EVALUATION OF THE 2-POISSON MODEL AS A BASIS FOR USING TERM FREQUENCY DATA IN SEARCHING

Vijay V. Raghavan
Department of Computer Science
University of Regina
Regina, Sask. S4S 0A2 CANADA

Hong-pao Shi
Department of Electronic Engineering
Sian Jiao-Tong University CHINA

C.T. Yu
Department of Electrical Engineering and Computer Science
University of Illinois at Chicago
Box 4348, Chicago, Ill., 60680 USA

ABSTRACT

The early work on the probabilistic models of retrieval assumed that the document representation is binary, indicating only the presence or absence of index terms. The 2-Poisson (TP) model which was proposed as a model of how the occurrence frequency of specialty words in a collection is distributed, has since been used to develop retrieval strategies that incorporate term frequency information. This work investigates the use of the TP model, in this context, further. It is shown that the search effectiveness, when no relevance information is assumed, can be further enhanced by using this model. Furthermore, when the term weights proposed in this work are used in conjunction with weights known as term significance weights, the results are very encouraging.

1. INTRODUCTION

Information retrieval systems, usually, deal with bibliographic databases, which are databases consisting of books, journal articles, technical reports, etc. Information retrieval system applications are, however, not limited to the library environment. A collection of office memos or correspondences, a court file of previous cases and rulings, and a government agency's file of patents, copyrights, etc. are all examples of textual data where the use of these systems can be beneficial.

In information retrieval, each stored document is normally reprsented by a set of keywords or terms, which collectively describe the content of the document. The task of assigning the terms to individual documents is known as indexing. When a user presents a request, which is also represented in the system by keywords, it is compared with the document representations and the documents are retrieved based on their degrees of similarity to the request. If a document is of interest to the user, it is termed relevant in relation to the given request; the document is non-relevant, otherwise.

In terms of the main activities involved, the process of indexing and index term weighting can be broken down into the following phases:

a) selection of terms which should be used to index a document; that is, deciding whether a term should or should not be assigned to a document,
b) deciding on the wight of importance of a term which has been assigned to a document,
c) deciding on the weight of importance of a term, which is present in a query, depending on the value of the term for searching. That is, the weight should indicate the usefulness of the term in separating the relevant documents from the non-relevant ones.

(i) Phase (a)

In systems where the indexing is performed automatically, the so-called full text indexing is commonly used [1]. With this approach all the words (except for a few common function words) present in a document or a document excerpt are assigned as index terms. Thus, the approach to the selection of terms is rather simplistic. Alternatively, the TP model, proposed by Bookstein, [2], and investigated by Harter, [3], can provide the basis for index term selection.

(ii) Phase (b)

For this phase several approaches have been proposed in the literature. If the indexing is done manually, the

indexers can assign a number to indicate how important a term is for a particular document. This type of weights was suggested by Maron and Kuhns, [4], in the context of their approach to indexing. In an automatic indexing environment, by far, the most popular type of weight is the within-document frequency weight (TF) proposed by Luhn, [5]. The suggestion is that the more often a term appears in a document the more important it is as an indicator of content. Other possibilities for determining term weights include the function proposed by Edmundson and Wyllys, [6], and the distribution separation measure of Harter, [3].

(iii) Phase (c)

Recent work on the probabilistic models [7-14] of retrieval represent the current state of knowledge in this phase of query term weighting. These models assume that relevance information can be made available. Typically, relevance information is obtained through a process called relevance feedback. That is, a number of promising documents with respect to a given query are retrieved using some basic search strategy. This search may be termed the initial search. The retrieved documents are then presented to the user for judgment as relevant or non-relevant. The information returned by the user following the initial search provides the basis for determining the importance of query terms.

The query term weights can also be obtained by methods that do not assume the availability of relevance information. In essence, these methods determine the value of a term in searching in terms no only of its occurrence frequencies within individual documents, but also its distribution in the collection of documents as a whole. Inverse document frequency (IDF) weights [15,16] and term discrimination value weights [17,18] have been considered attractive in this respect and have usually been the basis for initial search.

Croft and Harper, [19], demonstrate that, in fact, the probabilistic models of retrieval can guide us in determining appropriate term weights for searching. Specifically, they find that, when no relevance information is available, IDF-like weights provide a reasonable approximation of term relevance weights corresponding to the Binary Independence (BI) model [7,8]. Other such approximations have been identified by Yu et al., [21], and Robertson et al., [13].

In this paper possibilities of further enhancing the retrieval performance during an intial search are investiaged. The new weighting schemes considered are justified by unifying the arguments found in the works of Harter, [3], and Chow, [11]. In contrast to the results of Robertson el al., [13], results of this work imply that the TP model can lead to a better assignment of term weights for initial search than does the BI model. We also assess the appropriateness, in general, of using the TP model as a characterization of term frequency data.

The next section presents the motivation for this work. In section 3, earlier approaches on the basis of which the method we propose has evolved, are outlined. Section 4 develops the research plan and we perform the necessary experiments in section 5. The last two sections provide respectively an explanation of the findings of this work in terms of the performance of optimum (upper bound) retrieval strategies and the conclusions.

2. MOTIVATION

In information retrieval, a popular technique for performing initial sarch is to use some form of IDF weighting. Many experiments have been carried out with such weights and these weighting schemes have ben found to work quite well for a wide range of document collections.

