms corresponds to decreasing weight order — the most highly weighted term being designated by rank 1, the second most highly weighted by 2, and so on, down to the mth weighted term — the C_i^m possible sets of i matching terms out of m may be designated by vectors ranging from $(1, 2, \ldots, i)$ to $(m-i+1, m-i+2, \ldots, m)$. Such a vector, considered as an i-tuple, is known as an entry and can be used to determine the order of retrieval. That is, documents whose matching term set is specified by entry $(1, 2, \ldots i)$ are retrieved ahead of those with entry $(1, 2, \ldots, i-1, i+1)$, and so on, down to those with entry $(m-i+1, m-i+2, \ldots, m)$.

For convenience, single entries may be designated as zero-level blocks; the set of zero-level blocks which differ only in the right-most digit are first-level blocks; those differing in the two right-most digits are second-level blocks; and so on, down to the i-th level block which includes all C_i^m entries. The ordering among the entries — top-to-bottom, left column first — and the corresponding block structure are illustrated in Table 1 for $m=7$ and $j=4$. In this case the entries range from (1, 2, 3, 4) to (4, 5, 6, 7). The blocks are ordered according to their entries, that is, if E and F are two distinct jth level blocks containing entries e and f respectively, with e ordered before f, then block E is ordered before block F.

It remains to show that the precision weight method is superior to the standard query indexing system in which the query terms are not weighted. The process used for this purpose consists in computing the search precision for both the weighted and the unweighted retrieval systems at each recall level γ and comparing the results. The search precision in turn depends on Assumption 1 regarding the occurrences of query terms in relevant and nonrelevant documents, respectively, and on the resulting expected number of retrieved relevant, and retrieved nonrelevant documents for a given number of matching query-document terms (Appendix 1).

The recall points at which the precision is calculated are determined as follows. For the precision weight method, the order of retrieval of the documents — and therefore the ranks of the relevant documents — are strictly determined by the number of matching query-document terms; for documents with a common number of matching query terms the suborder is by entry number order, as previously explained.

For the standard unweighted method, the order of retrieval is also in decreasing order of the number of query-document term matches. However, no strict ranking exists within each set of documents exhibiting a common number of matching query terms. To determine a ranking within each of these document subsets, the following assumption must be made:

Assumption 2: Let $c(c > 1)$ relevant items and g nonrelevant items all exhibit the same coefficient with respect to some query Q, then it is assumed that g/c nonrelevant items are retrieved for each relevant retrieved. That is, the relevant items occur at even intervals among the set of nonrelevant in the ranked list of retrieved documents.

The only difference between the precision-weight method and the standard unweighted system is that the former allows a stricter ranking of the output documents for those items exhibiting a common number of query-document term matches. When the query terms are weighted in decreasing order of term precision, the relevant documents are, however, more likely to be retrieved early in the output order than when unweighted terms are used; hence the improvement in retrieval effectiveness.

The proof procedure is included in the next section.

3. The Effectiveness of the Precision Weight System

Consider a given query Q with a total of $|R|$ relevant documents. The query-document matching function induces an ordering among the retrieved documents as previously explained. Following the retrieval of each relevant document, the value of the recall goes up by $1/|R|$, reaching $|R|/|R|$ (that is, 1 following the retrieval of the last relevant item. Thus, in principle, a total of $|R|$ different recall points are possible for each query. Among all the possible recall points, some are of special interest, corresponding to the highest recall obtainable for a given number of matching query-document terms. In particular, for each set of documents exhibiting a common number of query-document term matches, a standard recall point is defined as the point corresponding to the retrieval of the last relevant document within that set of documents. The complete set of standard recall points for a given query may be designated by $\{s_1, s_2, \ldots, s_z\}$. The first three standard recall points (and the respective recall-precision values), corresponding to 7, 5, and 4 matching query terms, are shown for a typical sample query in Table 2. Ten relevant documents are assumed for the sample query of Table 2.

Let d_v be the minimum number of term matches between query Q and any document retrieved at recall point s_v for $1 \leq v \leq z$. It will now be shown that the retrieval precision obtained with the modified, weighted query Q* is not inferior to that obtained with the original query Q at any standard recall point s_v, or at any retrieval level between consecutive standard recall points.

Consider first the precision computation for the standard unweighted terms and query Q. At any standard recall point, say s_v, the retrieved documents can be classified into two types

i) documents having more than d_v terms in common with Q.

ii) documents having exactly d_v terms in common with Q.

Documents of type ii) can be further partitioned into smaller sets as follows. If query Q contains m terms, then there exist $y = C^m_{d_v}$ ways in which a document can have d_v terms in common with query Q. Each of the y distinct sets of terms may be represented by an entry a_ℓ in a d_vth level block of the type shown in the example of Table 1. The set of entries is $\bigcup_{\ell=1}^{y} a_\ell$, and the number of relevant (irrelevant) documents having exactly the set a_ℓ in common with Q may be denoted by a_ℓ' (a_ℓ''), respectively.

