Two Learning Schemes in Information Retrieval*

Clement T. Yu and Hirotaka Mizuno
Department of Electrical Engineering & Computer Science
University of Illinois at Chicago
Chicago, Illinois 60680

Two methods are given to improve weighting schemes by using relevance information of a set of queries. The first method is to estimate parameter values of two independence models in information retrieval --- the binary independence model and the non-binary independence model. The parameters estimated here are used to calculate optimal weights for terms in a different set of queries. Performance of this estimation is compared to the inverse document frequency method, the cosine measure, and the statistical similarity measure. The second method is to learn optimal weights of the non-binary independence model adaptively by a learning formula. Experiments are performed on three different document collections CISI, MEDLARS, and CRN4NUL for both methods, and results are reported. Both methods show improvements compared to the existing weighting schemes. Experimental results show that the second method gives slightly better performance than the first one, and has simpler implementation.

1. Introduction

In information retrieval systems, it is common to compute a similarity between a query and a document for retrieval of documents. Many methods of calculating similarities have been presented. Two well-known methods are the inverse document frequency method [Spar] and the cosine measure [SaMc]. Both methods are simple and intuitive. The inverse document frequency method assigns a higher weight to a term which occurs in few documents than a term in many documents. The similarity measure between a document and a query is simply the sum of the weights of the query terms in the document. The cosine measure is the cosine of the angle between the query and the document when they are represented as vectors. In spite of their intuitive appeal, they can not ensure optimality in retrieval.

The binary independence model [YuSa][RoSp] and the non-binary independence model[YuLe] are theoretical models which have been proposed to guarantee optimal retrieval. Optimality here means that a document with higher probability of relevance to a query is assigned a higher similarity than a document with lower probability of relevance. These two models are theoretical because they suppose a priori knowledge about some parameters used in calculation of similarities.

In this paper, the parameter values of these two independence models are estimated from document and term frequencies of terms. The estimation method of the parameter values is discussed and the performances of the estimated independence models are compared to that of the inverse document frequency method, the cosine measure, and the method which has been recently reported [WoYa] and called the statistical similarity measure. In section 2, outlines of two independence models are given. In section 3, we discuss the estimation method of the parameters, and introduce some constraints which should be satisfied by the estimated parameter values and the weighting functions derived from the parameters. The estimated independence models are applied to three different document collections and the performances are shown in section

This research is supported in part by National Science Foundation under IST-8501091 and IRI-8702177 and in part by OCLC.

4. In section 5, motivations of a learning scheme are described. In section 6, the learning scheme is given and its properties are discussed. In section 7, experimental results of the second method are given. The conclusions of the two methods are given in section 8.

2. Two Independence Models

2.1. Binary Independence Model

The binary independence model[RoSp][YuSa] ensures the optimality of retrieval described below.

Let p(R|D_i) be the probability that a document D_i is relevant to the query and f(Q,D_i) be the similarity between a query Q and a document D_i, where "R" in $p(R \mid D_i)$ stands for "relevance". Then the following condition is satisfied by the binary independence model.

$$f(Q,D_i) > f(Q,D_j) \longleftrightarrow p(R \mid D_i) > p(R \mid D_j)$$

The binary independence model makes the following two assumptions.

(1) Binary assumption : All document vectors are binary vectors. In other words, each element d_{ij} is 0 or 1. 0 means the absence of the term T_j in document D_i and 1 means the presence of the term T_j in document D_i.

(2) Independence assumption : All terms are statistically independent within the relevant document set and within the irrelevant document set.

Based on the above two assumptions, the weight assigned to term T_j is given by [RoSp]

$$w_1(T_j) = \log\frac{p_j}{1-p_j} - \log\frac{q_j}{1-q_j} \qquad (2.1)$$

where p_j is the probability that a relevant document has the term T_j and q_j is the probability that an irrelevant document has the term T_j. For a query Q having the following terms,

$$Q = (T_{k1}, T_{k2}, \cdots, T_{kl})$$

and for a document $D_i = (d_{i1}, d_{i2}, \cdots, d_{iM})$ the similarity between the two vectors is

$$sim(Q,D_i) = \sum_{n=1}^{l} d_{kn} . w_1(T_{kn}),$$

where d_{kn} is 0 or 1 because of the binary assumption.

In deriving the weighting function (2.1), we suppose a priori knowledge of p_j and q_j. These probability values are unknown in advance under ordinary circumstances. Our aim is to estimate p_j and q_j from some other parameters which are easily obtainable.

2.2. Non-Binary Independence Model

In the non-binary independence model, the independence assumption in the previous section is retained while the binary assumption is discarded. In other words, we take into account the term frequency of each term in each document. Then the optimal weight of a term T_j with frequency of occurrences d_{ij} in a document D_i is calculated by the following formula[YuLe].

$$w_2(T_j,d_{ij}) = \log\frac{p(tf=d_{ij} \mid R)}{p(tf=d_{ij} \mid I)} - \log\frac{p(tf=0 \mid R)}{p(tf=0 \mid I)} \qquad (2.2)$$

where p(tf=d_{ij}|R) is the probability that a relevant document has d_{ij} occurrences of the term T_j, and p(tf=d_{ij}|I) represents the probability that an irrelevant document has d_{ij} occurrences of the term T_j. "R" and "I" stand for "relevance" and "irrelevance", respectively. p(tf=0|R) and p(tf=0|I) are the probabilities with 0 occurrence of the term. For a query

$$Q = (T_{k1}, T_{k2}, \ldots, T_{kl})$$

and a document

$$D_i = (d_{i1}, d_{i2}, \cdots, d_{il})$$

the similarity between Q and D_i is

$$sim(Q,D_i) = \sum_{n=1}^{l} \delta_{kn} . w_2(T_{kn}, d_{kn})$$

where δ_{kn} is one when $d_{kn} \neq 0$, and zero when $d_{kn}=0$.

Again the probabilities p(tf=d_{ij}|R) and p(tf=d_{ij}|I) are usually unknown in advance. In addition to the estimation of p_j and q_j, the probabilities p(tf=d_{ij}|R) and p(tf=d_{ij}|I) are estimated from another available parameter in this paper.

3. Parameter Estimation of Two Independence Models

In order to estimate the probability values of the two independence models described in 2.2 and 2.3, we suppose that the probability values are correlated with document frequencies. Document frequency n_j is the number of documents which have term T_j. It seems to be a reasonable assumption that the probability values of a term T_j increases with its document frequency n_j.

Since document frequency n_j is a parameter which is easily obtainable, it becomes easy to predict the probability values in advance if a simple relationship between the probability values and document frequencies does exist.

An attempt to estimate parameters for the binary independence model has been made before[CrHa]. The estimation method given below is not only for the binary model but also for the non-binary model, and is more elaborate than the previous attempt.

3.1. Parameter Estimation of Binary Independence Model

Hypothesis I The probability values p_j and q_j in (2.1) are linearly correlated with the document frequency n_j of term T_j. In other words, there exist linear functions EP_1 and EQ_1 such that

$$p_j = EP_1(n_j), \qquad q_j = EQ_1(n_j).$$

Basing on the hypothesis I, we try to find the functions EP_1 and EQ_1 by the following steps.

Preconditions: We have a collection of documents D and a set of queries Q. For each query of Q, the set of relevant documents is known beforehand.

Step 1 : [Partition of queries] Divide a set of queries Q into two sets of queries Q_1 and Q_2 by some criterion. The query set Q_1 is used to estimate the probability values p_j and q_j from the document frequency n_j, while Q_2 is used to evaluate the performance of the estimated binary independence model. The following criterion is used to partition Q into Q_1 and Q_2 when it is applicable.

.criterion1:

Divide Q in such a way that each term of each query in Q_2 is contained in some query of Q_1.

If criterion 1 can not be satisfied by the queries in the collections, then use the following second criterion.

criterion2:

According to the empirical rule(the 2/3, 1/3 split)[Brei], Q_1 should contain 2/3 of all queries and Q_2 should contain the remaining 1/3 of queries.

Step 2 : [Data Plot] Count n_j(a document frequency),r_j(the number of relevant documents having term T_j), and s_j(the number of irrelevant documents having term T_j) for each term T_j of each query in Q_1. Plot the points (n_j,r_j/R) and(n_j,s_j/I) in each scattergram, where R is the total number of relevant documents and I is the total number of irrelevant documents. For a document frequency which has more than one corresponding r_j/R (or s_j/I), calculate their arithmetic mean and plot it. Having done this, we obtain two scattergrams, one for EP_1 and the other for EQ_1.

