Experiments with Generalized Binary Probabilistic Independence Model

S. K. M. Wong
Y. Y. Yao

TR 89-988
April 1989

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

Experiments with

Generalized Binary Probabilistic Independence Model

S.K.M. Wong and Y.Y. Yao

Abstract

This paper reports experiments with the generalized binary probabilistic independence model in which more complete statistical information is used. Two basic sets of experiments, referred to as nonpredictive and predictive, have been performed on two standard test collections. In the nonpredictive experiments, the complete relevance information was used to determine the optimal performance of the model. On the other hand, in the predictive experiments only partial relevance information was used to demonstrate the predictive power of the model. Although a simple method for estimating the parameters was used in the designed tests, significant improvements were obtained for both test collections. These preliminary results suggest that further work on the generalized model is worthwhile.

S.K.M. Wong, Y.Y. Yao, Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

1. Introduction

Two basic probabilistic models have been studied and tested extensively in information retrieval [1-10]. They are referred to as the probabilistic indexing model and the probabilistic retrieval model. In both systems, the probability of relevance is computed for each document with respect to a given query, and the documents are then ranked according to these probabilities. However, the probabilities used in these two models are estimated from different statistical data. In the probabilistic indexing model, the distributions of index terms in a set of queries, together with the relevance feedback information, are used to predict the relevance of a document to a user query. From a group of submitted queries, the system attempts to obtain a reasonable description for each document in the collection (i.e. indexing the documents from a sample of queries). The document descriptions obtained are used, later, to estimate the relevance relationship between any document in the collection and a *new* query. On the other hand, in the probabilistic retrieval model the system learns the query from the relevance judgment of a particular user on a set of documents (i.e. constructing a query from a sample of documents). The estimated query, based on the distributions of index terms in the relevant and nonrelevant document sets, enables the system to predict the relevance of a *new* document to the user.

The main disadvantage of these two probabilistic models is that each of them uses only part of the information that is available. Based on the general framework [8], Wong and Yao [11] have recently proposed a generalized binary probabilistic independence model in which a more complete statistical information is used. A quadratic discriminant function is obtained for the generalized model. This quadratic function can be reduced to the linear discriminant function used by either of the basic models under certain assumptions.

As a complement to the theoretical work [11], this paper reports experiments which are designed to test the feasibility and effectiveness of the generalized binary probabilistic independence model. Two sets of experiments, referred to as nonpredictive and predictive, have been performed on two standard test collections. Initially the complete relevance information was used to determine the optimal performance of the model. However, this has little practical value since the complete relevance information is not known. In order to test the predictive power of the model, two kinds of predictive experiments based on the odd-numbered documents test [2] and the query partition test [12] were conducted. Even though we used a simple method to estimate the parameters, significant improvements were obtained in these tests. These results suggest that further work on the generalized model is worthwhile.

2. Overview of the Generalized Binary Probabilistic Independence Model

In the generalized probabilistic independence model, it is assumed that both documents and queries are represented by binary vectors:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n), \quad \mathbf{y} = (y_1, y_2, \ldots, y_n) .$$

The components x_i and y_i are equal to 0 or 1, indicating the absence or presence of the ith index term in the document and query, respectively. With this representation, a retrieval function consistent with the probability ranking principle can be derived in the manner discussed below.

According to the Bayes decision procedure, a document described by $\mathbf{x}$ is judged to be relevant to a query described by $\mathbf{y}$ if

$$P(relevant \mid \mathbf{x}, \mathbf{y}) > P(nonrelevant \mid \mathbf{x}, \mathbf{y}) . \tag{2.1}$$

From the above decision rule, a discriminant function may be constructed as:

$$g(\mathbf{x}, \mathbf{y}) = \log \frac{P(relevant \mid \mathbf{x}, \mathbf{y})}{P(nonrelevant \mid \mathbf{x}, \mathbf{y})} . \tag{2.2}$$

Since

$$P(relevant \mid \mathbf{x}, \mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y} \mid relevant)\, P(relevant)}{P(\mathbf{x}, \mathbf{y})} ,$$

$$P(nonrelevant \mid \mathbf{x}, \mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y} \mid nonrelevant)\, P(nonrelevant)}{P(\mathbf{x}, \mathbf{y})} ,$$

eqn. (2.2) can be rewritten as:

$$g(\mathbf{x}, \mathbf{y}) = \log \frac{P(\mathbf{x}, \mathbf{y} \mid relevant)}{P(\mathbf{x}, \mathbf{y} \mid nonrelevant)} + \log \frac{P(relevant)}{P(nonrelevant)} , \tag{2.3}$$

where $P\,(relevant)$ and $P\,(nonrelevant)$ are the *a priori* probabilities.

