POSNIR: Probabilistic Natural Language Information Retrieval System

Changki Lee Seungwoo Lee Gary Geunbae Lee
Dept. of Computer Science & Engineering, POSTECH
San 31, Hyoja-Dong, Pohang, South Korea, 790-784
{leeck, pinesnow, gblee}@postech.ac.kr

Abstract

This paper describes our information retrieval system participated in CLIR task of NTCIR3 and reports the results with some observations. Most previous information retrieval models assume that terms of queries and documents are statistically independent from each another. However, independence assumption is obviously and openly understood to be wrong. We presented two method of incorporating the term dependences in probabilistic retrieval model to compensate the weakness of the linked dependence assumption.

Keywords: information retrieval, probabilistic model, term dependence

1. Introduction

Most previous information retrieval (IR) models assume that terms of queries and documents are statistically independent from each another. Although independence assumption is obviously and openly understood to be wrong [11], many IR models based on this assumption have been developed because the assumption leads to a formal representation of the model more easily, and most IR systems practically have worked well under this assumption.

Many researchers tried to remove the independent assumption and have incorporated various term dependence models with diverse techniques [1][2][7]. However, when a higher order model of term dependence is used, the easily reached formal representation of the model (in fact, the greatest merit of term independence) cannot be maintained and we face extreme difficulties in inducing the probabilities of the model. Nevertheless, it has been clarified that incorporation of a term dependence model actually improved the performance [3][7].

In this paper, we introduce two methods of incorporating term dependence in probabilistic retrieval model. First method is to incorporate term dependence in probabilistic retrieval model by adapting Bahadur-Lazarsfeld Expansion (BLE) using term co-occurrence information. Second method is to incorporate term dependence in probabilistic retrieval model by adapting a structural index system using dependency parse tree and the Chow Expansion (CE). Bahadur-Lazarsfeld Expansion and Chow Expansion were originally used in pattern recognition field [12].

This paper is organized as follows. In section 2, we discuss previous researches on diverse techniques to incorporate the term dependences in different retrieval models and compare them with our own research. In section 3, we describe our POSNIR system participated in CLIR task of NTCIR3 and report the results. In section 4, we describe Bahadur-Lazarsfeld Expansion theory and our adaptation of the theory to 2-Poisson model, particularly Okapi BM25. In section 5, we describe the Chow Expansion theory, the dependency parse tree, and our adaptation of the theory to 2-Poisson model for dependency structured indexing system. In section 6, we draw some conclusions and plans for future works.

2. Previous Researches

Robertson and S. Walker presented an IR model approximating 2-Poisson model, well known as Okapi BM series [8], which integrate within-document term frequency, document length and within-query term frequency. While these models have been widely used in IR, they are based on one important assumption, i.e., linked dependence assumption [11]:

$$\frac{\Pr(A,B \mid rel)}{\Pr(A,B \mid \overline{rel})} = \frac{\Pr(A \mid rel)}{\Pr(A \mid \overline{rel})} \frac{\Pr(B \mid rel)}{\Pr(B \mid \overline{rel})}$$

, where A, B are regarded as properties of documents, and *rel* designates the relevance set.

The linked dependence assumption is considerably weaker than the binary independence assumption, so in most cases, IR systems using the formula based on the linked dependence have shown relatively good experimental results. Nevertheless, it is also an unrealistic assumption, and many researches have tried to address the limitations of the linked dependence assumption by computing the term dependences using diverse techniques on the basis of different retrieval models.

Bollmann-Sdorra and Raghavan showed that, for retrieval functions such as dot products or the cosine used in the vector space model, weighted retrieval is incompatible with term independence in query space [3]. They also proved that the term independence in the query space even turned out to be undesirable.

Croft proposed an approach to integrate Boolean and statistical systems, where boolean queries are interpreted as a way of specifying term dependencies in the relevant set of documents [10].