Step 3 : [Estimation of EP_1 and EQ_1] Suppose that EP_1 and EQ_1 are linear functions of document frequencies, that is to say,

$$EP_1(n_j) = a + bn_j ,$$
$$EQ_1(n_j) = c + dn_j .$$

Then calculate the regression coefficients a,b,c, and d by solving the following normal equations.

$$\sum_{j=1}^{N_0}\frac{r_j}{R} = aN_0+b\sum_{j=1}^{N_0}n_j \qquad \sum_{j=1}^{N_0}n_j\frac{r_j}{R} = a\sum_{j=1}^{N_0}n_j+b\sum_{j=1}^{N_0}n_j^2$$

for a and b, and

$$\sum_{j=1}^{N_0}\frac{s_j}{I} = cN_0+d\sum_{j=1}^{N_0}n_j \qquad \sum_{j=1}^{N_0}n_j\frac{s_j}{I} = c\sum_{j=1}^{N_0}n_j+d\sum_{j=1}^{N_0}n_j^2$$

for c and d, where N_0 is the number of data points in the scattergrams obtained in step 2. Standard error of estimate s_{ep}^2 and s_{eq}^2 are calculated by the formula,

$$s_{ep}^2 = \frac{1}{N_0-2}\sum_{j=1}^{N_0}[\frac{r_j}{R}-(a+bn_j)]^2$$

$$s_{eq}^2 = \frac{1}{N_0-2}\sum_{j=1}^{N_0}[\frac{r_j}{I}-(c+dn_j)]^2$$

Step 4 : [Modification of EP_1 and EQ_1] We obtain the estimations of EP_1 and EQ_1 in the previous step. These estimations may be crude and have to be refined by some rules, because under some situations enough data may not be available to make accurate estimations. The constraints defined below are from reasonable assumptions about probabilities and weighting functions.

(A0) EP_1 and EQ_1 should be non-negative increasing functions of document frequencies.

(A1) At the total number of documents N, EP_1 and EQ_1 must be one.

(A2) A weighting function derived from EP_1 and EQ_1 should be a decreasing function of document frequencies.

From the constraint A0, we get the first condition for the regression coefficients a,b,c, and d as follows.

$$a \geq 0, \quad b>0, \quad c \geq 0, \quad d>0 \tag{3.1}$$

From the constraint A1, the second condition satisfied by the regression coefficients is obtained.

$$EP_1(N) = a+bN = 1, \qquad EQ_1(N) = c+dN = 1. \tag{3.2}$$

The estimated weighting function of the binary independence model is

$$EW(n) = \log\frac{a+bn}{1-a-bn} - \log\frac{c+dn}{1-c-dn}$$

Calculating the derivative of $EW(n)$, we get the sufficient conditions for the regression coefficients to satisfy the constraint A2 as follows.

$$d>b, \qquad (b-d)(\frac{a(1-a)}{b}-\frac{c(1-c)}{d}) \geq (a-c)^2 \tag{3.3}$$

We have to pay attention to the fact that the condition (3.3) is not a necessary but sufficient condition, therefore we may disregard this condition as far as we make sure that $EW(n)$ is a decreasing function in the range of concern (0<n<N).

Step 5 : [Evaluation of the estimations] Measure the performance of the estimated binary independence model in terms of recall and precision by using the query set Q_2.

Recall and precision are defined by

$$Recall = \frac{\text{the number of relevant documents retrieved}}{\text{the total number of relevant documents}}$$

$$Precision = \frac{\text{the number of relevant documents retrieved}}{\text{the number of documents retrieved}}$$

A weight of a term T_j in a query Q_2 is calculated by the estimated weighting function obtained in the step 3 and step 4 as follows.

$$w_1(T_j) = EW(n_j) = \log\frac{EP_1(n_j)}{1-EP_1(n_j)} - \log\frac{EQ_1(n_j)}{1-EQ_1(n_j)}$$

where n_j is a document frequency of a term T_j. In order to evaluate performance of the estimated binary independence model, the results of retrieval are compared to the inverse document frequency method, the cosine measure, and the statistical similarity measure.

3.2. Parameter Estimation of Non-Binary Independence Model

Hypothesis II The probability values $p(tf=d_{ij}\,|R)$ and $p(tf=d_{ij}\,|I)$ in (2.2) are correlated with document frequency n_j and term frequency d_{ij}. In other words, there exist some linear functions EP_2 and EQ_2 such that

$$p(tf=d_{ij}\,|R) = EP_2(n_j,d_{ij}),$$

$$p(tf=d_{ij}\,|I) = EQ_2(n_j,d_{ij}).$$

Then we try to estimate EP_2 and EQ_2 by the following steps.

Preconditions and step 1 are the same as in 3.1.

Step 2 : [Plot of Data] Count n_j, r_{jt}(the number of relevant documents which have term T_j of term frequency t), and s_{jt}(the number of irrelevant documents which have term T_j of term frequency t) for each term T_j of each query in the query set Q_1. Plot the points $(n_j, r_{jt}/R)$, and $(n_j, s_{jt}/I)$. For a document frequency and a term frequency which have more than one corresponding r_{jt}/R (or s_{jt}/I), calculate their arithmetic mean and plot it. Having done this, we obtain a number of scattergrams, each of which shows the relation of n_j to r_{jt}/R (or s_{jt}/I) of the term frequency t.

Step 3 : [Estimations of EP_2 and EQ_2] Based on the scattergrams obtained in step 2, estimate EP_2 and EQ_2 by the linear regression.

$$EP_2(n_j,t) = a_t + b_t n_j$$

$$EQ_2(n_j,t) = c_t + d_t n_j$$

where t is a term frequency ranging from 1 to F, and F is the maximum number of a term frequency. Each set of regression coefficients a_t, b_t, c_t, and d_t are calculated by the same kind of normal equations in step 3 of 3.1.

Step 4 : [Modifications of EP_2 and EQ_2] EP_2's and EQ_2's obtained in step 3 may not be accurate because we do not have enough data to make accurate estimations under usual circumstances. Therefore we need the step to compensate the scarcity of data in order to make the estimations more accurate. As in the step 4 of 3.1, we set up some constraints which are supposed to be satisfied by EP_2's and EQ_2's. Using the constraints, we modify the regression coefficients a_t, b_t, c_t, and d_t.

(B0) EP_2's and EQ_2's are non-negative increasing functions of document frequencies.

(B1) Sum of EP_2's is equal to EP_1, and sum of EQ_2's is equal to EQ_1. That is to say,

$$EP_1(n) = \sum_{t=1}^{F} EP_2(n,t), \qquad EQ_1(n) = \sum_{t=1}^{F} EQ_2(n,t).$$

where F is the maximum number of term frequencies.

(B2) A term of low term frequency is usually more common than a term of high term frequency. In other words,

$$EP_2(n,t) < EP_2(n,t-1), \qquad EQ_2(n,t) < EQ_2(n,t-1), \qquad (t>1).$$

(B3) Weighting functions derived from EP_2's and EQ_2's should be decreasing functions of document frequencies.

From the rule B0, B1, and B2, the following conditions for the regression coefficients are derived,

$$a_t \ge 0, \quad b_t > 0, \quad c_t \ge 0, \quad d_t > 0 \tag{3.4}$$

$$a = \sum_{t=1}^{F} a_t, \quad b = \sum_{t=1}^{F} b_t, \quad c = \sum_{t=1}^{F} c_t, \quad d = \sum_{t=1}^{F} d_t. \tag{3.5}$$

$$a_{t-1} \ge a_t, \quad b_{t-1} > b_t, \quad c_{t-1} \ge c_t, \quad d_{t-1} > d_t. \tag{3.6}$$

From the rule B3, we obtain the following sufficient conditions for a monotonic decreasing property of weighting functions.

$$(1 + \frac{b-d-bc-ad}{bd}) b_t c_t < a_t b_t, \tag{3.7}$$

$$[(b-c-A)(a_t d_t + b_t c_t) - A_t (b+d-bc-ad)]^2 < 4(A_t bd + (b-d-A) B_t d_t)(A_t (1-a-c) + (b-d-A) a_t c_t)$$

where $A = bc-ad$ and $A_t = b_t c_t - a_t d_t$. Again, since (3.7) is not a necessary condition but a sufficient one, we may discard (3.7) as far as we make sure that weighting functions are decreasing functions in the range of our concern.

Step 5 : [Evaluation of the estimations] As the same as the step 5 of 3.1, measure the performance of the estimated non-binary independence model in terms of recall and precision. A weight of a term T_j occurring t times in a document is calculated by the following formula.

$$w_2(T_j,t) = EW_t(n_j) = \log\frac{EP_2(n_j,t)}{EQ_2(n_j,t)} - \log\frac{1-EP_1(n_j)}{1-EQ_1(n_j)}$$

The performance is compared to that of the inverse document frequency method, the cosine measure, and the statistical similarity measure.

4. Experiments

Three different document collections are used to evaluate the performance of the two estimated independence models. The document collections are called CISI, MEDLARS, and CRN4NUL, respectively.

4.1. Preliminary Statistics of Document Collections

The following table gives some statistics of three document collections used in the experiments.

Table 4.1 Preliminary statistics

Collection	CISI	MEDLARS	CRN4NUL
N	1460	1033	424
Q	76	30	155
M	7737	6928	2652
AMD	100	51	53
AMQ	28	10	8

In the above table, the meanings of the symbols are as follows.

N	total number of documents
Q	. total number of queries
M	total number of terms
AMD	average number of terms contained in a document
AMQ	average number of terms contained in a query

4.2. Results of Estimated Binary Independence Model

The criterion 1 of the query partition can be applied to CISI and CRN4NUL collections. For MEDLARS, the criterion 2 is applied. The results of the query partitions are shown in Table 4.2. The numbers of queries contained in each set of queries are shown in the table.