In contrast, however, several researchers have shown that term weights for intial search could be derived from certain probabilistic models of retrieval, [11, 13, 19, 23]. Croft and Harper, [19], derive a form of IDF weights from the BI model. In Chow, [11], and Robertson el al., [13], a framework for deriving term weights from the TP model, assuming term independence and that no relevance information is available, is provided.

The experiments in [13] show, however, that the best results are obtained for an adhoc wighting scheme rather than that derivable for the TP model. We demonstrate in this work that such a discrepancy is likely to be due to inappropriate handling of certain types of terms.

Since the TP model incorporates additional information in the form of occurrence frequency distribution, we can also assess if term frequency data, as characterized by the TP model, represents an enhancement vis-a-vis the BI model.

3. RELATED EARLIER WORK

3.1 The 2-Poisson Model

The discussion here is based primarily on the two part paper by Harter, [3]. The emphasis of his work is on deciding whether a term is good enough to be assigned as an index term.

The 2-Poisson model comes about by the postulation that, for all specialty words, there exist exactly two levels of treatment of any specialty word. Furthermore, the 2-Poisson model assumes that any variation in the number of occurrences of such a word between

documents in the class corresponding to the same level of treatment is attributable to random fluctuation, as described by a Poisson distribution.

Notationally, therefore, the 2-Poisson model is characterized by two parameters u and v, representing the mean number of occurrences of the word in document classes I and II respectively, and a third parameter π , representing the proporation of documents in the collection which belong to class I. For concreteness we may assume that class I documents treat the subject to a relatively greater extent than do documents belonging to class II. Thus $u \geq v$, and the proporation of documents containing k occurrences of a particular specialty word is given by

$$f(k) = \pi \frac{e^{-u}u^k}{k!} + (1-\pi)\frac{e^{-v}v^k}{k!} \qquad (1)$$

The separation between the two Poisson distributions can be measured by

$$Z = \frac{u-v}{\sqrt{(u+v)}} \cdot \qquad (2)$$

Harter, then, identifies a measure of indexability which is a function of π and the probability that a document containing k occurences of a term belongs to class I.

3.2 Probabilistic models of retrieval -

Let $\underline{X}$ denote a document vector. It has been shown in [11], and several other papers, that an optimal retrieval rule, which has the objective of maximizing precision at any given level of recall, may be specified as

$$\text{retrieve } \underline{X} \text{ iff } \frac{P(\underline{X}|\text{Relevant})}{P(\underline{X}|\text{Non Relevant})} > \tau$$

for some threshold τ. The ratio of the two probabilities leads to the ranking of documents in descending order of their probability of relevance.

(i) Binary independence model

Each x_i, where $\underline{X} = (x_1, x_2, \ldots x_m)$ is assumed to be 1 or 0, depending on the presence or the absence of term i in the document. Furthermore, terms are assumed to be independently assigned within the set of relevant documents; the same asusmption is also made for the use of terms in the non-relevant documents. Then, we can show that

$$\frac{P(\underline{X}|\text{Relevant})}{P(\underline{X}|\text{Non Relevant})} = \frac{\prod_{i=1}^{m} p_i^{x_i}(1-p_i)^{(1-x_i)}}{\prod_{i=1}^{m} q_i^{x_i}\cdot(1-q_i)^{(1-x_i)}},$$

where p_i (respectively, q_i) is the probability that term i appears in a relevant (respectively, non-relevant) document. Applying log and dropping the term independent of $\underline{X}$, we find that ranking according to $P(\underline{X}/\text{Relevant})/P(\underline{X}/\text{Non Relevant})$ is equivalent to the ranking by

$$\sum_{i=1}^{m} x_i * \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad (3)$$

This summation, for reasons given in [9], is normally restricted to just the terms in the query.

(ii) TP independence model

Each document vector $\underline{X}=(x_1, x_2, \ldots x_m)$ is considered to be numeric, where x_i is the number of occurrences of term i in the document specified as $\underline{X}$. The frequencies of occurrences of terms, as in the earlier model, are assumed to be independently assigned in the set of documents in class I (and also in the set of documents in class II). Chow, [11], and Chow and Yu, [12], derived a ranking strategy based on this model, with the additional assumption that documents in class I correspond to relevant documents, and those in class II to non-relevant documents. Thus, the expected value of x_i among relevant (respectively non-relevant) documents is given by u_i (respectively v_i). Applying the same, optimal retrieval rule as in the binary independence model,

$$\log \frac{P(\underline{X}|\text{Relevant})}{P(\underline{X}|\text{Relevant})} = \log \frac{\prod_{i=1}^{m} u_i^{x_i} * e^{-u_i}/x_i!}{\prod_{i=1}^{m} v_i^{x_i} * e^{-v_i}/x_i!}$$

$$= \sum_{i=1}^{m}(v_i-u_i) + \sum_{i=1}^{m} x_i \log(u_i/v_i).$$

Ignoring the first term, which is not dependent on the document terms and restricting the function only to query terms, we can express the ranking function as

$$\sum_{\text{term } i \in \text{ query}} x_i \log(u_i/v_i) \qquad (4)$$

A function commonly encountered in information retrieval is the simple matching function defined as $\sum_i d_{ij} \cdot w_i$, where d_i and w_i are respectively the weights of the i^{th} term in document j and a given query. Thus, using ranking functions (3) or (4) for retrieval is equivalent to applying th simple matching function between the query and the

documents, where query term i has the weight $\log p_i(1-q_i)/q_i(1-p_i)$ or $\log u_i/v_i$, respectively.