Losee proposed a probabilistic model integrating boolean query in CNF (Conjunctive Normal Form), where most of the dependences exist between the disjunctions of the terms [5].

Losee also incorporated term dependence information in estimating Pr(d|rel) using the Bahadur-Lazarsfeld Expansion (BLE) [7], and documents were ranked by Expected Precision (EP) of the documents as follows:

$$\Pr(rel \mid d) = \frac{\Pr(d \mid rel)\Pr(rel)}{\Pr(d)}$$

, where d is the vector of a document and rel is a relevant set. Losee performed experiments using Cystic Fibrosis (CF) for spanning the degree of the terms and showed that the best performance was obtained when degree 3 and ±3 to ±5 window of the terms were used. However, in his experiment, he estimated the parameters (all probabilities and correlations for appropriate relevance class) using the 'retrospective' technique. That is, before the retrieval process, all parameters were estimated with the full knowledge of the characteristics of relevant and non-relevant documents. In general, however, we cannot know the relevant and non-relevant documents fully in real situations. Moreover, because of the relatively small sized test collection in his experiments, it is not sufficient to verify that this technique is actually effective in practical situations.

Van Rijsbergen explored one way of removing the independence assumption [1]. He constructed a probabilistic model incorporating dependences between index terms. The extent to which two index terms depend on one another is derived from the distribution of co-occurrences in the whole collection or in the relevant and non-relevant document sets, and used to construct a non-linear weighting function. In a practical situation, the values of some of the parameters of such a function must be estimated from small samples of documents. So a number of estimating rules were discussed and one in particular was recommended.

Turtle described a new formal retrieval model which uses probabilistic inference networks to represent documents and information needs [6]. Retrieval is viewed as an evidential reasoning process in which multiple sources of evidence about document and query content are combined to estimate the probability that a given document matches a query. This model generalizes several current retrieval models and provides a framework within which disparate information retrieval research results can be integrated. The chief advantage of the model is that it allows complex dependencies to be represented in an easily understood form and allows networks containing these dependencies to be evaluated without development of a closed form expression. However the model makes only limited use of term dependence information (phrase and thesaurus information) and should be extended to incorporate additional dependencies (e.g., term clustering).

Much of the works done within the TREC on the use of phrases and passages can be seen as seeking to capture dependencies by more informal means, though there may be other motivations as well [4]. Thus limiting candidate query expansion terms to those occurring in the passage neighborhoods of matching terms can be seen as a way of concentrating on the co-occurrence information so that it is more discriminating than the co-occurrence information computed over extended full texts.

3. POSNIR System

In this section, we describe our POSNIR system participated in CLIR task of NTCIR3 and report the results.

3.1 Keyword Extraction

For keyword extraction, we tagged the document collection and queries using POSTAG/K and POSTAG/J (the Korean and Japanese Part-Of-Speech tagger based on HMM) in Korean and Japanese. In the case of Chinese, we used Chinese segmenter POSTAG/C. The outputs of POSTAG/K and POSTAG/J are composed of lexis, POS tag, and lemma. From the result of the taggers, we selected lemma as keywords. Stop words are eliminated using two kinds of stop list: common stop list and query-specific stop list which must be removed from the query.

For constructing noun phrases, we made lexico-syntactic rules based on the POS-tag patterns. Some of the rules are described below.

Term1/{*NN*|*NP*} Term2/{*NN*|*NP*}
→ Term1_Term2
Term1/*JJ* Term2/{*NN*|*NP*} Term3/{*NN*|*NP*}
→ Term1_Term2_Term3

3.2 Initial Retrieval

Our retrieval system uses 2-Poisson model based on the probabilistic term distribution. The system retrieves top-raked documents after giving scores to each document of a target data collection with each query term list made from the keyword extraction process. For scoring, a rank system uses Okapi BM25 formula as shown below.