In order to estimate the probabilities in eqn. (2.3), different expressions for $P\,(\mathbf{x},\, \mathbf{y} \mid relevant)$ and $P\,(\mathbf{x},\, \mathbf{y} \mid nonrelevant)$ may be used. The following probabilistic independence assumptions are used in the proposed generalized model:

$$P\,(\mathbf{x},\, \mathbf{y} \mid relevant) = \prod_{i=1}^{n} P\,(x_i,\, y_i \mid relevant) \ ,$$

$$P\,(\mathbf{x},\, \mathbf{y} \mid nonrelevant) = \prod_{i=1}^{n} P\,(x_i,\, y_i \mid nonrelevant) \ . \tag{2.4}$$

This means that the *co-occurrence* of each index term in the document-query pairs, with respect to *relevant* and *nonrelevant*, is assumed to be independent of other terms. Before simplifying the discriminant function under the given independence assumptions, it is convenient to define the following symbols:

$$p_{i0} = P\,(\,x_i = 0,\, y_i = 0 \mid relevant\,),\quad q_{i0} = P\,(\,x_i = 0,\, y_i = 0 \mid nonrelevant\,)\ ,$$

$$p_{i1} = P\,(\,x_i = 0,\, y_i = 1 \mid relevant\,),\quad q_{i1} = P\,(\,x_i = 0,\, y_i = 1 \mid nonrelevant\,)\ ,$$

$$p_{i2} = P\,(\,x_i = 1,\, y_i = 0 \mid relevant\,),\quad q_{i2} = P\,(\,x_i = 1,\, y_i = 0 \mid nonrelevant\,)\ ,$$

$$p_{i3} = P\,(\,x_i = 1,\, y_i = 1 \mid relevant\,),\quad q_{i3} = P\,(\,x_i = 1,\, y_i = 1 \mid nonrelevant\,)\ , \tag{2.5}$$

where p_{ik} is the probability of co-occurrence of the ith index term conditioned on *relevant*, and q_{ik} is the probability conditioned on *nonrelevant*. Using these symbols, eqn. (2.4) can be expressed as:

$$P\,(\mathbf{x},\, \mathbf{y} \mid relevant) = \prod_{i=1}^{n} p_{i0}^{(1-x_i)(1-y_i)}\, p_{i1}^{(1-x_i)y_i}\, p_{i2}^{x_i(1-y_i)}\, p_{i3}^{x_i y_i} \ ,$$

$$P\,(\mathbf{x},\, \mathbf{y} \mid nonrelevant) = \prod_{i=1}^{n} q_{i0}^{(1-x_i)(1-y_i)}\, q_{i1}^{(1-x_i)y_i}\, q_{i2}^{x_i(1-y_i)}\, q_{i3}^{x_i y_i} \ . \tag{2.6}$$

Substituting $P\,(\mathbf{x},\, \mathbf{y} \mid relevant)$ and $P\,(\mathbf{x},\, \mathbf{y} \mid nonrelevant)$ into eqn. (2.3), one obtains:

$$g\,(\mathbf{x},\, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{p_{i2}\, q_{i0}}{p_{i0}\, q_{i2}} + y_i \log \frac{p_{i1}\, q_{i0}}{p_{i0}\, q_{i1}} + x_i\, y_i \log \frac{p_{i0}\, p_{i3}\, q_{i1}\, q_{i2}}{p_{i1}\, p_{i2}\, q_{i0}\, q_{i3}}$$

$$+ \sum_{i=1}^{n} \log \frac{p_{i0}}{q_{i0}} + \log \frac{P(relevant)}{P(nonrelevant)} \quad .$$

$$= \sum_{i=1}^{n} [\, a_i x_i + b_i y_i + c_i x_i y_i \,] + C \quad , \tag{2.7}$$

where

$$a_i = \log \frac{p_{i2}\, q_{i0}}{p_{i0}\, q_{i2}} \quad , \quad b_i = \log \frac{p_{i1}\, q_{i0}}{p_{i0}\, q_{i1}} \quad , \quad c_i = \log \frac{p_{i0}\, p_{i3}\, q_{i1}\, q_{i2}}{p_{i1}\, p_{i2}\, q_{i0}\, q_{i3}} \quad , \tag{2.8}$$

and

$$C = \sum_{i=1}^{n} \log \frac{p_{i0}}{q_{i0}} + \log \frac{P(relevant)}{P(nonrelevant)} \quad . \tag{2.9}$$