(a_ℓ'.and a_ℓ" may of course be computed using the Assumption 1 and the development of Appendix 1.)

The following quantities are now readily available:

a) The number of relevant (irrelevant) documents having exactly d_v terms in common with Q can be taken as $\sum_{\ell=1}^{y} a_\ell'$ $(\sum_{\ell=1}^{y} a_\ell'')$.

b) Assuming the number of relevant (irrelevant) documents having more than d_v terms in common with Q to be B' (B"), the total number of relevant documents retrieved at standard recall point s_v and at the last previc standard recall point s_{v-1} is $B' + \sum_{\ell=1}^{y} a_\ell'$ and B' respectively.

Consider now the precision of retrieval for Q at some recall point x, between s_{v-1} and s_v. The number of relevant documents retrieved at x should be less than $(B' + \sum_{\ell=1}^{y} a_\ell')$ but greater than B'. Without loss of generality, assume that the number of relevant documents retrieved by Q at $x = B' + \sum_{\ell=1}^{k} a_\ell'$ for some k, $1 \leq k < y$. In order to find the precision of retrieval, the number of documents retrieved must be known. Since every document of type ii) has the same likelihood of being retrieved by Q, the number of documents of type ii) that must be retrieved in order to retrieve these $\sum_{\ell=1}^{k} a_\ell'$ relevant documents can be assumed to be (by Assumption 2)

$$\left[\frac{(\sum_{\ell=1}^{y} a_\ell' + \sum_{\ell=1}^{y} a_\ell'')}{(\sum_{\ell=1}^{y} a_\ell')}\right] \cdot \left[\sum_{\ell=1}^{k} a_\ell'\right]. \qquad (6)$$

Thus the precision of retrieval at an arbitrary (nonstandard) recall point x — that is, the number of relevant retrieved at x divided by the total retrieved — will be equal to

$$\frac{B' + \sum_{\ell=1}^{k} a_\ell'}{\left\{B' + B'' + \left[\frac{\sum_{\ell=1}^{y} a_\ell' + \sum_{\ell=1}^{y} a_\ell''}{(\sum_{\ell=1}^{y} a_\ell')}\right] \cdot \left[\sum_{\ell=1}^{k} a_\ell'\right]\right\}} \qquad (7)$$

for the standard unweighted retrieval system.

Consider the precision for the weighted system using queries Q^* instead of Q. Unlike Q which treats every document of type ii) equally (in the sense that each has the same chance of being retrieved by Q), documents of type ii) are ordered linearly by Q^*, in increasing entry number order. In particular, documents exhibiting term set a_1 in common with Q are retrieved first, followed by those with a_2 in common with Q, and so on, until those with a_y in common are obtained.

Thus, to retrieve $\sum_{\ell=1}^{k} a_\ell'$ relevant documents out of the $\sum_{\ell=1}^{y} a_\ell'$ relevant ones of type ii), a total of

$$\sum_{\ell=1}^{k} a_\ell' + \sum_{\ell=1}^{k} a_\ell''$$

documents in all must be retrieved by Q^*. This implies that the precision at recall point x for Q^* is

$$\frac{B' + \sum_{\ell=1}^{k} a_\ell'}{(B' + B'' + \sum_{\ell=1}^{k} a_\ell' + \sum_{\ell=1}^{k} a_\ell'')}\,. \qquad (8)$$

To show that expression (8) is not smaller than expression (7), it is necessary and sufficient to demonstrate by comparing the respective denominators that

$$\frac{\sum_{\ell=1}^{k} a_\ell'}{\sum_{\ell=1}^{k} a_\ell''} \geq \frac{\sum_{\ell=1}^{y} a_\ell'}{\sum_{\ell=1}^{y} a_\ell''} . \qquad . (9)$$

It is sufficient to prove (10) as follows

$$\frac{\sum_{\ell=1}^{k} a_\ell'}{\sum_{\ell=1}^{k} a_\ell''} \geq \frac{\sum_{\ell=k+1}^{y} a_\ell'}{\sum_{\ell=k+1}^{y} a_\ell''} \qquad (10)$$

because when $x_1/y_1 \geq x_2/y_2$ and $x_1, x_2, y_1, y_2 \geq 0$, it is easy to show that

$$x_1/y_1 \geq (x_1 + x_2)/(y_1 + y_2) \geq x_2/y_2. \qquad (11)$$

The proof proceeds in two main steps. First the result is established for the case where the boundary indicator k coincides with the end of a block (lemma 1). This is done by showing that the ratio of relevant documents to irrelevant documents retrieved represented by the entries of a given block is at least as high as that represented by the entries of the next block. Thus, when k coincides with the end of a block, a repeated application of inequality (11) to the result of lemma 1 will prove (10). The result of lemma 1 is then used to prove the inequality for arbitrary k (lemma 2). A block of consecutive entries of the type shown in Table 1 may be designated by X; the corresponding expected number of relevant and irrelevant documents, that is, the expected number of documents exhibiting exactly the matching query terms specified by any entry in the block X may as before be identified by X' and X", respectively.