Run type	Query expansion	Avg Prec	Prec@10	Prec@100	R-Prec
K-K-C rigid	PRF	0.21570	0.30670	0.17870	0.24920
K-K-D rigid	No	0.17440	0.26000	0.12670	0.21370
K-K-D rigid	PRF	0.19570	0.31330	0.14500	0.24390
J-J-C rigid	No	0.24480	0.32380	0.15480	0.25580
J-J-D rigid	No	0.21280	0.29520	0.14980	0.21040
J-J-T rigid	No	0.23760	0.32140	0.15620	0.24380
C-C-C rigid	No	0.22030	0.31670	0.14330	0.25250
C-C-D rigid	No	0.17260	0.28100	0.11830	0.20150
C-C-T rigid	No	0.19280	0.28100	0.12950	0.22000

Table 1. Results of NTCIR3 CLIR task

$$w^{(1)} = \log\left(\frac{N-n+0.5}{n+0.5}\right), \quad (1)$$

$$Score(d,q) = \sum_{t\in q}\left(\frac{(k_1+1)\times tf_t}{k_1\times\left((1-b)+b\times\frac{dl_d}{avdl}\right)+tf_t}\times w^{(1)}\times\frac{(k_3+1)\times tf_q(q,t)}{k_3+tf_q(q,t)}\right) \quad (2)$$

, where N is the number of documents in the collection, n is the number of documents containing the term, tf_t is the term frequency of term t in a document d, dl_d is the document length, $avdl$ is the average document length, $tf_q(q,t)$ is the term frequency of query term t in the query q, and k_1 , b, k_3 are tunable constant parameters.

3.3 Query Expansion

Query expansion is achieved through PRF (Pseudo Relevance Feedback). In the process of PRF, top-ranked documents are regarded as relevant and TSV (Term Selection Value) is given to all single terms except stop words in them. Then, top-ranked single terms are expanded and added to the original query term list. In this process, the weights of both original and expanded query terms are re-weighted by Eq.(3) reflecting relevance and non-relevance information [9].

$$w^{(1)} = \frac{k_5}{k_5+\sqrt{R}}\left(k_4+\log\frac{N}{N-n}\right)+\frac{\sqrt{R}}{k_5+\sqrt{R}}\left(\log\frac{r+0.5}{R-r+0.5}\right) -\left(\frac{k_6}{k_6+\sqrt{S}}\log\frac{N}{N-n}\right)-\left(\frac{\sqrt{S}}{k_6+\sqrt{S}}\log\frac{s+0.5}{S-s+0.5}\right) \quad (3)$$

, where N, n is the same as in the Eq.(1), R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, S is the number of documents known to be non-relevant, s is the number of non-relevant documents containing the term, and k_5, k_6 are tunable constant parameters.

For TSV function, we developed a formula adapting TF (Term Frequency) factor as follows:

$$TSV = \left(\sum_{d\in R}\frac{tf_{t,d}}{k_1\left((1-b)+b\frac{dl_d}{avgdl}\right)+tf_{t,d}}\right)\times w^{(1)}$$

3.4 Experiments in NTCIR3

We performed experiments on three languages (Korean, Japanese, and Chinese). Because we do not have sufficient test collections of Japanese and Chinese to set parameters of our model, we used PRF only K-K-C and K-K-D run. So we did not perform query expansion at other runs except K-K-C and K-K-D runs. Table 1 summarizes the CLIR task results. The results indicate that when a query was expanded using PRF, the performance was better.

4. BLE Adaptation to 2-Poisson Model

We propose a method of incorporating the term dependence into probabilistic models, in particular 2-Poisson models, using Bahadur-Lazarsfeld Expansion (BLE) [7][12]. Document probabilities can also be estimated on the BLE. An exact probability is calculated when full expansion is used, but if the expansion is truncated, an estimate of the probability is computed. The expansion begins with the estimate of the independent probability, and then proceeds by multiplying this independent probability by a correction factor. Each successive approximation accounts for the correlations of one more higher order, but naturally requires the computation of additional terms. When the full expansion is used, the exact probability of a document is computed as follows:

$$\Pr(d) = \Pr_1(d)\left[1+\sum_{i<j}\rho_{ij}y_iy_j+\sum_{i<j<k}\rho_{ijk}y_iy_jy_k+\ldots+\rho_{12\ldots n}y_1y_2\ldots y_n\right] \quad (4)$$