Robertson et al., [13], considered the general case of the TP independence model in which the relevant and the non-relevant documents with respect to a query are not assumed to correspond to the two classes of documents over which the two Poisson distributions are defined for each query term. The expression obtained for the weight of a query term however involves several components and its computation is complicated.

4. RESEARCH AIMS AND METHODOLOGY

4.1 Aims

There are essentially three related aims. As far as the initial search is concerned, there is, first, the issue of whether IDF weighting is the best strategy. Secondly, we would like to establish whether term frequency, as modelled by the TP distribution, represents useful additional information. In other words, since TP model is the best known model of term frequency data, the question we wish to answer may be stated as: is it possible to obtain better performance by using TF?

The third aim is of a more general nature. We would like to determine whether the TP model is accurate. That is, given the fact that TP model aims to incorporate additional information, in comparison to the BI model, is it effective in doing that.

4.2 Methodology

In order to perform the investigations that we are interested in, we require a particular IDF weighting function.

Croft and Harper, [19], suggest an approach for initial search, based on BI model, that corresponds to the use of IDF weights. The ranking function (3) is rewritten as

$$\sum_{\text{term } i \in \text{query}} x_i \log\frac{p_i}{(1-p_i)} + \sum_{\text{term } i \in \text{query}} x_i \log\frac{q_i}{(1-q_i)} \quad (5)$$

Since no relevance information is available, $\log p_i/(1-p_i)$ is assumed to be a constant, C. Then, it is suggested that if the total collection of documents can be approximated to correspond to the set of non-relevant documents, and that q_i can be estimated by n_i/N, where n_i is the number of documents in which term i occurs (the document frequency) and N is the size of the collection. The expression (5) becomes $\sum_i x_i \cdot (\log((N-n_i)/n_i)+C)$.

For large N this can be written as

$$\sum_{\text{term } i \in \text{query}} x_i \ (\log(N/n_i)+C) \quad (6)$$

Clearly, this ranking function is equivalent to performing simple match retrieval with the weight of query term i, w_i, specified as $\log(N/n_i)+C$.

In order to develop a ranking function for initial search using the TP independence model, we begin with the parameter estimation scheme suggested by Harter, [3].

(i) Term weights based on Harter's parameter estimation

In section 3.2 (ii), it was shown that $\log(u/v)$ is an optimum term weight derivable, for a term in query, on the basis of the TP model, where u and v are the mean occurrence frequencies of the term respectively in the set of relevant and non-relevant documents.

For each term, estimators for u,v and π can be obtained by the method of moments. The values of u and v are shown to be the roots of the quadratic equation

$$ax^2 + bx + c = 0, \quad (7)$$

where the smaller root is taken to be v and the larger one is u. Let R_1, R_2 and R_3 denote the first three sample moments (about the origin) computed from the frequency data corresponding to a given term in an experimental document collection. Then, the coefficients in eq. (7) are

$$a = R_1-L,\ b = K-LM \text{ and } c = L^2-MK,$$

where $L = R_2-R_1$ and $K = R_3+2R_1-3R_2$.

Knowing u and v, π is given by

$$\pi = \frac{R_1-v}{u-v} \quad (8)$$

For the values to be meaningful in the context of language use, it is necessary that $u,v \geq 0$ and $0 \leq \pi \leq 1$. The degenerate cases are, therefore, handled as follows:

1. if $b^2-4ac \leq 0$, then set $u=R_1$ and $v=0$
2. if $v < 0$ then
   {set $v=0$
   if $L/R_1 < R_1$ then set $u=R_1$ else set $u=L/R_1$}
3. if $u<R_1$ or $v>R_1$ then set $u=R_1$ and $v=0$.

(ii) Proposed modifications to Harter's estimates

Harter's objective in providing a method to compute u,v and π was to obtain a measure that facilitates the decision as to the use or otherwise of a term as an index term. It appears, however, there are problems in directly using those estimates for determining term weights. In fact, Robertson et al., [13], use Harter's estimates and peform a retrieval experiment using function (4). This function was found to give worse results than an IDF function.

While there are several possible causes for the poor performance of Harter's estimates, it is felt that the way in which the degenerate cases are dealt with is a significant one. When the parameters u,v and π are computed it is desired that $v \geq 0$ and $u \geq R_1 \geq v$, since in such cases the 2-Poisson distribution represents a reasonable model of the terms' usage. Unfortunately, for the CRN4NUL* Collection, it was found that only 14.5% of the total of 2651 terms had the parameter values in the proper range. The corresponding percentage for the MEDNUL* Collection, which has 4726 terms, is 9.25. Thus, it should be expected that poor handling of degenerate cases will overshadow any performance gain that might otherwise be realized. Consequently, we consider the question of how one might handle the degenerate cases.

(ii.a) IDF approximation for degenerate cases

In this modification, the effect of degenerate cases is suppressed by employing weights derived from the TP model only if parameters estimates fall parameter estimates fall within the proper range. Thus, for term i, we have

$$w_i = \begin{cases} \log (u_i/v_i), & \text{if parameters are in proper range} \\ \log (N/n_i)+C & \end{cases} \quad (9)$$

Note that the weight for degenerate cases is the same as that derived earlier in this section and labelled as expression (6).

(ii.b) π-Approximation for degenerate cases

Instead of simply using IDF weights for degenerate cases, it is preferrable to specify weights which are derived from information such as the first moment, second moment, etc. By doing this we would better use all the available information.