Lemma 1: Let E and F be two consecutive (j-1)th level blocks in the same jth level block, with E ordered before F. If the query terms are randomly distributed across the relevant and irrelevant documents of the collection (Assumption 1), then $E'/E'' \geq F'/F''$.

Proof: The proof is by induction on j. When j=1, E and F are consecutive single entries. Let those entries be $(j_1, j_2, \ldots, j_{i-1}, j_i)$ and $(j_1, j_2, \ldots, j_{i-1}, j_{i+1})$ respectively, and let v and w designate $|R|^{m-1}$ and $|I|^{m-1}$, respectively.

In view of Assumption 1 regarding the query term distribution (see Appendix 1) one obtains

$$E'/E'' = [v\,(\prod_{k=1}^{i} r_{j_k}) \prod_{\ell\epsilon T_1} (|R|-r_\ell)] \;/\; [w\,(\prod_{k=1}^{i} h_{j_k}) \prod_{\ell\epsilon T_1} (|I|-h_\ell)]$$

$$= \frac{[v\,(\prod_{k=1}^{i-1} r_{j_k})(\prod_{\ell\epsilon T_1-\{j_{i+1}\}} (|R|-r_\ell))\, r_{j_i}(|R|-r_{j_{i+1}})]}{[w\,(\prod_{k=1}^{i-1} h_{j_k})(\prod_{\ell\epsilon T_1-\{j_{i+1}\}} (|I|-h_\ell))\, h_{j_i}(|I|-h_{j_{i+1}})]}$$

$$\geq \frac{[v\,(\prod_{k=1}^{i-1} r_{j_k})(\prod_{\ell\epsilon T_1-\{j_{i+1}\}} (|R|-r_\ell))\, r_{j_{i+1}}(|R|-r_{j_i})]}{[w\,(\prod_{k=1}^{i-1} h_{j_k})(\prod_{\ell\epsilon T_1-\{j_{i+1}\}} (|I|-h_\ell))\, h_{j_{i+1}}(|I|-h_{j_i})]}$$

$$= \frac{[v\,(\prod_{k=1}^{i-1} r_{j_k})\, r_{j_{i+1}} \prod_{\ell\epsilon T_1-\{j_{i+1}\}\cup\{j_i\}} (|R|-r_\ell)]}{[w\,(\prod_{k=1}^{i-1} h_{j_k})\, h_{j_{i+1}} \prod_{\ell\epsilon T_1-\{j_{i+1}\}\cup\{j_i\}} (|I|-h_\ell)]}$$

$= F'/F''$. This proves the inequality for j=1, and $T_1 = \{1, 2, \ldots, m\} - \{j_1, j_2, \ldots, j_i\}$.

Consider now a (j-1)th block E which includes g (j-2)th level blocks, that is, $E = \bigcup_{k=1}^{g} E_k$, where E_k is the kth (j-2)th level block of E. Let x and y be the first entries of E_1 (and therefore of E) and F respectively. Let $x = (x_1, x_2, \ldots, x_i)$ and $y = (y_1, y_2, \ldots, y_i)$. Since both of x and y are in the same jth level block, we have $x_k = y_k$ for $1 \leq k \leq i-j$. The fact that x and y are in consecutive (j-1)th level block forces $x_{i-j+1} = y_{i-j+1} - 1$. Furthermore, $x_{i-j+2} = x_{i-j+1} + 1$ and $y_{i-j+2} = y_{i-j+1} + 1$, since x and y are the first entries of E and F respectively. Thus, $y_{i-j+2} = x_{i-j+2} + 1$. Let $z = (z_1, z_2, \ldots, z_i)$ be the first entry in E_2. Since x and z are in the same (j-1)th level block, $z_k = x_k = y_k$ for $1 \leq k \leq i-j$, and $z_{i-j+1} = x_{i-j+1} = y_{i-j+1} - 1$. The fact that x and z are in consecutive (
level blocks makes $z_{i-j+2} = x_{i-j+2} + 1 = y_{i-j+2}$. Thus, it is easy to see that
mapping $Z((\ell_1, \ell_2, \ldots, \ell_{i-j}, \ell_{i-j+1}, \ell_{i-j+2}, \ldots, \ell_i)) =$
$(\ell_1, \ell_2, \ldots, \ell_{i-j}, \ell_{i-j+1} + 1, \ell_{i-j+2}, \ldots, \ell_i)$ from $\bigcup_{k=2}^{g} E_k$ to F is 1-1 an

By induction, $E_1'/E_2'' \geq E_2'/E_2'' \geq \ldots \geq E_g'/E_g''$. Thus,

$$E_1'/E_1'' \geq \max \{E_2'/E_2'', E_3'/E_3'', \ldots E_g'/E_g''\}$$

$$\geq (E_2' + E_3' + \ldots E_g') / (E_2'' + \ldots E_g''), \text{ using inequality (11}$$