$Pr_1(d)$ is the probability of a document by assuming the term independence, computed as:

$$\Pr_1(d) = \prod_{i=1}^{n} p_i^{x_i} (1-p_i)^{1-x_i}$$

, where d is represented as a binary vector $x=\{x_1,x_2,...,x_n\}$, $p_i=Pr(d_i = 1)$, and $1\text{-}p_i = Pr(d_i = 0)$. y_i is a normalization variable of x_i, computed as:

$$y_i = \frac{x_i - p_i}{\sqrt{p_i(1-p_i)}}$$

ρ_{ij} is a correlation coefficient of x_i and x_j, computed as:

$$\rho_{1,2,...i} = E[y_1 y_2 \cdots y_i] = \sum_d y_1 y_2 \cdots y_i \Pr(d)$$

The sum of correction factor of Eq. (4) may be arbitrarily truncated so that one can include all dependences with term pairs, term triples, and so on.

We incorporate the term dependence into the state-of-the-art 2-Poisson model, in particular Okapi BM25, using the BLE. We adapt the BLE to BM25 up to the second order as follows [18]:

$$Score(d,q) = \sum_{i \in q} qtf_i \frac{tf_i}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_i} w^{(1)} - \beta \log\left(1 + k_7 + \sum_{i<j} \rho_{ij} y_i y_j\right)$$

, where $w^{(1)}$ is the IDF estimated with no relevance information, and β and k_7 are constant parameters.

4.1 BLE Experiments

We will empirically demonstrate that the 2-Poisson model incorporating BLE term dependences gives a significantly better performance than the 2-Poisson model under the conventional linked dependence assumption. For a more convincing demonstration, we performed the experiments on two different languages, Korean and English. For Korean, we use the HANTEC2.0 collection [17] and 50 topics for test, and for English, we use WT10g collection for TREC evaluation and 50 topics of TREC-9 (Topic 451~450). All diverse experiments are carried out by varying models of POSNIR/K and POSNIR/E, which is a POStech Natural language Information Retrieval system for Korean and English respectively.

Term pairs are generated from single terms of original query term list. We extracted the term pairs within a sentence of query in our experiments. We consider a sentence as a window of documents as we did in the query. We can calculate ρ_{ij} by counting the number of documents where term i and term j appear in a sentence simultaneously. The counting is based on the assumption that the co-occurrence terms have a reasonably short distance in the document.

The results of our experiments are listed in the Table 2 and Table 3.

Cases	AvgP	P@5	P@10	R-P
BM25	0.2381	0.5440	0.4900	0.2826
BM25→BLE	**0.2464**	**0.5640**	**0.5020**	**0.2903**

Table 2. Performance on the HANTEC2.0 collection

Cases	AvgP	P@5	P@10	R-P
BM25	0.2027	0.3520	0.2860	0.2442
BM25→BLE	**0.2064**	0.3520	**0.2920**	**0.2495**

Table 3. Performance on the WT10g collection

Comparing 'BM25' with 'BM25→BLE', the performance of 'BM25→BLE' was better than 'BM25' on HANTEC2.0 as well as on WT10g collection. From the above results, the 2-Poisson model incorporating term dependence with BLE techniques significantly improves the performance.

5. Probabilistic Model for Dependency Structured Indexing System

In this chapter, we introduce a method of incorporating the term dependence into probabilistic models, in particular 2-Poisson models, using the Chow Expansion [14] and a dependency structured indexing system.

5.1 Chow Expansion Theory and Dependency Parse Tree

When the components of the vector $x=\{x_1,x_2,...,x_n\}$ are binary values, the problem of estimating a density becomes the problem of estimating the probability. Since there are 2^n possible vectors x, we must estimate 2^n probabilities, which is an enormous task.