Our overall approach in deriving this modification consists of expressing the ratio of the document frequency of the i^{th} term to the number of documents in the collection, n_i/N, in terms of the parameters in the TP model. Consider the 2-Poisson density function

$$f(k) = \pi_i \frac{e^{-u_i} u_i^k}{k!} + (1- \pi_i) \frac{e^{-v_i} v_i^k}{k!}$$

where u_i, v_i and π_i are the parameters and f(k) is the probability that term i has an occurrence frequency of k.

It is easy to see that $n_i = N \cdot \sum_{k>0} f(k)$. Thus,

$$n_i/N = \sum_{k>0} f(k) \quad (10)$$

In section 4.2(i) the method of parameter estimation proposed by Harter was presented. That method suggested that whenever a term is degenerate the value of the parameter v_i be set to 0. Thus, we may substitute the value 0 for v_i in eqn. (10) and obtain

$$n_i/N = \pi_i \cdot \sum_{k>0} \frac{e^{-u_i} u_i^k}{k!}$$

$$= \pi_i \cdot \left(1 - \frac{e^{-u_i} u_i^0}{0!}\right)$$

$$= \pi_i \left(1 - e^{-u_i}\right) \quad (11)$$

Given our assumption, that class I and II correspond respectively to the class of relevant and nonrelevant documents, in the derivation of ranking function (4),

$$f(k/\text{Document is relevant}) = \frac{e^{-u_i} u_i^k}{k!} .$$

By summing the above expression over all $k>0$, the probability that term i appears in a relevant document is determined to be $1-e^{-u_i}$. Thus, using the notation used in the BI model for this probability (i.e. denoting $1-e^{-u_i}$ by p_i) and the expression (11), we obtain

$$\log (N/n_i) = \log (1/ \pi_i) + \log (1/p_i) \quad (12)$$

For the initial search, since we do not have any relevance information, we may assume p_i to be the same for all i [19] and write eqn. (12) as

$$\log (N/n_i) = \log (1/ \pi_i) + C \quad (13)$$

In terms of the scheme in Harter's work for degenerate cases, two subcases

* These are the collections used in the experiments to be described. Further details on these collections are provided in the next section.

arise. First, is the subcase in which $v<0$ and $L/R_1>R_1$. Here, u and v are set respectively to L/R_1 and 0 and, from eqn. (8), it follows that

$$\pi = \frac{R_1}{L/R_1} = R_1^2/L \qquad (14)$$

In the other subcase $v<0$ and $L/R_1 \leq R_1$. Using eqn.(8), π works out to 1. Although this value may be quite appropriate in general, under our assumption that class I is the relevant class, $\pi = 1$ would imply that all documents are relevant. This is, clearly, not reasonable. We looked at the characterstics of these terms in more detail using experimental data and found that most of these terms appear with a term frequency of 1. In other words, the document frequency of these terms, n, is very close to the sum of the term frequencies, $N \cdot R_1$. As far as the value of v is concerned, we may follow Harter's suggestion and set it to zero, implying that the term does not appear in any of the non-relevant documents. Thus, the proportion of relevant documents, π , is at least n/N, which approximately equals R_1. For these reasons, for the subcase $v<0$ and $L/R_1 \leq R_1$, we propose that

$$\pi = R_1 \qquad (15)$$

Substituting eqns. (14) and (15) into (13), we have

$$w_i = \begin{cases} \log(u_i/v_i), \text{ if parameters are in proper ranage} \\ \log(L_i/R_{1i}^2)+C, \quad \text{if } v_i<0, \text{ and } L_i/R_{1i} > R_{1i} \\ \log(1/R_{1i})+C, \text{ otherwise} \end{cases} \qquad (16)$$

The investigations relating to the third objective are carried out by comparing the performance of the BI model to that of the TP model, assuming that complete relevance information is available. That is, we consider the best results achievable using the BI model against the corresponding results based on the TP model. These results will also be used to explain the variation observed in the performance of initial search strategies from one document collection to another.

5. EXPERIMENTS

5.1 General specifications

Two collections of documents are used for these experiments; the first is the collection of 424 documents in aerodynamics which was prepared from the abstracts of the documents used by the Cranfield Project (CRN4NUL), the other is the Medlar Collection of 450 documents in biomedicine (MEDNUL). For each of these collections a query collection of 24 queries is selected from those available within the SMART retrieval system [17]. These documents and queries have also been used by Salton and Yang, [16], and Salton et al., [23]. These collections include, for evaluation purposes, information for each query as to which of the documents are relevant. The standard recall and precision measures are used for comparing the performance of different strategies for weighting index terms. Recall is defined as the proporation of relevant documents retrieved and precision is the proportion of the retrieved documents actually relevant. The overall performance of a strategy is determined by processing the queries with that strategy and computing the average precision over all the queries for recall values 0.1, 0.2, ... and 1. The algorithm for averaging is consistent with that implemented in the SMART system.

The comparison of one method with another is accomplished by presenting the percentage improvement of both relative to a base strategy. The simple matching technique, in which both documents and queries are treated as binary vectors is used as the base. Since this method is also known as coordination level matching, the column of precision values corresponding to this method will be labelled COORD.

5.2 Experimental Evaluation of the Proposed Modifications

The first set of experiments are aimed at providing a perspective for the subsequent experiments with our proposed term weights. The results are presented in Table 1.