The discriminant function given by eqn. (2.7) is a quadratic function on the components of document vector $\mathbf{x}$ and query vector $\mathbf{y}$. For a given query, C and the sum of $b_i y_i$ are constants independent of any document. Therefore, a simpler discriminant function may be used:

$$g_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} [\, a_i x_i + c_i x_i y_i \,] \quad . \tag{2.10}$$

Consider now the problem of estimating the probabilities defined by eqn. (2.5). For each index term, the co-occurrence frequencies $(n_{ik},\, m_{ik})$ are summarized in Table 1.

	$x_i = 0$, $y_i = 0$	$x_i = 0$, $y_i = 1$	$x_i = 1$, $y_i = 0$	$x_i = 1$, $y_i = 1$
relevant	n_{i0}	n_{i1}	n_{i2}	n_{i3}
nonrelevant	m_{i0}	m_{i1}	m_{i2}	m_{i3}

Table 1. Contingency table of term co-occurrence frequencies

The parameters p_{ik}'s and q_{ik}'s may be computed from the formulas:

$$p_{ik} = \frac{n_{ik}}{\sum_{j=0}^{3} n_{ij}} , \quad q_{ik} = \frac{m_{ik}}{\sum_{j=0}^{3} m_{ij}} \quad (k = 0,\ 1,\ 2,\ 3) \ . \tag{2.11}$$

Substituting the above equations into eqn. (2.10) results in

$$g_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} [\, x_i \log \frac{n_{i2}\, m_{i0}}{n_{i0}\, m_{i2}} + x_i\, y_i \log \frac{n_{i0}\, n_{i3}\ \ m_{i1}\, m_{i2}}{n_{i1}\, n_{i2}\ \ m_{i0}\, m_{i3}} \,] \ . \tag{2.12}$$

It should be noted that the number of parameters involved in the generalized model is larger than that of the basic models. The estimation of these parameters depends on the historically accumulated set of data. In the case of a small sample, it may be even more difficult to have an accurate estimation of the required parameters.

The following section reports the results obtained from a series of experiments with the generalized model.

3. Experiments

The experiments were carried out by using two document collections ADINUL and the CRN4NUL in the SMART system [13,14]. The ADINUL collection has 82 documents and 35 queries in the field of documentation and the CRN4NUL has 424 documents and 155 queries on aerodynamics. The collections also include information regarding the relevance of documents for each query. The whole or part of this relevance information was used to estimate the required parameters.

Based on the relevance information used, the experiments can be classified into two categories, nonpredictive and predictive. In the nonpredictive tests, complete relevance information was used to determine the optimal performance of the model. This kind of experiments is sometimes referred to as a retrospective and an upper bound experiment [2,5]. On the other hand, predictive tests used only partial relevance information to demonstrate the predictive power of the model. The predictive experiments can be further divided into two groups. In one group, each document collection is divided into even-numbered and odd-numbered subsets [2]. The parameters estimated from even-numbered subset were applied to the odd-numbered subset. However, if a query has no even-numbered relevant document, or at the other extreme no odd-numbered relevant document, the query may not be suitable for the intended predictive tests. For this reason, queries that are close to these two extremes should be removed in the tests. In the experiments performed, only queries which have almost the same number of even-numbered and odd-numbered relevant documents were selected. Using this query selection criterion, 16 out of 35 and 104 out of 155 queries were chosen for ADINUL and CRN4NUL, respectively. In the other group, based on the query partition algorithm suggested by Raghavan and Yu [12,15], the queries in the collection were partitioned into two sets, the *base* set and the *evaluation* set. The partitioning strategy ensures that every term in the queries of the evaluation set is contained in at least one query in the base set. The details of the algorithm can be found in [12], from which we obtained 31 and 4 queries, respectively, in the base set and evaluation set for the ADINUL collection. For the CRN4NUL collection the corresponding numbers are 105 and 45. The estimated parameters from the base set of queries were then applied to the evaluation set of queries.