Table 4.2 Query partitions

Collection	CISI	MEDLARS	CRN4NUL
Q_1	61	20	110
Q_2	15	10	45

The regression coefficients a, b, c, and d for each collection are obtained by using the query set Q_1 as in the following table. The scattergrams and the linear regressions of CRN4NUL collection are shown in Fig. 4.1.

Table 4.3 Regression coefficients

Collection	CISI	MEDLARS	CRN4NUL
a	0.04209	0.05437	0.07145
b	0.00089	-0.00021	-0.00034
c	-0.00054	-0.00140	-0.00113
d	0.00068	0.001	0.00240

Since the coefficient c of every collection is negative and does not satisfy (3.1), c is modified to 0. Then, the second condition of (3.2) becomes $dN = 1$. The first condition of (3.2) is not satisfied by the coefficients a and b of every collection as follows.

$$EP_1(N) = a+bN = 0.04209+0.00089\times1460 = 1.3415 \quad (CISI),$$

$$EP_1(N) = a+bN = 0.05437-0.00021\times1033 = -0.1626 \quad (MEDLARS),$$

$$EP_1(N) = a+bN = 0.07145-0.00034\times424 = -0.07271 \quad (CRN4NUL).$$

Denoting the modified coefficients by a', b', c', and d', we obtain the constraints for these coefficients as follows.

$$a' = 1-b'N \qquad c' = 0 \qquad d' = \frac{1}{N} \tag{4.1}$$

where N is the total number of documents. If we use (4.1) to choose the regression coefficients, then it becomes impossible to satisfy (3.3) which is derived from the rule A2, because R.H.S. of the second inequality of (3.3) is

$$(a'-c')^2 = a'^2.$$

On the other hand, its L.H.S. becomes

$$-(b'-d')(\frac{c'(1-c')}{d'}-\frac{a'(1-a')}{b'}) = -a'^2.$$

Since this condition is sufficient, we can use (4.1) as far as we make sure that a weighting function derived from (4.1) is a decreasing function as document frequency increases in the range of our concern.

We still have ambiguity in deciding a' and b'. If we rely on a rather than b, then b' must be $(1-a)/N$, in order to satisfy the first condition of (3.2). Here we bring in a empirical rule, that is to say, a ratio of d to d' is equal to a ratio of $(1-a)/N$ to b'. In other words,

$$\frac{1-a}{N} : b' = d : d'.$$

This means that if we interpret $(1-a)/N$ as an experimental result for gradient of p_j, then ratio of gradients of p_j and q_j before and after the modification must be equal. Now using the rule, we decide a' and b' by the following formula.

$$b' = \frac{1-a}{N^2 d} \qquad a' = 1-b'N \tag{4.2}$$

Even though (4.2) does not give any optimal choice of a' and b', and other choices of a' and b' may give better performance than (4.2), (4.2) seems to be a reasonable choice.

The regression coefficients modified by (4.1) and (4.2) are as follows.

Table 4.4 Modified coefficients

Collection	CISI	MEDLARS	CRN4NUL
a'	0.03494	0.0796	0.0892
b'	0.000661	0.000891	0.002148
c'	0.0	0.0	0.0
d'	0.000685	0.000968	0.00236

The estimated weighting functions for three collections derived from Table 4.4 are shown in Fig.4.2. The performance of the estimated binary independence model using the regression coefficients of Table 4.4 is measured in terms of recall and precision. The results are shown in Table 4.5 with the performances of the other three methods. In Table 4.5, all numbers appeared are average precisions over the queries contained in Q_2. Average improvements of the estimated binary independence model compared to the other three methods are shown in Table 4.6. This table shows that the estimated binary independence model retrieves documents, for instance, 4.9 % better than the inverse document frequency method, and 26.5 % better than the cosine measure in CISI collection. Note that the results are reported for query set Q_2 only, where parameter estimation is not performed.

Table 4.5(a) CISI

average precisions over 15 queries

recall	eBIM	IDFM	COSINE	SSM
0.1	0.3670	0.3135	0.2330	0.2379
0.2	0.2776	0.2637	0.1665	0.1906
0.3	0.2075	0.1979	0.1541	0.1628
0.4	0.1639	0.1680	0.1323	0.1391
0.5	0.1477	0.1343	0.1172	0.1216
0.6	0.1265	0.1162	0.1042	0.1060
0.7	0.1093	0.1063	0.0924	0.0953
0.8	0.0937	0.0945	0.0834	0.0811
0.9	0.0762	0.0739	0.0716	0.0672
1.0	0.0471	0.0473	0.0484	0.0478

4.3. Results of Estimated Non-Binary Independence Model

The regression coefficients a_t, b_t, c_t, and d_t estimated for each collection by using Q_1 are shown in Table 4.7. In this table, "tf" means a term frequency and "pts" means the number of data points used to estimate the regression coefficients for each term frequency. Scattergrams and linear regressions of CRN4NUL collection are shown in Fig. 4.3.

Table 4.5(b) MEDLARS

average precisions over 10 queries

recall	eBIM	IDFM	COSINE	SSM
0.1	0.8261	0.8241	0.8595	0.8700
0.2	0.7898	0.8064	0.7133	0.7678
0.3	0.6948	0.6747	0.5424	0.6454
0.4	0.6521	0.6476	0.5039	0.5237
0.5	0.5340	0.5560	0.4210	0.4554
0.6	0.4362	0.4329	0.3718	0.4226
0.7	0.3462	0.3157	0.3204	0.3413
0.8	0.2796	0.2755	0.2676	0.2756
0.9	0.1811	0.1781	0.1858	0.1911
1.0	0.0651	0.0842	0.1189	0.1048

Table 4.5(c) CRN4NUL

average precisions over 45 queries

recall	eBIM	IDFM	COSINE	SSM
0.1	0.6174	0.6464	0.5865	0.6449
0.2	0.5054	0.4998	0.4352	0.4903
0.3	0.4034	0.3865	0.3675	0.4014
0.4	0.3548	0.3290	0.2867	0.2978
0.5	0.3007	0.2863	0.2341	0.2432
0.6	0.2520	0.2515	0.2026	0.2200
0.7	0.2026	0.2015	0.1658	0.1826
0.8	0.1607	0.1577	0.1441	0.1510
0.9	0.1377	0.1350	0.1204	0.1289
1.0	0.1261	0.1246	0.1131	0.1199

Table 4.6 Improvements by eBIM (%)

Collection	CISI	MEDLARS	CRN4NUL
IDFM	4.9	-1.1	2.0
COSINE	26.5	7.4	16.7
SSM	22.8	1.1	8.6

Table 4.7(a) Regression coefficients of CISI

estimated by using 61 queries

tf	a_t	b_t	c_t	d_t	pts
1	0.04851	0.00034	0.01382	0.00036	191
2	-0.00022	0.00024	-0.00366	0.00014	188
3	0.00535	0.00010	-0.00268	0.00007	181
4	0.00342	0.00005	-0.00217	0.00004	157
5	0.01432	-0.00001	-0.00141	0.00003	126
6	0.00515	0.00001	-0.00119	0.00002	84
7	0.00413	0.00001	-0.00059	0.00001	70
8	0.01004	-0.00001	-0.00022	0.00001	45
9	0.00102	0.00001	0.00007	0.00001	35
10	0.02001	-0.00005	0.00027	0.0	28
11	0.00078	0.00001	0.00059	0.0	13
12	0.03193	-0.00008	0.00004	0.0	7
13	0.01193	0.00003	-0.00019	0.0	9

Using the rules B0, B1, B2, and B3, we modify the regression coefficients of Table 4.7 and make a modified set of regression coefficients as shown in Table 4.8. First of all, the regression coefficients estimated by less than 20% of maximum pts are discarded, because we can not expect enough reliability in such data. In order to decide the coefficients uniquely, the following set of formula is used in addition to (3.4)-(3.7).

$$c''_t = 0 \quad d''_t = \frac{Nd_t + c_t}{N} \quad b''_t = \frac{Nd''_t - a_t}{N} \quad a''_t = a_t \tag{4.3}$$