The columns labelled CH correspond to the use of function (6). Croft and Harper suggest that best results were obtained for $C \approx 1$. We, therefore, chose C=1. The results, which use the Harter's method of parameter estimation and handling degenerate cases, are labelled H with TF and H without TF. When TF is used, the weight function is that defined in (4); in the other case (i.e. TF is not used), x_i is restricted to 0 or 1 and the weight is simply $\log(u_i/v_i)$. The end conditions are handled as follows:

$$w_i = \begin{cases} 0 \text{ if } u_i = 0 \\ 9999 \text{ if } v_i = 0 \end{cases}$$

For the CRN4NUL Collection CH weights are 8 to 15 percent better than H; the corresponding figures for the MEDNUL are 15 to 23 percent. It is also important to note, although according to theory TF should be used, H without TF is better than H with TF for both the collections. These results are consistent with those

obtained by Robertson et al. [13]. Thus, when the degenerate cases are not handled properly, the weight function derived from the TP model is not as effective as CH.

Tables 2 and 3 exhibit the performance of the methods proposed here for dealing with degenerate cases. The IDF approximation given by eqn. (9) and is labelled IDF-APRX and the other modification, specified by eqn. (16) and is based on the generalization of IDF, is labelled Π-APRX. For both IDF-APRX and Π-APRX, experiments were performed for C values of 0,1,2, and 3.

For the CRN4NUL Collection (table 2), the performance of both IDF-APRX and Π-APRX are better without TF, than with TF. These two methods, when used without TF, are better than CH. The performance of IDF-APRX and Π-APRX are, however, quite sensitive to the value of C. When compared on a consistent basis, that is C=1 as in CH, IDF-APRX and Π-APRX are only marginally better than CH, with the percentage improvements ranging from 1.3 to 1.7%. Among the C values tested, the best performance for IDF-APRX and Π-APRX are obtained respectively for C=2 and C=3. The corresponding percentage improvements over CH for these choices are 5.4 and 8.5.

The results for MEDNUL Collection appear in table 3. The performance figures with TF were worse than those for without TF in this case also, but are not reported. The results obtained were best for C=1 and IDF-APRX and Π- APRX are better than CH, respectively, by 5.7% and 7%. Thus, for both collections, when the degenerate cases are better handled, the functions based on the TP model outperform CH.

5.3 Comparison With Other Earlier Methods

Besides the IDF function (labelled CH) - originally due to Sparck Jones - tested in the previous section, many other functions have been investigated in the literature. In this section we select several important ones among them for comparison with our proposal. The functions selected are:

(i) Cosine Similarity,

(ii) the version of IDF used in several papers by Salton,

which is

$$w_i = TF_i/n_i,$$

and

(iii) the function found to give very good results by Robertson et al., given by

$$w_i = Z_i * \log (N/n_i)$$

where Z_i is the separation measure defined in eqn. (2).

The functions of (i) and (ii) above can be viewed as having both a query term weight component (query TF_i, or $1/n_i$) and a document term weight component (document TF_i). The function proposed by Robertson et al., on the other hand, is a hybrid query term weight that combines Z, derived from Harter's measure of term quality, with $\log(N/n_i)$. Since Π-APRX and IDF-APRX do not perform well with TF, it seems worthwhile to consider a more complex query term weight which combines Z with one of our functions. Thus, the weight function Π-APRX * Z, with C=3, is evaluated against the functions listed above.

The results are presented in table 4. The functions by Salton and Robertson et al. are labelled respectively S and RVP. It is interesting that the various weight functions considered here impact the two collections very differently. When compared to CH, all the functions in table 4 are either as good or better for CRN4NUL Collection; they are all worse for the MEDNUL collection. For the CRN4NUL Collection, Π-APRX * Z is better than the other functions by between 9% (for RVP) to 21% (for Cosine). The best result for the MEDNUL Collection corresponds to Π-APRX without Z or TF.

6. DISCUSSION

In order to better understand and explain the results of the earlier section, we decided to perform retrospective searches in which actual parameter values are used. These are upper bound results in the sense that they correspond to the performance when complete relevance information is assumed.

For the BI model, the number of relevant and non-relevant documents in which term i appears, r_i and s_i respectively, with respect to a given query are computed. Then,

$$w_i = \frac{r_i + c'}{|R| - r_i + c'} \Big/ \frac{s_i + c'}{|I| - s_i + c'} .$$

In the case of TP model, for a given query, the average occurrence frequency in the relevant and non-relevant sets, u_i and v_i, are specified as follows:

$$u_i = [(\sum_{j \in R} f_{ij}) + c'] \; / \; (|R| + c')$$

$$v_i = [(\sum_{j \in R} f_{ij}) + c'] \; / \; (|I| + c')$$

In the above, R and I are respectively the set of relevant and non-relevant documents and c' is used to handle end conditions. It was found that the smaller the value of c', the better is the retrieval performance. The results reported in table 5 correspond to c' = 1.0 E -30.

Table 5 shows that, for CRN4NUL collection, the TP model with TF performs better than TP without TF. In contrast, using TF does not help in the MEDNUL Collection. This aspect explains why the various functions tested in table 4, which use document term weight component, perform better than CH for CRN4NUL Collection but not for MEDNUL. The results obtained for TP with Z is, however, surprising and does not help explain the effect, of using Z with Π-APRX, observed in table 4.