In the derived discriminant function (2.10), the sum is over the entire set of index terms. Since the number of index terms is very large, such a summation is computationally expensive. Therefore, two modified functions were used. The first alternative was obtained by summing over the index terms that appear in at least one query. This is similar to the earlier practice in the conventional probabilistic model for reducing the dimensionality [4,5]. The second version is essentially based on the same argument by introducing the following independence assumptions:

$$\begin{aligned} P(x_i = 0, y_i = 0 \mid relevant) &= P(x_i = 0 \mid relevant)\, P(y_i = 0 \mid relevant), \\ P(x_i = 1, y_i = 0 \mid relevant) &= P(x_i = 1 \mid relevant)\, P(y_i = 0 \mid relevant), \\ P(x_i = 0, y_i = 0 \mid nonrelevant) &= P(x_i = 0 \mid nonrelevant)\, P(y_i = 0 \mid nonrelevant), \\ P(x_i = 1, y_i = 0 \mid nonrelevant) &= P(x_i = 1 \mid nonrelevant)\, P(y_i = 0 \mid nonrelevant). \quad (3.1) \end{aligned}$$

Substituting eqn. (3.1) into eqn. (2.8), one immediately obtains:

$$\begin{aligned} a_i &= \log \frac{P(x_i = 1 \mid relevant)\, P(x_i = 0 \mid nonrelevant)}{P(x_i = 0 \mid relevant)\, P(x_i = 1 \mid nonrelevant)}, \\ c_i &= -a_i + \log \frac{p_{i3}\, q_{i1}}{p_{i1}\, q_{i3}}. \end{aligned} \quad (3.2)$$

Since $P(x_i = 1 \mid relevant)$ and $P(x_i = 1 \mid nonrelevant)$ are the probabilities that term t_i occurs in the relevant and nonrelevant documents for an arbitrary query, it may be reasonable to assume that they are equal. Thus, $a_i = 0$ and from eqns. (2.10) and (2.11), we arrive at another discriminant function:

$$g_2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \, y_i \log \frac{n_{i3} \, m_{i1}}{n_{i1} \, m_{i3}} \; . \qquad (3.3)$$

In this function, the sum is over all query terms. It is also interesting to note that this function is essentially the dot product used in vector space model except for the correction factor, $\log (n_{i3} \, m_{i1}) / (n_{i1} \, m_{i3})$.

Another important consideration for the experiments is that there is a possibility that some of the entries in Table 1 may be zero. For this reason, the original contingency table was modified as shown in Table 2.

	$x_i = 0$, $y_i = 0$	$x_i = 0$, $y_i = 1$	$x_i = 1$, $y_i = 0$	$x_i = 1$, $y_i = 1$
relevant	$n_{i0} + .5$	$n_{i1} + .5$	$n_{i2} + .5$	$n_{i3} + .5$
nonrelevant	$m_{i0} + .5$	$m_{i1} + .5$	$m_{i2} + .5$	$m_{i3} + .5$

Table 2. Modified contingency table of term co-occurrence frequencies

This table was used for both the nonpredictive and predictive tests.

The standard recall and precision measures were used for performance evaluation. Recall is defined as the proportion of relevant documents retrieved and precision is the proportion of the retrieved documents actually relevant. The overall performance was determined by computing the average precision over all the queries for recall points 0.1 , 0.2 , . . . , and 1.0 . As a reference point, experiments were also carried out using the coordinate match in which documents and queries are represented by binary vectors. The coordination match is defined as follows:

$$coord(\mathbf{x}, \mathbf{y}) = \sum_{i=i}^{n} x_i \, y_i \quad , \tag{3.4}$$

where **x** is a document vector and **y** is a query vector. The percentage improvement over that of coordination match indicates the effectiveness of the proposed model.

3.1. Nonpredictive tests

In all experiments, the discriminant function (2.12) was modified such that the sum is over the set of index terms which are present in at least one query. The required parameters were estimated from Table 2. Tables 3(a) and 3(b) summarize the results for ADINUL and CRN4NUL collections. In these tables, the results of function (2.12) is labeled by Gprob1 and the results of function (3.3) is labeled by Gprob2. For the ADINUL collection, the average improvements are 146.8 and 148.5 per cent, respectively, for functions (2.12) and (3.3). The corresponding figures for CRN4NUL collection are 69.8 and 84.1 per cent, respectively. The results indicate that improved performance was attained by the generalized model with complete relevance information.