The first condition of (4.3) is derived from (3.4), (3.5), and (4.1). (3.4) and (3.5) are

$$c'_t \geq 0 \quad , \quad c' = \sum_{t=1}^{F} c'_t.$$

On the other hand, c' is 0 by (4.1). Therefore all c'_t's must be 0. The second condition of (4.3) is based on the fact that probability values $p(tf = d_{ij} | I)$ at N seem to be reliable because

$$\sum_{t=1}^{8}(c_t + Nd_t) = 0.9801 \approx 1 \quad (CISI),$$

$$\sum_{t=1}^{9}(c_t + Nd_t) = 1.0003 \approx 1 \quad (MEDLARS),$$

$$\sum_{t=1}^{8}(c_t + Nd_t) = 0.9894 \approx 1 \quad (CRN4NUL).$$

Table 4.7(c) Regression coefficients of CRN4NUL estimated by using 110 queries

tf	a_t	b_t	c_t	d_t	pts
1	0.08776	0.00073	0.01131	0.00123	101
2	0.08743	0.00030	-0.00301	0.00054	101
3	0.07063	-0.00013	-0.00267	0.00026	98
4	0.05260	-0.00011	-0.00070	0.00013	87
5	0.06180	-0.00031	0.00040	0.00007	77
6	0.03777	-0.00019	0.00107	0.00004	62
7	0.03653	-0.00012	0.00117	0.00003	46
8	0.04788	-0.00021	0.00239	0.00001	36
9	0.03432	-0.00010	0.00120	0.00002	17
10	0.02545	-0.00007	0.00200	0.00001	17
11	0.10018	-0.00076	0.00146	0.00001	10
12	0.02234	-0.00016	0.00202	0.0	5
13	0.07231	-0.00038	0.00121	0.00001	5
14	0.03541	-0.00052	0.00154	0.00001	3

Table 4.7(b) Regression coefficients of MEDLARS estimated by using 20 queries

tf	a_t	b_t	c_t	d_t	pts
1	0.07478	0.00027	-0.0007	0.00059	93
2	0.05337	0.00020	-0.00194	0.00020	92
3	0.04951	-0.00006	-0.00051	0.00008	90
4	0.04028	-0.00007	0.00002	0.00004	77
5	0.04313	-0.00014	0.00030	0.00002	60
6	0.03243	-0.00008	0.00055	0.00001	60
7	0.04757	-0.00017	-0.00045	0.00002	40
8	0.03423	-0.00012	0.00030	0.00001	24
9	0.02769	-0.00008	0.00072	0.0	21
10	0.00941	0.00001	0.00105	0.0	13
11	0.00508	0.00001	0.00096	0.0	15
12	0.03935	-0.00021	0.00015	0.00001	13
13	-0.06018	0.00074	0.00331	-0.00003	3

Taking c_t+Nd_t as the actual value of $p(tf=d_{ij}\,|I)$ at N, we obtain the following equation.

$$c''_t+Nd''_t = Nd''_t = c_t+Nd_t.$$

This leads to the second condition of (4.3). The third condition of (4.3) is based on the assumption that the probability values of $p(tf=d_{ij}\,|R)$ and $p(tf=d_{ij}\,|I)$ are the same at N. This means that a proportion of a probability value of each term frequency to the total(total is 1) at N is the same in $p(tf=d_{ij}\,|I)$ and $p(tf=d_{ij}\,|R)$. Again we rely on a_t's rather than b_t's, then b''_t must satisfy

$$a_t + Nb''_t = c''_t + Nd''_t = Nd''_t.$$

If one of the following things happened, then a''_t or b''_t or d''_t are set half of a''_{t-1} or b''_{t-1} or d''_{t-1}, respectively. This rule is obtained by observing the experimental results of d_t, that is $d_t \approx 0.5d_{t-1}$.

- $a''_t<0$ or $b''_t<0$ or $d''_t<0$
- $a''_t>a''_{t-1}$ or $b''_t>b''_{t-1}$ or $d''_t>d''_{t-1}$
- $a''_t<0.1a''_{t-1}$ or $b''_t<0.1b''_{t-1}$ or $d''_t<0.1d''_{t-1}$

In order to satisfy (3.5), the final coefficients used to calculate weights are given by

$$a'_t = \frac{a'}{\sum_{k=1}^{F} a''_k}a''_t \qquad b'_t = \frac{b'}{\sum_{k=1}^{F} b''_k}b''_t \qquad c'_t = 0.0 \qquad d'_t = \frac{d'}{\sum_{k=1}^{F} d''_k}d''_t \tag{4.4}$$

So far, since we do not consider B3, we have to make sure that weighting functions derived from (4.3) and (4.4) are decreasing functions of document frequency. Again, although (4.3) and (4.4) do not guarantee any optimality of choices, it usually gives reasonable-performances. The weighting functions derived from Table 4.8(c) are shown in Fig. 4.4.

Table 4.8(a) Modified coefficients of CISI

tf	a'_t	b'_t	c'_t	d'_t
1	0.020001	0.000359	0.0	0.000374
2	0.010001	0.000147	0.0	0.000139
3	0.002206	0.000069	0.0	0.000069
4	0.001410	0.000039	0.0	0.000039
5	0.000705	0.000021	0.0	0.000029
6	0.000353	0.000017	0.0	0.000019
7	0.000176	0.000007	0.0	0.000010
8	0.000088	0.000004	0.0	0.000005

Table 4.8(h) Modified coefficients of MEDLARS

tf	a'_t	b'_t	c'_t	d'_t
1	0.026152	0.000634	0.0	0.00060
2	0.018665	0.000180	0.0	0.00020
3	0.017315	0.000039	0.0	0.00008
4	0.008754	0.000019	0.0	0.00004
5	0.004478	0.000010	0.0	0.00002
6	0.002379	0.000005	0.0	0.000011
7	0.001189	0.000002	0.0	0.000005
8	0.000595	0.0000012	0.0	0.00003
9	0.000074	0.0000006	0.0	0.00001

Table 4.8(c) Modified coefficients of CRN4NUL

tf	a'_t	b'_t	c'_t	d'_t
1	0.023502	0.001457	0.0	0.00127
2	0.023414	0.000453	0.0	0.00054
3	0.018915	0.000121	0.0	0.00026
4	0.009627	0.000060	0.0	0.00013
5	0.005582	0.000030	0.0	0.00007
6	0.003592	0.000015	0.0	0.00004
7	0.003102	0.000008	0.0	0.00003
8	0.001466	0.000004	0.0	0.00002

Table 4.9(a) CISI

average precisions over 15 queries

recall	eNBIM	IDFM	COSINE	SSM
0.1	0.3784	0.3135	0.2330	0.2379
0.2	0.2923	0.2637	0.1665	0.1906
0.3	0.2046	0.1979	0.1541	0.1628
0.4	0.1771	0.1680	0.1323	0.1391
0.5	0.1502	0.1343	0.1172	0.1216
0.6	0.1324	0.1162	0.1042	0.1060
0.7	0.1111	0.1063	0.0924	0.0953
0.8	0.0992	0.0945	0.0834	0.0811
0.9	0.0785	0.0739	0.0716	0.0672
1.0	0.0481	0.0473	0.0484	0.0478

Table 4.9(b) MEDLARS

average precisions over 10 queries

recall	eNBIM	IDFM	COSINE	SSM
0.1	0.8397	0.8241	0.8595	0.8700
0.2	0.8175	0.8064	0.7133	0.7678
0.3	0.7685	0.6747	0.5424	0.6454
0.4	0.6443	0.6476	0.5039	0.5237
0.5	0.5557	0.5560	0.4210	0.4554
0.6	0.4473	0.4329	0.3718	0.4226
0.7	0.3940	0.3157	0.3204	0.3413
0.8	0.2887	0.2755	0.2676	0.2756
0.9	0.1494	0.1781	0.1858	0.1911
1.0	0.0764	0.0842	0.1189	0.1048

Table 4.9(c) CRN4NUL

average precisions over 45 queries

recall	eNBIM	IDFM	COSINE	SSM
0.1	0.6449	0.6464	0.5865	0.6449
0.2	0.5444	0.4998	0.4352	0.4903
0.3	0.4501	0.3865	0.3675	0.4014
0.4	0.3935	0.3290	0.2867	0.2978
0.5	0.3104	0.2863	0.2341	0.2432
0.6	0.2586	0.2515	0.2026	0.2200
0.7	0.2052	0.2015	0.1658	0.1826
0.8	0.1561	0.1577	0.1441	0.1510
0.9	0.1262	0.1350	0.1204	0.1289
1.0	0.1155	0.1246	0.1131	0.1199

Table 4.10 Improvements by eNBIM (%)

Collection	CISI	MEDLARS	CRN4NUL
IDFM	8.3	2.4	4.2
COSINE	30.8	11.0	19.4
SSM	27.0	4.4	11.0

The performance of the estimated non-binary independence model using the regression coefficients of Table 4.8 is shown in Table 4.9 with the performances of the other three methods. The numbers appearing in Table 4.9 are the average precisions over the queries contained in Q_2. Improvements of the estimated non-binary independence model compared to the other three method are shown in Table 4.10. This table shows that the estimated non-binary independence model retrieves documents, for instance, 8.3% better than the inverse document frequency method, and 30.8% better than the cosine measure in CISI collection.

5. Motivations of the Second Method

The first method may be unable to make the most of relevance information. For instance, the method can not distinguish terms of the same document frequency and/or term frequency. Another disadvantage is that we have to bring in some empirical rules to determine the coefficients uniquely. In order to compensate these disadvantages, a formula is given to learn optimal weights of the non-binary independence model. The formula is applied to the same three document collections. Effects of a coefficient used in the formula and the number of learning times on retrieval performances are measured in terms of average improvements compared to the inverse document frequency method.