Comparing BI model to TP model with TF, it is seen that for the CRN4NUL and MEDNUL Collections the BI model is better respectively by 7% and 20%. Although it is difficult to make conclusive statements because of the sensitivity of these results to c', it appears that the TP model is still not exploiting the term frequency data well enough.

Based on our findings, it is believed that the use of Π-APRX or IDF-APRX as the query term weight is better than using CH. Deriving this weighting scheme, however, only solves a part of the problem. That is, according to the ranking function (4) obtained on the basis of the TP model, a document term weight component must also be employed. Unfortunately, the experiments above show that the use of the latter component leads to performance degradation and, in that sense, contradict the theory on which ranking function (4) is based.

When this paper was in final stages, we came to know of recent work by Croft, [22], in which a particular method of exploiting within-document frequency information has been shown to be effective. We felt it would be interesting to see if that method might help resolve the second part of the problem, explained above.

Croft, [22], suggests that the within-document frequency information should be normalized before it is used to derive document term weights. Specifically, if TF_{di} is the within-document frequency of term i in document d, then the normalized within-document frequency (denoted, n_{di}) is calculated as $TF_{di}/\max\{TF_{d1}, TF_{d2}, \ldots,\}$. Then, the document term weight is given by $K+(1-K)n_{di}$, where K is a constant between 0 and 1. This weight is termed as the term significance weight.

We performed a number of experiments, which essentially correspond to repeating experiments already reported except that the document term weight proposed by Croft are used in place of TF. These are presented in table 6. For both CRN4NUL and MEDNUL collections, N-TF (TF_{di}'s normalized according to Croft) is used in conjunction with CH, CR, Π- APRX and TP. CH and Π - ARPX are as previously defined. TP stands for, as in table 5, log (u/v) computed using complete relevance information. CR is a variation of ranking function (6), adopted in [22], given by

$$\text{term} \sum_{i \in \text{query}} x_i \ (\log((N-n_i)/n_i)+C).$$

For the CRN4NUL collection, the above experiments were performed for K values of 0.3, 0.5 and 0.7. However, the results were not really sensitive to K. Therefore, table 6 only reports results for K=0.5. For MEDNUL collection the experiments were performed only for K=0.5.

It is worth noting, at this point, that for the collections used here, the term weights specified in the documents are all multiples of 12. Thus, if TF_{di} is the term weight, then $TF_{di}/12$ is the number of times the term i appears in document d. Since this interpretation is employed in all our experiments, the earlier results labelled 'with TF' actually mean that document term weight is defined as $TF_{di}/\min\{TF_{d1}, TF_{d2}, \ldots\}$.

First we see, from tables 3,4, and 6, that CH*N-TF is better than CH by 27.6% and 10.7% respectively for CRN4NUL and MEDNUL collections. The results based on complete relevance information are also encouraging. Specifically, TP*N-TF is better than the BI model (table 5) by 20.4% and 13.4% respectively for the two collections. These results suggest not only that the term significance weights proposed by Croft are effective, but also that, with proper specification of document term weights, the TP model has the potential to yield better retrieval strategies than the BI model. This statement should however be viewed in the light of the fact that only 14.5% (for CRN4NUL) and 9.25% (for MEDNUL) of terms properly fit the 2-Poisson distribution. The results corresponding to CR*N-TF are presented since this is the function adopted in [22].

Comparing the results for Π- APRX*N-TF to those of CR*N-TF, it is seen that the former is significantly better (7.3%) for the CRN4NUL collection. In the case of MEDNUL collection, Π- APRX*N-TF is only better than CR*N-TF at recall values above 0.5 and, on the average, the performance is just marginally better.

Although in some cases the percentage improvements for the proposed scheme are small, they must be viewed in the light of earlier results (table 1) which concluded that weights based on TP model lead to significant degradation in performance. Furthermore, the cases where the improvement is not significant only point to the fact that one should search for better methods of computing the parameters u, v and π and of handling degenerate cases. This is because, the retrospective experiments (using complete relevance feedback) indicate that there is still room for improvement.

7. CONCLUSION

Salton, [1], recommends the use of IDF weighting to perform the initial search. It has also been argued, in the context comparing IDF to weighting schemes that lead to a different shape of weight distribution, that the performance of IDF is not likely to be improved upon, [20]. In this work, improvements to a weighting scheme based on the TP model are suggested and it is shown that this modification obtains better results, for the collections tested, than the weighting schemes currently in use. Our results indicate further research is needed not only in the direction of obtaining better tem weights based on the TP model, but also in the development of alternate models of within-document frequency information.

REFERENCES:

1. G. Salton, A blueprint for automatic indexing, ACM-SIGIR Forum, vol. XVI, no. 2,, 1981, pp. 22-38.

2. A. Bookstein and D.R. Swanson, Probabilistic models for automatic indexing, Journal of the American Society for Information Science, vol. 25, 1974, pp. 312-319.

3. S.P. Harter, A probabilistic approach to automatic keyword indexing, Journal of the American Society for Information Science, vol. 26, 1975, Part I: pp. 197-205, Part II: pp. 280-289.

4. M.E. Maron and J.L. Kuhns, On relevance, probabilistic indexing and information retrieval, Journal of the ACM, vol. 7, 1960, pp. 216-244.

5. H.P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of research and development, vol. 1, no. 2, Oct. 1957, pp. 309-317.

6. H.P. Edmundson and R.E. Wyllys, Automatic abstracting and indexing-survey and recommendations, Communications of the ACM, vol. 4, no. 5, May 1961, pp. 226-234.