This implies

$$E'/E'' \geq \min \{E_1'/E_1'', (E_2' + \ldots E_g') / (E_2'' + \ldots E_g'')\}$$

$$= \frac{[v \, (\pi_{k=1}^{i-j} \, r_{\ell_k})(r_{\ell_{i-j+1}})(|R| - r_{\ell_{i-j+1}} + 1)(\Sigma_{\ell} \Big\{ (\pi_{k=i-j+2}^{i} \, r_{\ell_k}) \; \pi_{s \varepsilon T_{1_\ell}} \; (|R| - r_s}{[w \, (\pi_{k=1}^{i-j} \, h_{\ell_k})(h_{\ell_{i-j+1}})(|I| - h_{\ell_{i-j+1}} + 1)(\Sigma_{\ell} \Big\{ (\pi_{k=i-j+2}^{i} \, h_{\ell_k}) \; \pi_{s \varepsilon T_{1_\ell}} \; (|I| - h_s}$$

where ℓ is an entry in $\bigcup_{k=2}^{g} E_k$, $T_{1_\ell} = \{1, 2, \ldots, m\} - \{\ell_1, \ell_2, \ldots, \ell_i\}$ and $\sum_{\ell}$ is summing over all entries in $\bigcup_{k=2}^{g} E_k$. Letting $f = (f_1, f_2, \ldots, f_i)$ be the corresponding entry of ℓ we obtain $f_{i-j+1} = \ell_{i-j+1} + 1$ and $f_k = \ell_k$, $1 \leq k \leq i$ and $k \neq i-j+1$. Since the terms are arranged in decreasing precision values, the previous expression is greater than or equal to

$$\geq \frac{[v\,(\prod_{k=1}^{i-j} r_{\ell_k})(r_{f_{i-j+1}})(|R|-r_{f_{i-j+1}})(\sum_{\ell}\left\{(\prod_{k=i-j+2}^{i} r_{\ell_k})\prod_{s\epsilon T_{1_\ell}}(|R|-r_s)\right\}))]}{[w\,(\prod_{k=1}^{i-j} h_{\ell_k})(h_{f_{i-j+1}})(|I|-h_{f_{i-j+1}})(\sum_{\ell}\left\{(\prod_{k=i-j+2}^{i} h_{\ell_k})\prod_{s\epsilon T_{1_\ell}}(|I|-h_s)\right\}))]}$$

$$= \frac{[v\,(\prod_{k=1}^{i-j} r_{f_k})(r_{f_{i-j+1}})(|R|-r_{f_{i-j+1}})(\sum_{f}\left\{(\prod_{k=i-j+2}^{i} r_{f_k})\prod_{s\epsilon T_{1_f}}(|R|-r_s)\right\}))]}{[w\,(\prod_{k=1}^{i-j} h_{f_k})(h_{f_{i-j+1}})(|I|-h_{f_{i-j+1}})(\sum_{f}\left\{(\prod_{k=i-j+2}^{i} h_{f_k})\prod_{s\epsilon T_{1_f}}(|I|-h_s)\right\}))]}$$

$= F'/F''$, where $T_{1_f} = \{1, 2, \ldots, m\} - \{f_1, f_2, \ldots, f_i\}$ and $\sum_{f}$ is summing over all entries in F. ▌

The proof for the general case (for arbitrary values of k) is given in the next lemma.

Lemma 2: Let $y = C_i^m$, that is, y represents the number of entries in a block structure for i matching terms out of m; and let a_ℓ designate the ℓth entry in an ith level block E. Then under the same Assumption 1 as before regarding the distribution of query terms in the documents $(\sum_{\ell=1}^{k} a_\ell')/(\sum_{\ell=1}^{k} a_\ell'') \geq (\sum_{\ell=k+1}^{y} a_\ell')/(\sum_{\ell=k+1}^{y} a_\ell'')$ for $1 \leq k < y$.

Proof: For any k, $1 \le k < y$, a_k is the last entry of an $(i-j)$th level block in E for some j, $1 \le j \le i$. (A zeroth level block is an entry with itself as the last entry in the block). In order to avoid excessive use of symbols, the lemma is proved for cases $j=1$ and $j=2$ only. It is easy to see that the proof can be extended for any j, $1 \le j \le i$. Let E have g $(i-1)$th level blocks, i.e. $E = \bigcup_{x=1}^{g} E_x$.