It is interesting to note that the performance of eqn. (3.3) is better than that of eqn. (2.12). This supports the hypothesis that the query terms are more useful than the non-query terms. These results also suggest that further investigation on the selection of a more reliable set of index terms is required.

ADINUL 35 queries 82 documents			
Recall	Precision		
	Coord	Gprob1	Gprob2
0.10	0.3578	0.6159	0.7369
0.20	0.3264	0.5959	0.7028
0.30	0.2702	0.5388	0.6130
0.40	0.1918	0.5115	0.5695
0.50	0.1820	0.5040	0.5511
0.60	0.1529	0.4165	0.4042
0.70	0.1124	0.3233	0.2938
0.80	0.1056	0.2944	0.2646
0.90	0.0918	0.2457	0.2115
1.00	0.0918	0.2422	0.2112
Average percent improvement over coordination match		+146.8	+148.5

Table 3 (a) Nonpredictive test for ADINUL collection

CRN4NUL 155 queries 424 documents			
Recall	Precision		
	Coord	Gprob1	Gprob2
0.10	0.4876	0.6285	0.7288
0.20	0.3992	0.5748	0.6589
0.30	0.3306	0.5037	0.5744
0.40	0.2728	0.4449	0.4901
0.50	0.2438	0.4118	0.4525
0.60	0.1890	0.3324	0.3692
0.70	0.1382	0.2616	0.2748
0.80	0.1087	0.2043	0.2099
0.90	0.0801	0.1576	0.1602
1.00	0.0764	0.1458	0.1528
Average percent improvement over coordination match		+69.8	+84.1

Table 3 (b) Nonpredictive test for CRN4NUL collection

3.2. Predictive tests

Tables 4(a) and 4(b) present the results of the odd-numbered document tests. The parameters estimated from the even-numbered documents were used to retrieve the odd-numbered documents. The coordination match was also carried on the odd-numbered documents. The average improvements over coordination match are 28.7 and 29.8 per cent for the ADINUL collection and -.9 and 17.6 per cent for the CRN4NUL collection. In fact, the scheme of dividing documents into even and odd-numbered subsets is rather arbitrary. The even-numbered documents are not necessarily be related to the odd-numbered documents. This may explain why we obtained less improved results for the CRN4NUL collection. It may be more reasonable to first form clusters of documents and then to divide the documents in each cluster into a training part and an evaluation part.

Tables 5(a) and 5(b) summarize the results of the query partition tests. Significant improvement is observed for both the ADINUL collection and the CRN4NUL collection. For ADINUL, the improvements are 30.3 and 61.8 per cent over the coordination match method. The improvements for the CRN4NUL collection are 13.3 and 36.8 per cent. The comparison of Tables 4 and 5 shows that we obtained better improvements in the query partition tests than the odd-numbered documents tests. This may be explained by the observation that the base set and evaluation set of queries, obtained by the query partition algorithm, are related.

ADINUL 16 queries 41 documents			
Recall	Precision		
	Coord	Gprob1	Gprob2
0.10	0.4954	0.6446	0.6329
0.20	0.4829	0.6446	0.6173
0.30	0.4576	0.6384	0.5860
0.40	0.4529	0.6193	0.5635
0.50	0.4458	0.6152	0.5635
0.60	0.3807	0.4554	0.4617
0.70	0.3799	0.4197	0.4576
0.80	0.3249	0.4165	0.4563
0.90	0.3236	0.4067	0.4551
1.00	0.3233	0.4048	0.4546
Average percent improvement over coordination match		+28.7	+29.8

Table 4 (a) Odd-numbered documents test for ADINUL collection

CRN4NUL 104 queries 212 documents			
Recall	Precision		
	Coord	Gprob1	Gprob2
0.10	0.6320	0.6850	0.7557
0.20	0.6003	0.6390	0.7280
0.30	0.5635	0.6089	0.6914
0.40	0.5354	0.5471	0.6510
0.50	0.5279	0.5369	0.6385
0.60	0.4689	0.4341	0.5323
0.70	0.4389	0.4033	0.5011
0.80	0.4257	0.3927	0.4822
0.90	0.4135	0.3876	0.4744
1.00	0.4134	0.3866	0.4735
Average percent improvement over coordination match		-0.9	+17.6