6. Learning Scheme

6.1. Learning Formula

A learning formula used in the experiments is as follows.

$$w_{k+1}(T_j,tf) = w_k(T_j,tf) + \frac{c}{n_j}[w_{opt}(T_j,tf) - w_k(T_j,tf)] \qquad (6.1)$$

where $w_k(T_j,tf)$ is a weight of j^{th} term T_j with term frequency tf after learning k times, $w_{opt}(T_j,tf)$ is optimal weight of j^{th} term T_j with term frequency tf by the non-binary independence model, c is a learning coefficient, and n_j is document frequency of term T_j. Although $w_{opt}(T_j,tf)$ is determined not only by T_j and tf but also by a query in a learning query set, we do not express it explicitly here for simplicity of expression. $w_k(T_j,tf)$ and $w_{opt}(T_j,tf)$ will be abbreviated as w_k and w_{opt} in the sequel.

The reason why document frequency n_j is incorporated in this formula is that weights of term T_j are learned $m \times n_j$ times at every iteration of learning, where m is the number of occurrences of term T_j in the learning query set. This means that weights of high document frequency terms are learned many more times than weights of low document frequency term. Document frequency n_j in the formula (6.1) prevents this inequality in learning.

6.2. Properties of Learning Formula

It can be easily derived from (6.1) that the formula expressing w_k by w_0 (initial weight) is as follows.

$$w_k = [1 - (1-\frac{c}{n_j})^k](w_{opt} - w_0) + w_0. \qquad (6.2)$$

From (6.2), we recognize that if $\frac{c}{n_j}$ is close to one, then $(1 - \frac{c}{n_j})^k$ rapidly approaches to zero as k increases, and this means w_k rapidly approaches to w_{opt} with times of learning.

If $\frac{c}{n_j}$ is very small and k is also small, then the following approximation holds.

$$(1 - \frac{c}{n_j})^k \approx 1 - \frac{kc}{n_j}.$$

And (6.2) becomes

$$w_k \approx w_0 + \frac{kc}{n_j}(w_{opt} - w_0).$$

This means that k steps of learning using c is almost equivalent to one step of learning using kc, provided $\frac{c}{n_j}$ is very small and k is also small.

This relation of learning coefficients and the number of learning steps will be observed in the results of the experiments in section 7.

If w_{opt} is unique, then w_k never overshoots or undershoots w_{opt} as far as $\frac{c}{n_j}$ is less than one. However, as pointed out in 6.1, the values of w_{opt} may be different for different query, and this may cause w_k to overshoot or undershoot some values of w_{opt}.

6.3 Learning Procedure

The following is the learning procedure used in the experiments.

1° Assign a weight by the inverse document frequency method to each query term contained in a learning query set as an initial weight.

2° Using (6.1) and the learning query set, modify weights in documents.

3° Using another query set, evaluate performance of the weights modified in 2°. Compare the performance of the weights to the performance of the inverse document frequency method.

4° Go to 2° until the learning is executed sufficient times.

The following example illustrates the learning procedure described above.

Example Suppose that term T_1 has the following distribution of term frequencies in total ten documents.

Term-Doc	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
T_1	0	1	2	0	0	1	0	1	2	0

Then initial weights of documents are

Term-Doc	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
T_1	0	log2	log2	0	0	log2	0	log2	log2	0

Suppose that D_{1-6} are non-relevant andD_{7-10} are relevant to the example query. Then

$$w_{opt}(T_1,1) = \log\frac{3}{4} \approx -0.29,$$

$$w_{opt}(T_1,2) = \log\frac{3}{2} \approx 0.4.$$

$w_k(T_1,1)$ is learned three times, and $w_k(T_1,2)$ is learned two times during one iteration of learning(i.e. going through all documents having the term). Therefore during first time of learning with coefficient 0.2, the weights are changed after one iteration of learning as follows.

$$w_1(T_1,1) = log\,2 + \frac{0.2}{5}(log\frac{3}{4} - log\,2) = 0.65,$$

$$w_2(T_1,1) = 0.65 + \frac{0.2}{5}(log\frac{3}{4} - 0.65) = 0.61,$$

$$w_3(T_1,1) = 0.61 + \frac{0.2}{5}(log\frac{3}{4} - 0.61) = 0.57.$$

And

$$w_1(T_1,2) = log\,2 + \frac{0.2}{5}(log\frac{3}{2} - log\,2) = 0.68,$$

$$w_2(T_1,2) = 0.68 + \frac{0.2}{5}(log\frac{3}{2} - 0.68) = 0.67.$$

The weights in the documents are changed as follows.

Term-Doc	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
T_1	0	0.57	0.67	0	0	0.57	0	0.57	0.67	0

□

7. Experimental Results of the Second Method

The experiments are performed on three different document collections, CISI, MEDLARS, and CRN4NUL. The query sets for learning and evaluation are the same as those in the experiments of section 4, that is to say, Q_1 and Q_2, respectively.

Five different learning coefficients ranging from 0.002 to 0.1 are experimented. Learning curves of each learning coefficient *are shown in Fig.7.1. The improvement in the figure is measured compared to the* inverse document frequency method. The precision-recall graph of the learning formula after 10 times of learning with coefficient 0.016 in CISI collection is shown in Fig.7.2 with the precision-recall of the inverse document frequency method.

Table 7.1 shows recall-precision of the formula after each time of learning in CISI. From this table, we can observe that one step of learning with coefficient 0.016 is almost equal to two steps of learning with coefficient 0.008 as pointed out in section 6.

Table 7.2 shows average improvements of the scheme using coefficient 0.016 after 10 times of learning. The improvements are measured compared to the inverse document frequency method.

Table 7.2 Improvements by the scheme (%)

CISI	MEDLARS	CRN4NUL
7.8	1.0	7.3

Table 7.1(a) Precision-Recall of CISI(c =0.008)

precision improvement compared to IDFM (%)

recall-step	1	2	3	4	5	6	7	8	9	10
0.1	1.3	1.9	3.1	3.4	3.7	4.3	3.0	1.6	1.6	3.1
0.2	3.8	5.7	7.2	6.9	7.2	7.8	8.5	9.6	9.9	10.3
0.3	3.3	2.6	2.2	2.1	2.0	1.7	2.1	2.0	1.5	2.3
0.4	0.3	-0.5	-0.6	-0.1	1.2	1.7	2.1	2.2	2.2	2.9
0.5	3.8	4.9	5.2	6.2	6.8	6.8	8.1	8.8	9.5	9.7
0.6	3.0	3.8	5.4	6.3	6.7	7.6	7.6	7.8	7.6	7.9
0.7	2.8	4.2	4.6	5.4	5.8	5.4	5.3	5.8	5.4	5.7
0.8	4.1	3.6	3.2	2.9	1.6	1.7	1.0	0.8	0.9	1.3
0.9	0.2	1.2	2.5	2.8	3.5	3.8	4.7	5.2	6.3	6.1
1.0	1.5	1.6	1.6	1.7	2.3	2.2	2.2	2.1	2.1	2.1
average	2.4	2.9	3.5	3.8	4.1	4.3	4.5	4.6	4.7	5.1

Table 7.1(b) Precision-Recall of CISI(c =0.016)

precision improvement compared to IDFM (%)

recall-step	1	2	3	4	5	6	7	8	9	10
0.1	1.9	3.7	4.3	1.4	3.1	4.7	5.4	6.5	8.9	10.1
0.2	5.7	7.2	7.4	9.3	10.5	9.3	11.7	11.1	11.6	12.4
0.3	2.4	2.1	1.7	2.0	2.3	2.8	4.2	5.3	5.9	7.0
0.4	-0.6	0.0	1.8	2.3	2.9	3.4	3.0	3.2	3.8	4.4
0.5	4.8	6.1	6.8	8.9	9.7	10.3	10.4	11.4	12.2	13.7
0.6	3.8	6.3	7.4	7.5	7.9	7.8	9.8	11.3	10.7	11.7
0.7	4.2	5.3	5.4	5.7	5.7	7.1	7.1	7.2	6.9	7.0
0.8	3.9	2.8	1.2	0.8	1.3	1.6	2.6	3.0	2.7	2.9
0.9	1.2	2.8	3.8	5.2	6.0	5.7	6.2	6.2	6.5	6.3
1.0	1.6	1.7	2.2	2.1	2.1	2.3	2.3	2.3	2.5	2.5
average	2.9	3.8	4.2	4.5	5.2	5.5	6.3	6.7	7.2	7.8

Table 7.3 summarizes the best results of improvements above the inverse document frequency method, for each collection we obtained in the experiments. We do not have a systematic method to determine the coefficients and the learning times shown in the table.

Table 7.3 Best results

Collection	CISI	MEDLARS	CRN4NUL
best result(%)	15.35	2.37	7.49
coefficient	0.2	0.2	0.1
learning times	4	3	1

From Fig.7.1-2, we observe the following results.

(1) Improvement in performance is observed in every collection when the learning coefficient is small(0.002-0.016).

(2) The improvement of MEDLARS is small compared to those of CISI and CRN4NUL.

(3) The performance is rapidly deteriorated in CRN4NUL after the first learning with the coefficient 0.1.