If the components of x are statistically independent, the problem is greatly simplified. In this case we can write

$$\Pr(x) = \prod_{i=1}^{n} \Pr(x_i) = \prod_{i=1}^{n} p_i^{x_i} (1-p_i)^{1-x_i}$$

, where $p_i = \Pr(x_i=1)$ and $1\text{-}p_i = \Pr(x_i=0)$.

It is natural to ask whether or not there are any compromise positions betweens being completely accurate, which requires estimating 2^n probabilities, and being forced to assume statistical independence,

which reduce the problem to one of estimating only n probabilities. One answer is provided by finding an expansion for Pr(x) and approximating Pr(x) by partial sum, e.g. the Rademacher-Walsh Expansion and the Bahadur-Lazarsfeld Expansion [12].

Another interesting class of approximation to a joint probability distribution Pr(x) is based on the identity

$$\begin{aligned}\Pr(x) &= \Pr(x_1,\cdots,x_n) \\ &= \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(x_3 \mid x_2,x_1)\cdots\Pr(x_n \mid x_{n-1},\cdots,x_1)\end{aligned}$$

Suppose the variables are not independent, but we can number the variables so that $\Pr(x_i|x_{i-1},\ldots,x_1)$ is solely dependent on some preceding variable $x_{j(i)}$. Then we obtain the product expansion

$$\Pr(x) = \Pr(x_1)\Pr(x_2 \mid x_{j(2)})\Pr(x_3 \mid x_{j(3)})\cdots\Pr(x_n \mid x_{j(n)})$$

By substituting 0 or 1 for x_i and $x_{j(i)}$, we can verify that

$$\Pr(x_i \mid x_{j(i)}) = \left[p_{i|j(i)}^{x_i}(1-p_{i|j(i)})^{1-x_i}\right]^{x_{j(i)}}\left[p_i^{x_i}(1-p_i)^{1-x_i}\right]^{1-x_{j(i)}}$$

, where $p_{i|j(i)} = \Pr(x_i=1|x_{j(i)}=1)$ and $p_i = \Pr(x_i=1|\ x_{j(i)}=0)$.

By letting $p_i = \Pr(d_i=1)$, taking the logarithm, and collecting terms, we obtain the Chow Expansion [14].

$$\begin{aligned}\log\Pr(x) &= \sum_{i=1}^{n}\log(1-p_i) + \sum_{i=1}^{n}x_i\log\frac{p_i}{1-p_i} \\ &+ \sum_{i=2}^{n}x_{j(i)}\log\frac{1-p_{i|j(i)}}{1-p_i} + \sum_{i=2}^{n}x_i x_{j(i)}\log\frac{p_{i|j(i)}(1-p_i)}{(1-p_{i|j(i)})p_i}\end{aligned}$$

Chow and Liu suggest the construction of a tree such that the mutual information between a variable and the variable immediately above it are maximized.

A dependency relationship [15] is an asymmetric binary relationship between a word called head (or governor, parent), and another word called modifier (or dependent, daughter). Dependency grammars represent sentence structures as a set of dependency relationships. Normally the dependency relationships from a tree connect all the words in a sentence. A word in the sentence may have several modifiers, but each word may modify at most one word. The root of the dependency tree does not modify any word. It is also called the head of the sentence.

Since a dependency parse tree represents the term dependence relations in the syntactic structure, we can apply this dependency parse tree which is generated by linguistic dependency parser in the Chow Expansion, instead of the mutual information MST (minimum spanning tree).

We developed a simple dependency parser for Korean to apply dependency parse trees to the Chow Expansion. Our dependency parser uses some heuristics which are generally used in dependency parsing [16] (e.g. Non-crossing condition, Constraint under surface information and nearest modifyee principle). Our dependency parser shows about 70% precision.