For $j=1$, a_k is the last entry of E_x for some x, $1 \le x \le g$. By lemma 1, we have $E_1'/E_1'' \ge E_2'/E_2'' \ge \ldots \ge E_g'/E_g''$. Thus

$$\left(\sum_{\ell=1}^{k} a_\ell'\right) / \left(\sum_{\ell=1}^{k} a_\ell''\right) = \left(\sum_{\ell=1}^{x} E_\ell'\right) / \left(\sum_{\ell=1}^{x} E_\ell''\right) \ge \min_{1\le\ell\le x} \{E_\ell'/E_\ell''\}$$

$$= E_x'/E_x'' \ge E_{x+1}'/E_{x+1}'' \ge \max_{x+1\le\ell\le g} \{E_\ell'/E_\ell''\}$$

$$\ge \left(\sum_{\ell=x+1}^{g} E_\ell'\right) / \left(\sum_{\ell=x+1}^{g} E_\ell''\right) = \left(\sum_{\ell=k+1}^{y} a_\ell'\right) / \left(\sum_{\ell=k+1}^{y} a_\ell''\right).$$

For $j=2$, let a_k be the last entry of the t th $(i-2)$th level block within the xth $(i-1)$th level block. Let $E_x = \bigcup_{\ell=1}^{w} F_\ell$, i.e. E_x has w $(i-2)$th level blocks. Applying the result of lemma 1 and proceeding as above, one obtains

$$\left(\sum_{\ell=1}^{t} F_\ell'\right)/\left(\sum_{\ell=1}^{t} F_\ell''\right) \ge \left(\sum_{\ell=t+1}^{w} F_\ell'\right)/\left(\sum_{\ell=t+1}^{w} F_\ell''\right).$$

Using (11), this can be rewritten as

$$\left(\sum_{\ell=1}^{t} F_\ell'\right)/\left(\sum_{\ell=1}^{t} F_\ell''\right) \ge \left(\sum_{\ell=1}^{t} F_\ell' + \sum_{\ell=t+1}^{w} F_\ell'\right)/\left(\sum_{\ell=1}^{t} F_\ell'' + \sum_{\ell=t+1}^{w} F_\ell''\right)$$

$$= E_x'/E_x'' \ge \left(\sum_{\ell=t+1}^{w} F_\ell'\right)/\left(\sum_{\ell=t+1}^{w} F_\ell''\right).$$

Thus,

$$\Big(\sum_{\ell=1}^{k} a_\ell'\Big) / \Big(\sum_{\ell=1}^{k} a_\ell''\Big) = \Big(\sum_{\ell=1}^{x-1} E_\ell' + \sum_{\ell=1}^{t} F_\ell'\Big) / \Big(\sum_{\ell=1}^{x-1} E_\ell'' + \sum_{\ell=1}^{t} F_\ell''\Big)$$

$$\geq \min \Big\{ \sum_{\ell=1}^{x-1} (E_\ell') / \sum_{\ell=1}^{x-1} (E_\ell''),\ \sum_{\ell=1}^{t} (F_\ell') / \sum_{\ell=1}^{t} (F_\ell'')\Big\}$$

$$\geq E_x'/E_x'' \geq \max \{E_x'/E_x'',\ E_{x+1}'/E_{x+1}''\}$$

$$\geq \Big(\sum_{\ell=t+1}^{w} F_\ell' + \sum_{\ell=x+1}^{g} E_\ell'\Big) / \Big(\sum_{\ell=t+1}^{w} F_\ell'' + \sum_{\ell=x+1}^{g} E_\ell''\Big)$$

$$= \Big(\sum_{\ell=k+1}^{y} a_\ell'\Big) / \Big(\sum_{\ell=k+1}^{y} a_\ell''\Big). \blacksquare$$

The result of lemma 2 together with the previous discussion shows that the precision weight process is superior to the standard unweighted process at any arbitrary (nonstandard) recall point. It remains to be shown that the result is also true for any standard recall point. The next theorem establishes this fact and summarizes the results.

Theorem 1: Let the terms of a given query $Q = \{1, 2, \ldots, m\}$ be arranged in decreasing order of their precision values. There exists an assignment of weights to the terms of Q, consistent with the precision weight method, such that if assumption 1 holds, the indexing method is superior to the standard unweighted method at any recall point.

Proof: It is sufficient to show the case for the standard recall points. Consider the set D of documents having d_v or more terms in common with the original query (i.e. the standard recall point s_v). If D_j is any relevant document retrieved by Q and D_i any irrelevant document not retrieved by Q at recall s_v, then by definition of retrieval at s_v, the number of term matches of D_j with Q is greater than or equal to d_v and that of D_i with Q is les than d_v. By Appendix 2, $f(Q^*, D_j) > f(Q^*, D_i)$. Thus, all the relevant documents D can be retrieved by Q^* without retrieving any irrelevant document previously not retrieved by Q at recall s_v. This implies that the precision of retrieval for Q^* at s_v is at least as high as that for Q. ∎