One reason of (2) is that the size of learning data of MEDLARS is small compared to the other collections. In MEDLARS, 20 queries are used for learning in contrast to 61 queries of CISI and 110 queries of CRN4NUL. Another reason is that a term in Q_2 does not always appear in Q_1. As described in 4.2., criterion1 can not be satisfied by MEDLARS in the partition of the query set, and only 22 query terms out of 88 total query terms contained in Q_2 appear in Q_1.

The reason of (1) and (3) is that a weight of a query term of every term frequency is approaching to its optimal weight after each time of learning as far as a learning coefficient is small. On the other hand, when a learning coefficient is large, a weight of some query terms may overshoot its optimal weight and is overestimated in a query as shown in the following example.

Example Suppose that term T_j has document frequency one, and three queries use term T_j in a learning query set. Optimal weights for the queries and the initial weight are as follows(Fig.7.3(a)).

$$w_{opt}^{(1)} = 2.0$$

$$w_{opt}^{(2)} = 4.0$$

$$w_{opt}^{(3)} = 3.5$$

$$w_0 = 1.0$$

If the learning coefficient c is set 0.2, then in the first time of learning, w_k is modified as follows.

$$w_1 = 1.0+0.2(2.0-1.0) = 1.2$$

$$w_2 = 1.2+0.2(4.0-1.2) = 1.76$$

$$w_3 = 1.76+0.2(3.5-1.76) = 2.11$$

w_3 already exceeds $w_{opt}^{(1)}$. After the second time of learning w_6 becomes 2.68 and w_9 is 2.96 after the third time of learning(Fig.7.3(b)).□

The above example shows that a learning with large coefficient or many times of learning may cause overshootings of optimal weights for some queries. This means that an unimportant term in some queries may be overestimated and deteriorate performance. If we call the coefficient a critical coefficient when using it shows good improvement after first few learnings, but deteriorates performance after that, then the critical coefficient for CRN4NUL is around 0.1. The critical coefficients for CISI and MEDLARS are not obvious like CRN4NUL, and they may be about 2.5 for CISI and 0.6 for MEDLARS as shown in Fig.7.4. Because of the equivalence of the learning coefficient and times of learning as shown in 6.2., the concept of the critical coefficient corresponds to (coefficient) × (learning times). This means that even though a small coefficient is used in learning, it may deteriorate performance after many times of learning.

8. Conclusions

The parameter values used in the binary independence model and the non-binary independence model are estimated by using linear regressions and some constraints and rules in the first method. The optimal weighting functions are derived from the estimations and their performances are measured and compared to the performance of the inverse document frequency method, the cosine measure, and the statistical similarity measure. In the experiments using three different collections of documents, the estimated binary independence model shows on the average about 1.9%, 16.9%, and 10.8% better performance than the inverse document frequency method, the cosine measure, and the statistical similarity measure, respectively. The estimated non-binary independence model shows on the average about 5%, 20.3%, and 14.1% better performance than the inverse document frequency method, the cosine measure, and the statistical similarity measure, respectively.

Even though the two estimated models presented here show better performances than the three other methods, the discrepancy of the performances between the estimated models and the theoretical models is immense (Table 8.1), and there seems to be much room to improve the estimation method. One reason of this discrepancy is lack of sufficient data to make estimations accurate. Another reason is that there is much ambiguity in deciding the regression coefficients uniquely. We use some empirical rules in addition to the constraints on probability values and weighting functions to decide regression coefficients uniquely. We do not think that we choose the best possible set of rules and constraints. There might be more reasonable set of rules and constraints. For instance, the rule that weights of high term frequency should be larger than those of low term frequency may be a more reasonable constraint than the constraint B2 of section 3.2., or we should take into account a suitable model of term frequency distributions[BoSw]. Another possible reason is that we discard some data of high term frequencies in the experiments of the estimated non-binary model. Aggregating the data might be a more reasonable way to estimate the coefficients than discarding them in step 4 of 3.2.

Table 8.1 CISI

performances of theoretical and estimated models(average over 15 queries)

recall	eBIM	BIM	eNBIM	NBIM
0.1	0.3670	0.4845	0.3784	0.7718
0.2	0.2776	0.4208	0.2923	0.6798
0.3	0.2075	0.3228	0.2046	0.5394
0.4	0.1639	0.2809	0.1771	0.4684
0.5	0.1477	0.2472	0.1502	0.3905
0.6	0.1265	0.2234	0.1324	0.3417
0.7	0.1093	0.1646	0.1111	0.2856
0.8	0.0937	0.1207	0.0992	0.2057
0.9	0.0762	0.1007	0.0785	0.1664
1.0	0.0471	0.0515	0.0481	0.1034

A scheme for learning optimal weights of the non-binary model is presented. The experimental results show that this scheme improves performance on the average 5.4% better than the inverse document frequency method. This result is slightly better than the estimated non-binary model. Learning with a small coefficient seems to be almost constantly improving performance with times of learnings. On the other hand, a learning with a large coefficient is deteriorating performance with times of learnings.

Validity of the schemes presented here is based on the assumption that there exists some similarity between significances of terms in a learning query set and that of the queries in a query set for evaluation. In other words, we expect that a valuable query term in a learning query set is also valuable in the other set, and non-valuable term in a learning query set is also non-valuable in the other set. This seems to be true to some extent. However, the degree of significance of the term in a query may not be the same as that in another query. That is why we obtain some improvement in the experiments, but it is not sufficiently significant. Another difficulty is that mechanical ways to determine the parameters used in the schemes are hard to find out, as in the determination of the regression coefficients in the first method and the appropriate learning coefficient in the second method.

Future directions of research are as follows:

(1) Seek a different formula which does not show deterioration as the number of learnings increases.

(2) Seek a way to differentiate the usage of a term by one user from that of a different user. A methodology to differentiate the usage of a term in one context from the term in another context is sketched in [Yu].

(3) Eliminate dependencies of terms. An elimination of term dependencies is attempted in [WoZW], experimented in [WoYa], and achieves some improvement.

(4) Incorporate normalization by document size into the schemes reported here. We compared our results mainly with the basic methods-- the inverse document frequency method and the cosine measure. Other methods might give better results than the basic methods. For instance, the method based on the inverse document frequency and improved by normalization[Crof], or the method based on the cosine measure and improved by taking into account term frequencies[Salt2] gives better performance than the basic methods. The experiments shown in this paper are not intended to show that these learning schemes achieve the best performance among the known retrieval methods. Rather it shows a more fundamental fact, that is, learning does help to improve retrieval performance. We expect better performance if normalization by document size is incorporated into the learning methods.

References

[BoSw] Bookstein, A. and Swanson, D. "A Decision Theoretic Foundation for Indexing", J. of the American Society for Information Science, January 1975, pp45-50.

[Brei] Breiman, L. et al. "Classification and regression trees", WADSWORTH INTERNATIONAL GROUP, 1984, p11.

[CrHa] Croft, W.B. and Harper, D.J. "Using Probabilistic Models of Document Retrieval without Relevance Information", J. of Documentation, Vol. 35, 1979 pp285-295.

[Crof] Croft, W.B. "Experiments with Representation in a Document Retrieval System", Information Technology:Research and Development, Vol. 2, January 1983, pp1-21.

[RoSp] Robertson, S.E. and Sparck Jones, K. "Relevance Weighing of Search Term", J. of the American Society for Information Science, May-June 1976, pp129-146.

[Salt] Salton, G. (ed.) "The SMART Retrieval System --- Experiments in Automatic Document Processing", Englewood Cliffs, NJ:Prentice-Hall, 1971.

[Salt2] Salton, G. "Recent Trend in Automatic Information Retrieval", ACM SIGIR conference, 1986, pp1-10.

[SaMc] Salton, G. and McGill, M. J. "Introduction to Modern Information Retrieval", New York:McGraw-Hill, 1983 pp121-131.

[Spar] Sparck Jones, K. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", J. of Documentation, Vol. 28, No. 1, March 1972, pp11-20.

[WoYa] Wong, S.K.M. and Yao, Y.Y. "A Statistical Similarity Measure", ACM SIGIR 1987, pp3-12.

[WoZW] Wong, S.K.M., Ziarko, W., and Wong, P.C.N. "Generalized Vector Space Model in Information Retrieval", ACM SIGIR conference, 1985,pp18-25.

[Yu] Yu, C.T. "A Framework for Effective Retrieval", Technical Report, the University of Illinois at Chicago, 1987.

[YuLe] Yu, C.T. and Lee, T.C. "Non-Binary Independence Model", ACM SIGIR conference, 1986, pp265-268.

[YuSa] Yu, C.T. and Salton, G. "Precision Weighting - an effective automatic indexing method", J.ACM, 1976, pp76-88.