5.2 Adapting the Chow Expansion to the Dependency Structured Indexing System

Chow and Liu suggest the construction of a MST using mutual information for a dependence tree which is originally used in the Chow Expansion. However, we use a dependency parse tree which is generated by linguistic dependency parser instead of the mutual information MST, because a dependency parse tree intuitively and linguistically represents the term dependence relations in the syntactic structure. We consider the use of syntactic structure as one way of relaxing the independence assumption of the terms. So, we use a dependency structured indexing system which consists of the dependency parse tree and Chow Expansion to relax the independence assumption.

We adapt the Chow Expansion to the 2-Poisson model using dependency structured indexing system as follows [19]:

$$\begin{aligned}&Score(d,q) \\ &= \sum_{i\in q} qtf_i \frac{tf_i}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_i}\left(x_i\log\frac{N}{n_i} + x_{j(i)}k_7\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)} + x_i x_{j(i)}k_8\log\frac{q_i q_{j(i)}}{q_{i,j(i)}}\right)\end{aligned}$$

, where d is represented as a binary vector $x=\{x_1,x_2,\ldots,x_n\}$. k_1, b, k_7 and k_8 are constant parameters.

5.3 Chow Expansion Experiments

We evaluated the 2-Poisson model incorporating the Chow Expansion term dependences with the ETRI-KEMONG test collection which is a Korean encyclopedia published by the Kemong company [13]. The test set contains 46 natural language queries and the relevance information of the entry lists related to each query.

Figure 1 shows the dependency structured indexing system. Through the dependency structure analysis of the documents, we extract single terms and dependency relationship for indexing. In retrieving, keyword extraction from the query is performed in the same manner as the indexing process.

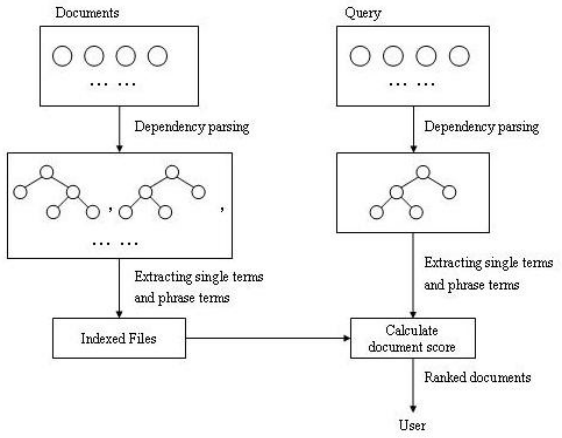

Figure 1. Dependency structured indexing system

The results of our experiments for all the cases are listed in the Table 4.

Cases	AvgP	P@5	P@10	R-P
BM25	0.6052	0.4391	0.2804	0.5241
BM25→CE	**0.6195**	0.4391	**0.2826**	**0.5534**

Table 4. Performance on the ETRI-KEMONG collection

Comparing 'BM25' with 'BM25→CE', the performance of 'BM25→CE' was better than 'BM25' on ETRI-KEMONG collection. From the above results, the probabilistic model for dependency structured indexing system seems to improve the performance, but more experiments should be called for.

6. Conclusions

We described our POSNIR system participated in CLIR task of NTCIR3 and report the results. And we presented two other extended methods of incorporating the term dependences in probabilistic retrieval model to compensate the weakness of the previous linked dependence assumption. Due to the time constraint, however, we couldn't apply these new methods to this year's NTCIR task, but we carried out some independent experiments to verify the proposed models. From the results, improvement of the performance was observed on the document collections in two extended models.

The disadvantage in using the BLE is that the retrieval cost becomes very high because co-occurrence information between the two terms must be obtained at the search time when the user query is given. The longer the size of the query is, the greater the number of term pairs are, which incurs a much higher retrieval cost. To reduce this cost, useless term pairs can be removed at a certain threshold. However, this is not the general solution, so effective algorithms or auxiliary DB's to pre-obtain the co-occurrence information will need to be developed in the future.