4. Implementation of the Method

The precision of a term with respect to a query is normally not known before the retrieval of documents has taken place. Furthermore, the term precision is difficult to determine accurately. However, exact values of the term precision are unimportant, since the corresponding values are used only as a ranking device for output documents. Thus only relative magnitudes of the term precision need be obtained, and these can be approximated as follows. It may be assumed that the collection of documents can be partitioned into a number of sub-collecti containing "similar" documents. Furthermore, consider a number of "typical" queries, $\hat{Q}$, containing the terms for the given subcollections. For each term, k, in the typical queries, one can compute the average values of the following ratios over all queries in $\hat{Q}$: $r_k'/(|R \cap S_Q|-r_k')$ and $h_k'/(|I \cap S_Q|-h_k')$* where r_k' (h_k') is the number of relevant (irrelevant) documents containing term k and retrieved by query Q in $\hat{Q}$, and S_Q is the set of documents retrieved by one of the typical queries in $\hat{Q}$. The exact precision value of term k, $(r_k/(|R|-r_k))/(h_k/|I|-h_k))$ may then be approximated by $(r_k'/(|R \cap S_Q|-r_k'))/(h_k'/(|I \cap S_Q|-h_k'))$.

* This type of information is readily available in a retrieval system based on user-system interaction of the relevance feedback type. [6,7]

5. Experimental Results

Assumption 1 regarding the distribution of query terms in the relevant and nonrelevant documents may not always be completely satisfied in practice. Thus theorem 1 may not be valid in the most general situation. Experimental results are given to illustrate the effectiveness of the precision weight method for a practical case where Assumption 1 is not necessarily valid.

The experiment is performed using a subset of ten queries in conjunction with the Cranfield collection of documents in aerodynamics (CRN 2 NUL). Let Δ_1 be a value of Δ so chosen as to satisfy condition (i) of lemma 3 in Appendix 2, while transforming condition (ii) into an equality instead of a strict inequality as in lemma 3. Setting $\Delta_2 = \Delta_{1/2}$, both conditions (i) and (ii) are then satisfied. The tabulated retrieval results for each of the ten queries is shown in Table 3. The precision of retrieval is given for each query at intervals of $1/|R|$, where $|R|$ is the number of relevant documents for the query. The percentage improvement (or deterioration) obtained for the sample queries with the precision method over the original unweighted queries is given at intervals of $1/|R|$, as well as over the whole recall range. The comparison of the retrieval performance for the modified queries with that of the original queries is presented in the following terms:

(i) For each query, an average improvement is determined (averaged over all the recall points).

(ii) Of the 10 queries tested, a retrieval improvement is obtained at each recall point for 8 queries.

(iii) The average improvement for all the queries is 91.6%.

Thus, one concludes that Assumption 1 has a good chance of being satisfied and that the precision weight method is a useful indexing device. Efficient implementations of the method and extensions to other more desirable query-document matching functions remain to be discovered.

References

[1] G. Salton, C.S. Yang and C.T. Yu, Contributions to the Theory of Indexing, Information Processing 74, North Holland Publishing Co., Amsterdam, 1974, p. 584-590.

[2] G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, Computer Science Dept., Technical Report 74-208, Cornell University, Ithaca, July 1974, to appear in Journal of the ASIS.

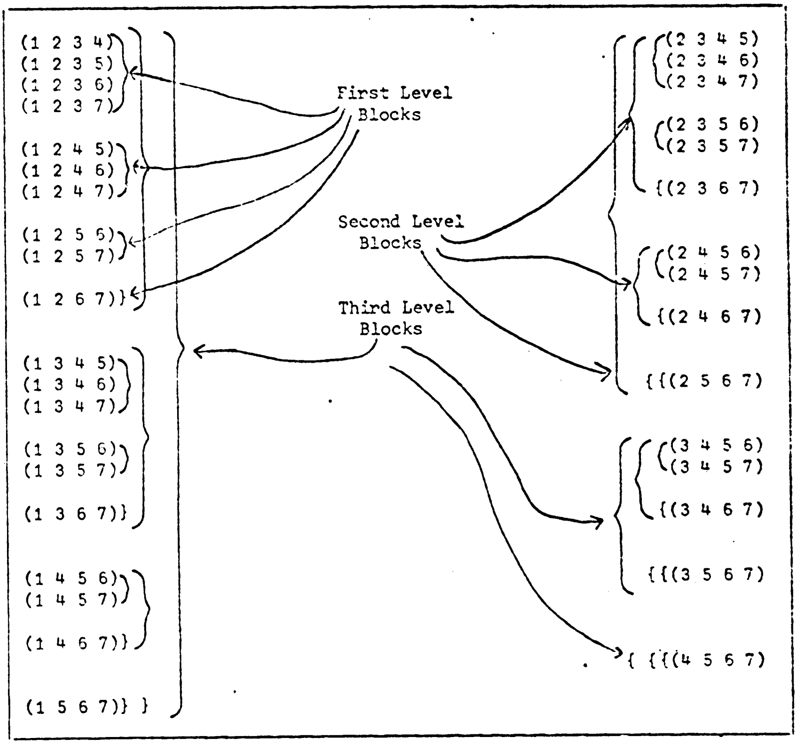
[3] G. Salton, Recent Studies in Automatic Text Analysis and Document Retrieval, Journal of the ACM, Vol. 20, No. 2, April 1973, p. 258-278.