Fig. 4.2

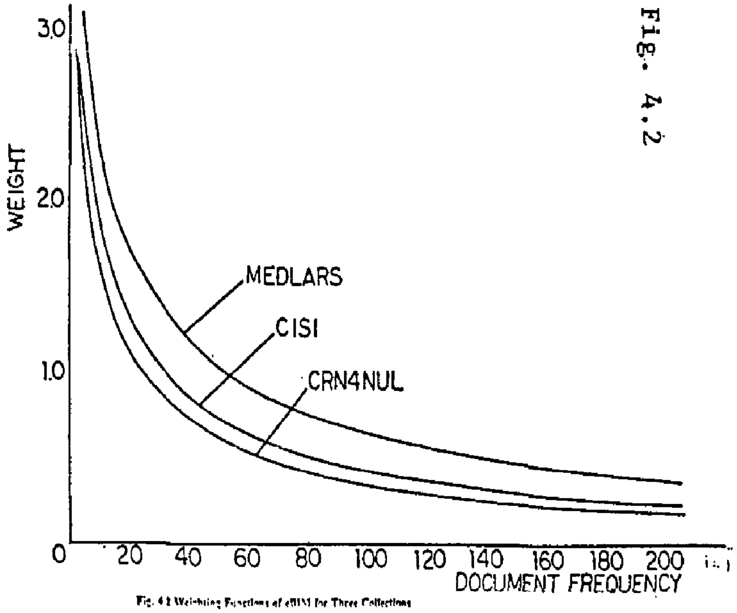

Fig. 4.2 Weighting Functions of cBIM for Three Collections

Fig. 4.1(a)

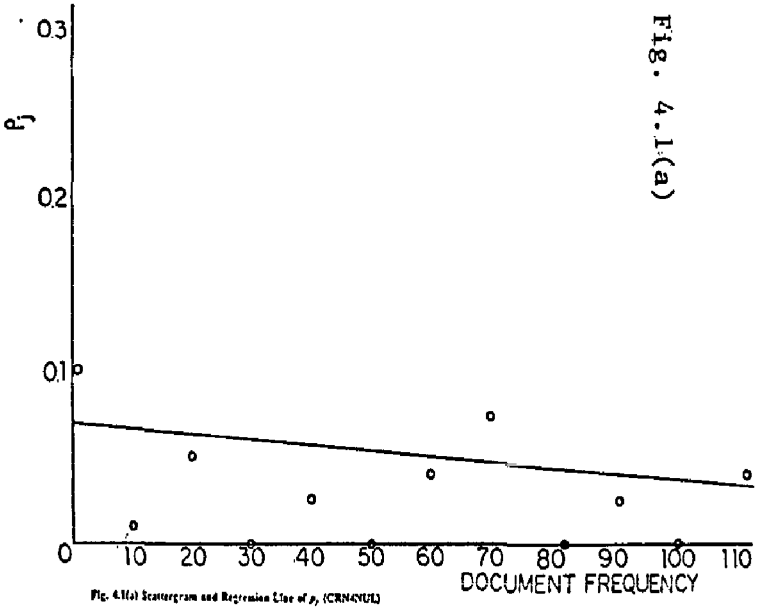

Fig. 4.1(a) Scattergram and Regression Line of p_j (CRN4NUL)

Fig. 4.3(a)

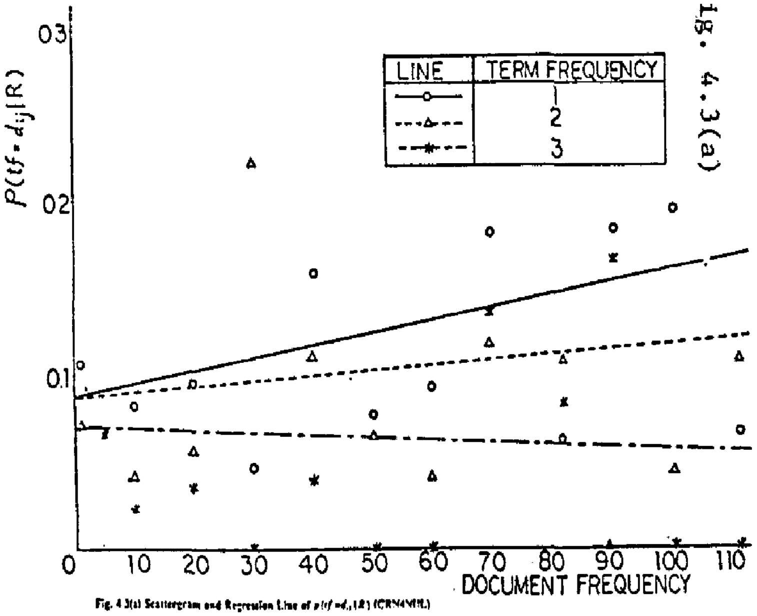

Fig. 4.3(a) Scattergram and Regression Line of $p(tf=d_{ij}|R)$ (CRN4NUL)

Fig.4(b)

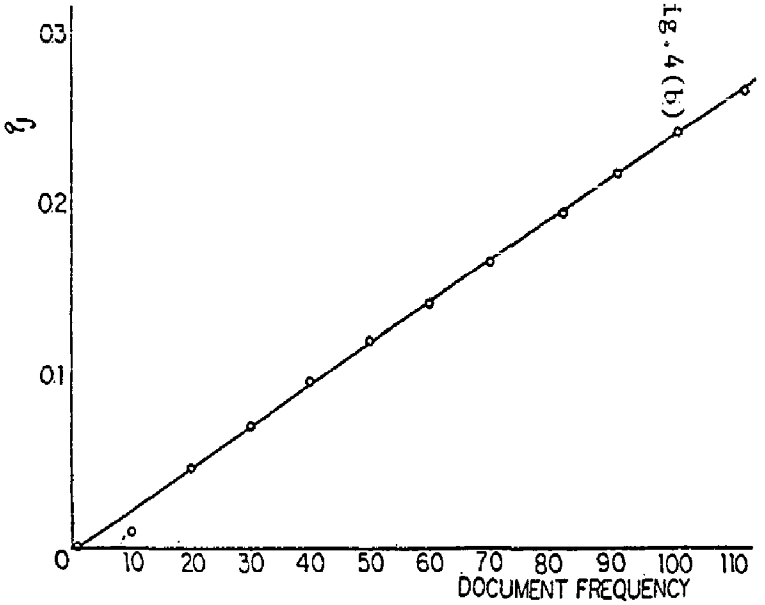

Fig. 4.3(b)

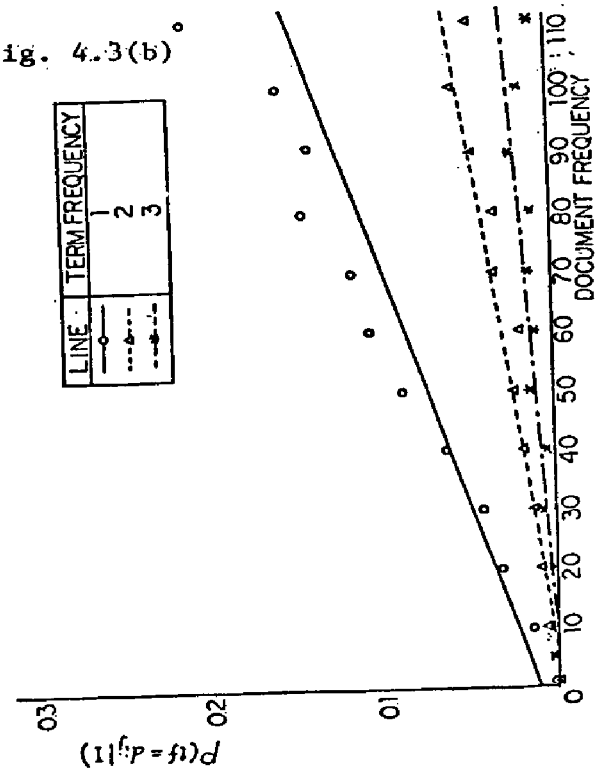

Fig. 4.3(b) Scattergram and Regression Line of $p(t|d_j,1)$ (CRANFIELD)

Fig. 4.4

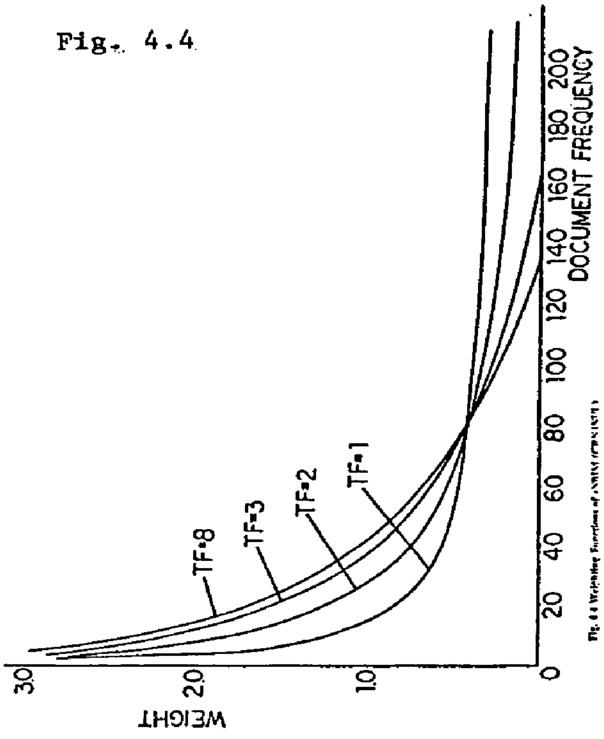

Fig. 4.4 Weighting Functions of [illegible] (CRANFIELD)