Table 4 (b) Odd-numbered documents test for CRN4NUL collection

ADINUL 4 queries 82 documents			
Recall	Precision		
	Coord	Gprob1	Gprob2
0.10	0.5694	0.6667	0.8333
0.20	0.5023	0.5671	0.6806
0.30	0.4845	0.5372	0.6412
0.40	0.3222	0.4204	0.6383
0.50	0.2779	0.4061	0.6258
0.60	0.2699	0.4061	0.4586
0.70	0.2528	0.3561	0.4385
0.80	0.2407	0.3561	0.4142
0.90	0.1898	0.2383	0.2518
1.00	0.1898	0.2282	0.2518
Average percent improvement over coordination match		+30.3	+61.8

Table 5 (a) Query partition test for ADINUL collection

CRN4NUL 45 queries 424 documents			
Recall	Precision		
	Coord	Gprob1	Gprob2
0.10	0.4556	0.4982	0.6156
0.20	0.3461	0.4407	0.5030
0.30	0.3140	0.3706	0.4325
0.40	0.2810	0.3410	0.3799
0.50	0.2483	0.3120	0.3370
0.60	0.2144	0.2535	0.3038
0.70	0.1678	0.1769	0.2122
0.80	0.1297	0.1316	0.1692
0.90	0.1059	0.1111	0.1479
1.00	0.1026	0.1037	0.1456
Average percent improvement over coordination match		+13.3	+36.8

Table 5 (b) Query partition test for CRN4NUL collection

4. Conclusion

This paper reports a series of experiments designed to test the effectiveness of the generalized binary probabilistic independence model. The preliminary results are encouraging and they seem to provide some support to the soundness of the proposed method. One of the main drawbacks in the generalized model lies in its independence assumption. Different methods aimed at removing such a strong assumption are currently being investigated.

Acknowledgment

The final version of this paper was prepared while the authors visiting the Department of Computer Science, Cornell University. They wish to thank J. Hopcroft for providing the opportunity to work at Cornell and they are particularly indebted to G. Salton for his encouragement and hospitality.

References

[1] Maron, M.E.; Kuhns, J.L. "On Relevance, Probabilistic Indexing and Information Retrieval." *Journal of the Association for Computing Machinery.* 7(3):216-244; 1960.

[2] Robertson, S.E.; Sparck Jones, K. "Relevance Weighting of Search Terms." *Journal of the American Society for Information Science.* 27:129-146; 1976.

[3] Yu, C.T.; Salton, G. "Precision Weighting -- An Effective Automatic Indexing Method." *Journal of the Association for Computing Machinery.* 23(1):76-88; 1976.

[4] Van Rijsbergen, C.J. "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval." *Journal of Documentation.* 33:106-119; 1977.

[5] Harper, D.J.; Van Rijsbergen, C.J. "An Evaluation of Feedback in Document Retrieval Using Co-occurrence Data." *Journal of Documentation.* 34(3):189-216; 1977.

[6] Croft, W.B; Harper, D.J. "Using Probabilistic Models of Document Retrieval Without Relevance Information." *Journal of Documentation.* 35:285-295; 1979.

[7] Van Rijsbergen, C.J. *Information Retrieval.* London: Butterworth; 1979.

[8] Robertson, S.E.; Maron, M.E.; Cooper, W.S. "Probability of Relevance: A Unification of Two Competing Models for Document Retrieval." *Information Technology: Research and Development.* 1(1):1-21; 1982.

[9] Maron, M.E.; Curry, S.; Thompson, P. "A Inductive Search System: Theory, Design, and Implementation." *IEEE Transections on System, Man, and Cybernetics.* SMC-16(1):21-28; 1986.

[10] Fuhr, N. "Two Models of Retrieval with Probabilistic Indexing." *Proceedings of the ACM SIGIR Conference On Research and Development in Information Retrieval.* 249-257; 1986.

[11] Wong, S.K.M.; Yao, Y.Y. "A Generalized Binary Probabilistic Independence Model." To appear in *Journal of the American Society for Information Science.*

[12] Raghavan, V.V.; Yu, C.T. "Experiments on the Determination of the Relationships Between Terms." *ACM Transactions on Database Systems,* 4(2):240-260; 1979.

[13] Salton, G. (editor). *The SMART Retrieval System --- Experiments in Automatic Document Processing.* Englewood Cliffs, NJ: Prentice-Hall; 1971.

[14] Salton, G.; McGill, M.H. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill; 1983.

[15] Yu, C.T.; Raghavan, V.V. "Single-Pass Method for Determining the Semantic Relationships Between Terms." *Journal of the American Society for Information Science.* 28:345-354; 1977.