[4] K. Spark Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, p. 11-21.

[5] G. Salton, and C.T. Yu, The Construction of Effective Vocabularies for Information Retrieval, Proceedings of the SIGPLAN - SIGIR Interface Meeting, Association for Computing Machinery, p. 48-60, 1974.

[6] J.J. Rocchio, Jr., Relevance Feedback in Information Retrieval, in The SMART System — Experiments in Automatic Document Processing, G. Salton, editor, Chapter 14, Prentice Hall, Englewood Cliffs, N.J., 1971.

[7] G. Salton, The Performance of Interactive Information Retrieval, Information Processing Letters, Vol. 1, No. 2, July 1971, p. 35-41.



Complete Fourth Level Block for Four

Matching Terms Out of Seven

Table 1

Document Rank	Relevance Indicator (R means relevant)	Number of Matching Query-Document Terms	Standard Recall Point ✓	Recall	Precision
1	R	7		0.1	
2	N	7	✓	0.1	0.5
3	N	6		0.1	
4	N	6		0.1	
5	R	5		0.2	
6	N	5	●	0.2	0.33
7	R	5		0.3	
8	N	5	✓	0.3	0.37
9	R	4		0.4	
10	N	4	●	0.4	0.42
11	R	4	✓	0.5	0.45
12	N	3			
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	

Typical Precision Computation at

Standard Recall Points

(assumption: total number of relevant is 10)

Table 2

✓ standard recall points

● additional points for which precision is computable

Query No.	Recall	Original Query	Modified Query	of each Recall Point	Improvement
33	0.50	0.500	1.000	100.0	83.3
	1.00	0.033	0.055	67.0	
34	0.20	0.100	0.250	150.0	81.7
	0.40	0.067	0.108	61.0	
	0.60	0.061	0.111	82.0	
	0.80	0.040	0.080	100.0	
	1.00	0.032	0.037	16.0	
35	0.20	1.000	1.000	0.0	8.2
	0.40	0.857	0.666	-22.3	
	0.60	0.818	0.750	-8.3	
	0.80	0.800	0.800	0.0	
	1.00	0.416	0.714	71.4	
36	0.25	0.200	0.250	25.0	86.5
	0.50	0.137	0.307	124.0	
	0.75	0.125	0.375	200.0	
	1.00	0.040	0.042	5.0	
37	0.25	1.000	1.000	0.0	71.5
	0.50	0.333	0.666	100.0	
	0.75	0.230	0.428	86.1	
	1.00	0.038	0.076	100.0	
38	0.14	1.000	1.000	0.0	8.0
	0.28	1.000	1.000	0.0	
	0.42	1.000	1.000	0.0	
	0.57	0.705	0.666	-5.5	
	0.71	0.600	0.714	19.0	
	0.85	0.545	0.545	0.0	
	1.00	0.350	0.500	42.9	
39	0.20	0.166	1.000	500.0	183.0
	0.40	0.129	0.285	121.0	
	0.60	0.120	0.300	150.0	
	0.80	0.058	0.137	136.0	
	1.00	0.034	0.036	5.8	
40	0.50	0.200	1.000	400.0	275.0
	1.00	0.133	0.333	150.0	
41	0.20	0.083	0.250	200.0	73.7
	0.40	0.026	0.035	34.6	
	0.60	0.029	0.046	58.6	
	0.80	0.031	0.054	74.2	
	1.00	0.025	0.025	0.0	
42	0.125	0.333	0.500	50.0	43.1
	0.250	0.333	0.500	50.0	
	0.375	0.375	0.400	6.7	
	0.500	0.400	0.440	10.0	
	0.625	0.416	0.500	20.0	
	0.750	0.428	0.545	27.3	
	0.875	0.318	0.466	46.5	
	1.00	0.057	0.133	133.3	
				average % of improvement over the 10 queries =	91.6

Comparison of the Modified Queries with
the Original Queries
Table 3

Appendix 1 Expected Number of Retrieved Documents

Consider query Q_j with terms $\{1, 2, \ldots, m\}$, and let R be the set of relevant documents. It is assumed that the distribution of the terms in Q across the relevant document set R is uniform; that is

i) $r_{j_k}/|R|$ is the probability that a relevant document contains term j_k

ii) $(|R|-r_\ell)/|R|$ is the probability that a relevant document does not contain term ℓ.

Assuming that the assignment of the terms is independent, the probability that a relevant document contains exactly the set of terms $\{j_1, j_2 \ldots, j_i\}$ is then

$$[(\mathop{\pi}_{k=1}^{i} r_{j_k}) /|R|^i] \cdot [(\mathop{\pi}_{\ell\epsilon T_1} (|R|- r_\ell))/|R|^{m-i}]$$

$$= \frac{(\mathop{\pi}_{k=1}^{i} r_{j_k})(\mathop{\pi}_{\ell\epsilon T_1} (|R|-r_\ell))}{|R|^m} \qquad (12)$$

where T_1 extends over the whole set of query terms $\{1, 2, \ldots, m\}$ less the terms included in the initial product. To obtain the expected number of relevant documents containing exactly the terms $\{j_1, j_2, \ldots, j_i\}$, the foregoing expression (12) must be multiplied by R, giving

$$\frac{(\mathop{\pi}_{k=1}^{i} r_{j_k})(\mathop{\pi}_{\ell\epsilon T_1} (|R|-r_\ell))}{|R|^{m-1}} \qquad (13)$$

which is expression (1) of section 2. Identical arguments establish expression (2) as the expected number of nonrelevant documents containing exactly terms $\{j_1, j_2, \ldots, j_i\}$.