7. C.T. Yu and G. Salton, Precision weighting - an effective automatic indexing method, Journal of the ACM, vol. 23, 1976, pp. 76-88.

8. S.E. Robertson and K. Sparck Jones, Relevance weighting of search terms, Journal of the American society for Information Science, vol. 27, 1976, pp. 129-146.

9. C.J. Van Rijsbergen, A theoretical basis for the use of co-occurrence data in information retrieval, Journal of documentation, vol. 33, 1977, pp. 106-119.

10. D.J. Harper and C.J. Van Rijsbergen, An evaluation of feedback in document retrieval using co-occurence data, Journal of documentation, vol. 34, 1978, pp. 189-216.

11. D. Chow, Constructing feedback queries, M.Sc. thesis, 1978, University of Alberta, Edmonton.

12. D. Chow and C.T. Yu, Constructing feedback queries, Journal of the ACM, vol. 29, no. 1, January, 1982, pp. 127-151.

13. S.E. Robertson, C.J. Van Rijsbergen and M.F. Porter, Probabilistic models for indexing and searching. Third international converence on information storage and retrieval, 1980, Cambridge, England.

14. S.E. Robertson, M.E. Maron and W.S. Cooper, Probability of relevance: A unification of two competing models of document retrieval, Information Technology: Research and Development, vol. 1, no. 1, 1982, pp. 1-21.

15. K. Sparck Jones, A statistical interpretation of term specificity in retreival, Journal of documentation, vol. 28, no. 1, March 1972, pp. 11-21.

16. G. Salton and C.S. Yang, On the specification of term values in automatic indexing, Journal of documentation, vol. 29, no. 4, Dec. 1973, pp. 351-372.

17. G. Salton, Dynamic information and library processing, Prentice-Hall, Englewood Cliffs, New Jersey, 1975, chapter 3.

18. G. Salton, C.S. Yang and C.T. Yu, Contribution to the theory of indexing, Information Processing, vol. 74, 1974, pp. 584-590, Amsterdam: North Holland Publishing Co.

19. W.B. Croft and D.J. Harper, Using probabilistic models of document retrieval without relevance information, Journal of documentation, vol. 35, 1979, pp. 285-295.

20. H. Wu and G. Salton, A comparison of search term weighting: term relevance vs. inverse document frequency, Proceedings of the fourth international conference on information storage and retrieval, ACM-SIGIR Forum, vol. XVI, no. 1,

1981, pp. 30-39.

21. C.T. Yu, K. Lam and G. Salton, Term weighting in information retrieval using the term precision model, Journal of the ACM, vol. 19, no. 1, January 1982, pp. 152-170.

22. W.B. Croft, Experiments with representation in a document retrieval system. Technical Report 82-21, COINS, Univ. of Massachusetts, Amherst, Mass., 1982.

23. G. Salton, H. Wu and C.T. Yu, The measurement of term importance in automatic indexing. Journal of the American Society for Information Science, Vol. 32, no. 3, May 1981, pp. 175-186.

Table 1

Evaluation of term weights based on Harter's parameter estimates.

R	CRN4NUL				MEDNUL			
	COORD	CH	H with TF	H without TF	COORD	CH	H with TF	H without TF
0.1	0.5668	0.7209	0.5042	0.6382	0.6977	0.7426	0.6947	0.669
0.2	0.4031	0.5637	0.4327	0.4824	0.6139	0.6988	0.6061	0.5922
0.3	0.3178	0.4462	0.4118	0.4262	0.5142	0.6427	0.5287	0.5667
0.4	0.2853	0.403	0.3709	0.3661	0.4611	0.6277	0.4782	0.5435
0.5	0.2376	0.3437	0.3032	0.3129	0.4156	0.5795	0.4321	0.5153
0.6	0.2086	0.2899	0.2447	0.2447	0.3429	0.4841	0.3542	0.415
0.7	0.1439	0.2135	0.2167	0.1811	0.2865	0.4116	0.3214	0.3479
0.8	0.1129	0.1702	0.1548	0.1495	0.2032	0.3015	0.231	0.2518
0.9	0.0787	0.1266	0.137	0.1329	0.150	0.2303	0.1957	0.2204
1.0	0.0722	0.1201	0.1187	0.1261	0.0865	0.1569	0.1537	0.1585
		45.9%	32.8%	37.8%		38.9%	15.2%	24.2%

Table 2

Evaluation of the proposed modications to term weights using the CRN4NUL collection.

R	CH	IDF - APRX.			Π-APRX.		
		with TF C=1	without TF		with TF C=1	without TF	
			C=1	C=2		C=1	C=3
0.1	0.7209	0.6375	0.7305	0.7169	0.6329	0.7082	0.7280
0.2	0.5637	0.4877	0.5608	0.5406	0.4959	0.5565	0.5450
0.3	0.4462	0.4253	0.4807	0.4517	0.4234	0.4705	0.4708
0.4	0.4030	0.3813	0.4444	0.4237	0.3775	0.4295	0.4307
0.5	0.3437	0.3356	0.3710	0.3594	0.3343	0.3656	0.3657
0.6	0.2899	0.2777	0.3055	0.3012	0.2791	0.3106	0.3054
0.7	0.2135	0.2292	0.2071	0.2220	0.2239	0.2134	0.2263
0.8	0.1702	0.1532	0.1661	0.1818	0.1471	0.1745	0.1856
0.9	0.1266	0.0969	0.1185	0.1360	0.0966	0.1194	0.1407
1.0	0.1201	0.0828	0.1104	0.1276	0.0835	0.1132	0.1307
	45.9%	30.8%	47.2%	51.3%	29.9%	47.6%	54.4%

Table 3

Evaluation of the proposed modications to term weights using the MEDNUL collection.