Fig.7.1(a)

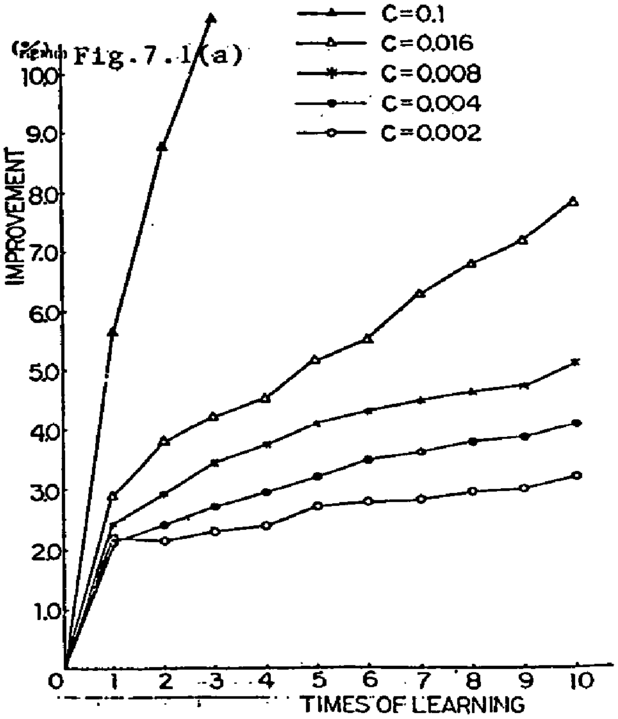

Fig. 7.1(a) Learning Curves (CISI)

Fig.7.1(b)

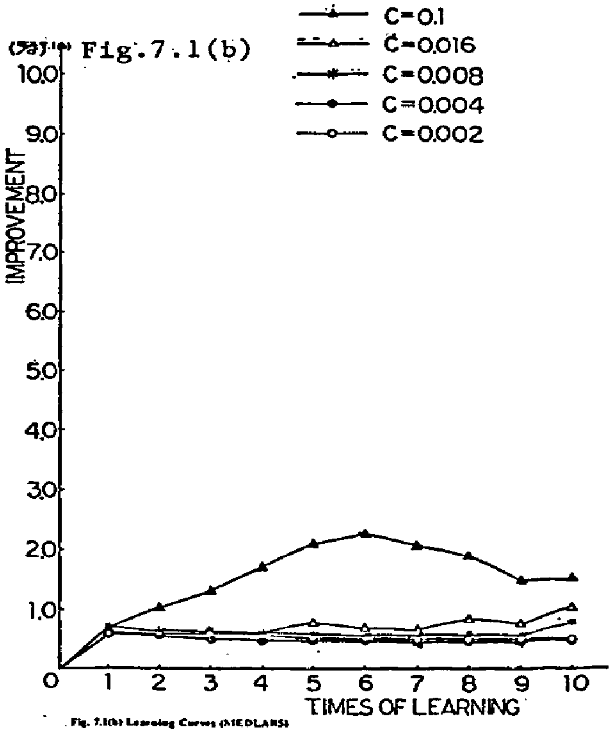

Fig. 7.1(b) Learning Curves (MEDLARS)

Fig.7.1(c)

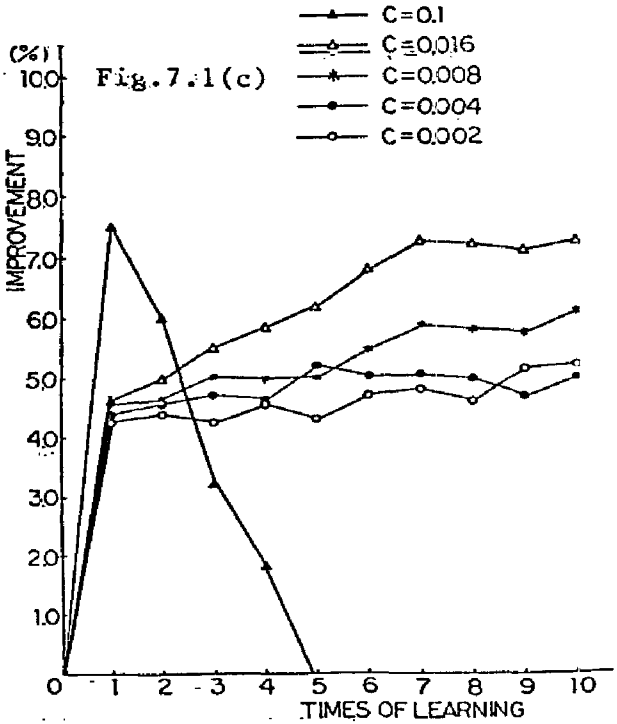

Fig. 7.1(c) Learning Curves (CRN4NUL)

Fig. 7.2

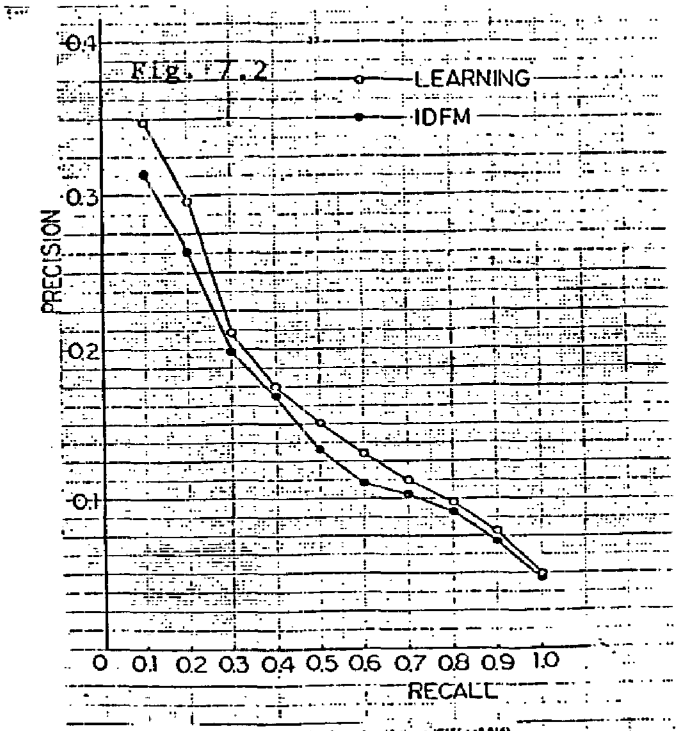

Fig. 7.2 Precision-Recall after Learning (8 times (CISI,c=0.016)

Fig. 7.3

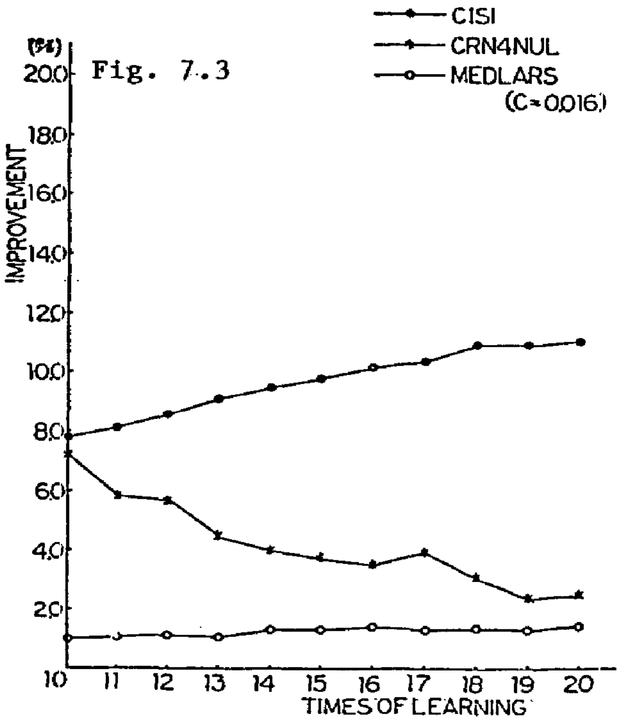

Fig. 7.3 Learning Curves after 10 Times of Learning

Fig. 7.4.

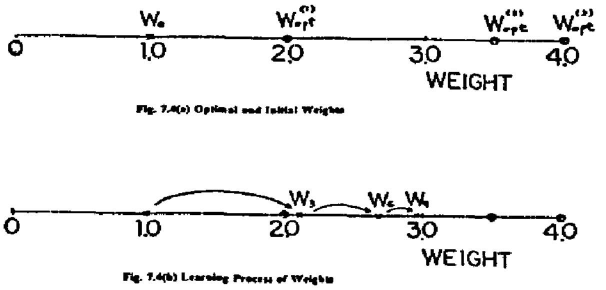

Fig. 7.4(a) Optimal and Initial Weights

Fig. 7.4(b) Learning Process of Weights