Appendix 2 The Weighting Function

Lemma 3 Let $\{1, 2, \ldots, m\}$ be the set of terms included in query Q arranged in decreasing order of the term precision. There exists an assignment of weights $p_1 > p_2 \ldots > p_m = 1$ to the terms of Q such that if $\{j_1, j_2, \ldots, j_k \mid j_1 < j_2 < \ldots < j_k\}$ and $\{i_1, i_2, \ldots, i_\ell \mid i_1 < i_2 < \ldots < i$ are two sets of terms in common between the query and two documents D_j and D_i, respectively, with $k \geq \ell$, then the following statements are true:

i) if $k > \ell$, then $f(Q^*, D_j) > f(Q^*, D_i)$,

ii) if $k = \ell$ and there exists a z such that $j_s = i_s$ for $1 \leq s < z$, and $j_z < i_z$ (or $j_1 < i_1$), then $f(Q^*, D_j) > f(Q^*, D_i)$,

where Q^* is the modified (weighted) query derived from Q.

Proof: For $1 \leq i \leq m$, the weights p_i are defined as

$$p_i = 1 + x_i \Delta \tag{14}$$

The variables x_i, $3 \leq i \leq m - 1$ are given by the recursive formula

$$x_{m-i} = \left(\sum_{j=1}^{\lfloor \frac{i+1}{2} \rfloor} x_{m-i+j} - \sum_{j=1}^{\lfloor \frac{i+1}{2} \rfloor -1} x_{m+1-j} \right) + 1$$

with $x_m = 0$, $x_{m-1} = 1$, and $x_{m-2} = 2$. The constant Δ which depends on m is so chosen that $\left(\sum_{j=1}^{\lfloor \frac{m}{2} \rfloor} x_j - \sum_{j=1}^{\lfloor \frac{m}{2} \rfloor} x_{m+1-j} \right) \Delta < 1$, and $\Delta > 0$.

It is sufficient to consider the case where $k+\ell \leq m$, for if $k+\ell > m$, then the terms in common between the two documents can be deleted. For case (i), where $k>\ell$, it is sufficient to consider $k = \ell+1$.

$$
\begin{aligned}
f(Q^*, D_j) - f(Q^*, D_i) &= \sum_{s=1}^{\ell+1} p_{j_s} - \sum_{s=1}^{\ell} p_{i_s} \\
&\geq p_m + \left(\sum_{s=1}^{\ell} p_{j_s} - \sum_{s=1}^{\ell} p_{i_s} \right) \\
&\geq p_m + \left(\sum_{s=1}^{\ell} p_{m+1-s} - \sum_{s=1}^{\ell} p_s \right) \\
&\geq 1 + \left(\sum_{s=1}^{\lfloor \frac{m}{2} \rfloor} p_{m+1-s} - \sum_{s=1}^{\lfloor \frac{m}{2} \rfloor} p_s \right) \\
&= 1 + \left(\sum_{s=1}^{\lfloor \frac{m}{2} \rfloor} x_{m+1-s} - \sum_{s=1}^{\lfloor \frac{m}{2} \rfloor} x_s \right) \Delta > 0.
\end{aligned}
$$

For case (ii), where $k=\ell$, one obtains

$$
f(Q^*, D_j) - f(Q^*, D_i) = \sum_{s=1}^{\ell} p_{j_s} - \sum_{s=1}^{\ell} p_{i_s} = \sum_{s \geq z}^{\ell} p_{j_s} - \sum_{s \geq z}^{\ell} p_{i_s} .
$$

Letting $j_z = m-i$ for some i, the difference in query document similarities becomes

$$
\begin{aligned}
\left(x_{m-i} + \sum_{s>z}^{\ell} x_{j_s} - \sum_{s \geq z}^{\ell} x_{i_s} \right) \Delta &\geq \left(x_{m-i} + \sum_{s=1}^{\ell-z} x_{m+1-s} - \sum_{s=1}^{\ell-z+1} x_{m-i+s} \right) \Delta \\
&\geq \left(x_{m-i} + \sum_{s=1}^{\lfloor \frac{i+1}{2} \rfloor -1} x_{m+1-s} - \sum_{s=1}^{\lfloor \frac{i+1}{2} \rfloor} x_{m-i+s} \right) \Delta > 0. \blacksquare
\end{aligned}
$$