The disadvantage in using the Chow Expansion is that the retrieval cost of dependence tree becomes very high because the dependence tree of the user query is obtained by dependency parser at the search time and co-occurrence information between the two terms are obtained by dependency parser at the indexing time. To reduce this cost, very fast and robust dependency parser will need to be developed in the future.

Another future project will be to apply the BLE and Chow Expansion to PRF. Many researches on query expansions using PRF have verified a significant performance improvement. But it is not yet known whether BLE and Chow Expansion techniques indeed work well on the ARF expanded queries or not, and this question is another interest point to the BLE and Chow Expansion based term dependency model.

References

[1] C.J. Van Rijsbergen. A theoretical basis for use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106-119, 1977.

[2] Clement T. Yu, Chris buckley, K. Lam, and Gerard Salton. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2(4):129-154, 1983.

[3] Peter Bollmann-Sdorra and Vijay V. Raghavan. On the Necessity of Term Dependence in a Query Space for Weighted Retrieval. *Journal of the American Society of Information Science*, 49(13): 1161-1168, 1998.

[4] K. Spark Jones, S. Walker and S.E. Robertson. A probabilistic model of information retrieval: Development and status. Technical Report 446, University of Cambridge Computer Laboratory, 1998.

[5] Robert M. Losee and Abraham Bookstein. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. *Information Processing and Managements*, 24(3):315-321, 1988.

[6] H.R. Turtle and W.B. Croft. Inference Networks for document Retrieval. In Intern. Conf. on Research and Development in Information Retrieval, pages 1-24, SIGIR, 1990.

[7] Rorbert M. Losee. Term dependence: Truncating the Bahadur-Lazarsfeld expansion. *Information Processing and Managements*, 30(2):293-303, 1994.

[8] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *In Proceedings of the ACM SIGIR'94*, 232-241, 1994.

[9] S.E. Robertson and Walker S. On relevance weights with little relevance information, *In Proceedings of the ACM SIGIR'97*. 16-24, 1997.

[10] W. Bruce Croft. Boolean queries and term dependencies in probabilistic retrieval models. *Journal of the American Society for Information Science and Technology*, 37(2):71-77, 1986.

[11] William S. Cooper. Some Inconsistencies and Misnomers in Probabilistic Information Retrieval. *In Proceedings of the ACM SIGIR'91*, 57-61, 1991.

[12] Richard O. Duda, Peter E. Hart. Pattern Classification and Scene Analysis. *A Wiley-Interscience publication*, 111-113, 1973.

[13] Kemong. *The Kemong Company new encyclopedia*. Seoul: Kemongsa Publishing Co. 1992.

[14] Chow, C., and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory, IT-14(3), 462-467*. 1968.

[15] David Hays. Dependency theory: a formalism and some observations. *Language*, 40:511-525, 1964.

[16] Sadao Kurohashi, Makoto Nagao. KN Parser: Japanese Dependency/Case Structural Analyzer. *Proceedings of the Workshop on Sharable Natural Language Resources*, pp48-55. 1994.

[17] Suk-Hoon Lee, Sung Hyon Myaeng, Ji-Young Kim, Dong-Hyun Jang, Jeong-Hyun Seo, Hyun Kim. Packaging Hangul Test Collection as an Evaluation System of Information Retrieval, *In Proceeding of The 5th Korea Science & Technology Infrastructure Worksho*p, 31-48, 2000 (in Korean).

[18] Bong-Hyun Cho, Changki Lee, Gary Geunbae Lee. Exploring Term Dependences in Probabilistic Information Retrieval Model. *Information Processing and Managements*, (accepted).

[19] Changki Lee and Gary Geunbae Lee. Probabilistic Information Retrieval Model for Dependency Structured Indexing System, *In Proceedings of the ACM SIGIR'02 Workshop on Mathematical/Formal Methods in Information Retrieval*, 2002.