R	CH	without TF, C=1	
		IDF-APRX	Π-APRX
0.1	0.7426	0.7493	0.7504
0.2	0.6933	0.7078	0.6915
0.3	0.6427	0.6366	0.6353
0.4	0.6277	0.6285	0.6261
0.5	0.5795	0.5700	0.5924
0.6	0.4841	0.4906	0.4997
0.7	0.4116	0.4364	0.4456
0.8	0.3015	0.3298	0.3380
0.9	0.2303	0.2538	0.2551
1.0	0.1569	0.1722	0.1723
	38.9%	44.6%	45.9%

Table 4

Comparison of a proposed function with several known weight functions.

R	CRN4NUL				MEDNUL			
	COSINE	S	RVP	Π-APRX with Z	COSINE	S	RVP	Π-APRX with Z
0.1	0.723	0.6943	0.7245	0.7346	0.7191	0.7826	0.7528	0.737
0.2	0.5754	0.6173	0.5808	0.5895	0.6933	0.6841	0.6815	0.6666
0.3	0.4433	0.525	0.4868	0.5141	0.6192	0.6313	0.6144	0.5716
0.4	0.3977	0.4425	0.4442	0.4702	0.5927	0.5831	0.5969	0.560
0.5	0.3538	0.3644	0.387	0.4025	0.5607	0.5468	0.5417	0.513
0.6	0.2753	0.2781	0.3282	0.3545	0.4805	0.4708	0.448	0.4441
0.7	0.1986	0.2309	0.2423	0.2548	0.4156	0.4014	0.4025	0.3887
0.8	0.1703	0.1827	0.1892	0.2067	0.3101	0.2829	0.3088	0.3159
0.9	0.1317	0.1392	0.1356	0.144	0.2277	0.2089	0.2221	0.2255
1.0	0.1253	0.1123	0.1278	0.1373	0.1629	0.1618	0.1471	0.1559
	46.0%	53.8%	58.6%	67.5%	37.8%	34.8%	33.9%	32.2%

Table 5

Evaluation of different ways of implementing TP model relative to BI model, using weights based on complete relevance information derived retrospectively.

R	CRN4NUL				MEDNUL			
	BI	TP			BI	TP		
		with TF	without TF	with Z		with TF	without TF	with Z
0.1	0.8620	0.7973	0.8794	0.8157	0.8811	0.9541	0.8793	0.8557
0.2	0.7590	0.7313	0.7893	0.7302	0.8521	0.8642	0.8590	0.8316
0.3	0.6673	0.6569	0.6232	0.6289	0.8164	0.8274	0.7907	0.7952
0.4	0.6237	0.5857	0.5949	0.5933	0.8088	0.7996	0.7835	0.7880
0.5	0.5726	0.5193	0.5207	0.5297	0.7783	0.7225	0.7390	0.7254
0.6	0.4330	0.4316	0.4073	0.4277	0.7143	0.6524	0.6863	0.6545
0.7	0.3471	0.3563	0.3200	0.3338	0.6742	0.5712	0.6502	0.6276
0.8	0.2668	0.2484	0.2387	0.2446	0.5290	0.4492	0.5253	0.5180
0.9	0.1918	0.1924	0.1848	0.1793	0.4056	0.3191	0.3970	0.4052
1.0	0.1710	0.1687	0.1657	0.1649	0.2780	0.2345	0.2680	0.2731
	117.8%	110.5%	106.8%	106.4%	103.2%	88.0%	102.8%	101.2%

Table 6

Evaluation of document term weights defined as $K+(1-K)n_{di}$, where n_{di} is the normalized within-document frequency, used in conjunction with various query term weights.

R	CRN4NUL K=0.5				MEDNUL K=0.5			
	CH* N-TF C=1	CR* N-TF C=1	Π-APRX* N-TF C=3	TP* N-TF	CH* N-TF C=1	CR* N-TF C=1	Π-APRX* N-TF C=1	TP* N-TF
0.1	0.7740	0.7725	0.7718	0.9485	0.8419	0.8419	0.8074	0.9460
0.2	0.6527	0.6535	0.6721	0.8670	0.7667	0.7667	0.7606	0.9250
0.3	0.5898	0.5978	0.5766	0.7527	0.7196	0.7199	0.6851	0.8430
0.4	0.4845	0.4870	0.4957	0.7000	0.6495	0.6498	0.6422	0.8069
0.5	0.4057	0.4119	0.4432	0.6110	0.6142	0.6138	0.6133	0.7891
0.6	0.3366	0.3406	0.3568	0.4924	0.5124	0.5103	0.5534	0.7577
0.7	0.2650	0.2674	0.2862	0.3864	0.4388	0.4398	0.4635	0.6933
0.8	0.2059	0.2065	0.2242	0.2599	0.3072	0.3104	0.3362	0.5619
0.9	0.1549	0.1556	0.1647	0.2091	0.2564	0.2567	0.2572	0.4384
1.0	0.1352	0.1359	0.1434	0.1878	0.1709	0.1759	0.1793	0.3197
	73.5%	74.8%	82.1%	138.2%	49.6%	50.4%	52.7%	121.6%