PROBABILISTIC INFORMATION RETRIEVAL MODEL FOR DEPENDENCY STRUCTURED INDEXING SYSTEM

Changki Lee, Gary Geunbae Lee
Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31 Hyoja dong, Nam Gu, POHANG, 790-784, KOREA
Phone: +82-54-279-5581, Fax: +82-54-279-2299, {leeck, gblee}@postech.ac.kr

Abstract

Most previous information retrieval (IR) 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, so we present a new method of incorporating term dependence in probabilistic retrieval model by adapting a structural index system using dependency parse tree and the Chow Expansion to compensate the weakness of the assumption. In this paper, we describe a theoretic process to apply the Chow Expansion to the general probabilistic models and the state-of-the-art 2-Poisson model, and we re-examine the weight of phrase terms. Through the experiments on document collections, ETRI-KEMONG in Korean, we demonstrate that the incorporation of term dependences using the Chow Expansion contribute to the improvement of performance in Probabilistic IR systems.

Key words

Term dependence, Phrasal indexing, Chow Expansion, Probabilistic model, 2-Poisson model

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 [18], 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][3][10][11]. 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 [6][10][17].

Many approaches that have traditionally been regarded as a tool for increasing precision or relaxing the independent assumption is the use of phrases for indexing and retrieval of documents. Phrases have been found to be a useful indexing unit by most of the leading groups participating at NIST and DARPA sponsored Text REtrieval Conferences for performance evaluations of IR systems [7].

In this paper, we propose a new method of incorporating term dependence in probabilistic retrieval model by adapting a structural index system using dependency parse tree and the Chow Expansion which was originally used in pattern recognition field [21]. We focused on two different aspects: First, we describe a theoretic process in applying the Chow Expansion to the original probabilistic model [14] and state-of-the-art 2-Poisson model [13], then we re-examine the weight of the phrase terms. Second, we try to empirically verify a statistically significant improvement of the performance for our proposed models in practical Korean IR system. The experiments were conducted on a standard document collection of Korean, ETRI-KEMONG [26].

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 the Chow Expansion theory and the dependency parse tree, and in section 4, describe our adaptation of the theory to probabilistic IR models and 2-Poisson models, particularly Okapi BM25. In section 5, we illustrate our retrieval procedure using the model and describe

comprehensive experiment results and analyses. In section 6, we draw some conclusions and plans for future works.

2. PREVIOUS RESEARCHES

Robertson & Sparck Jones originally proposed a probabilistic retrieval model based on the distribution of query terms in relevant and non-relevant documents [14] (as Eq. 1), and for which Robertson and S. Walker presented a formula combining prior weights and pure evidence-based estimates [15]. The estimated formula actually approximates Inverse Document Frequency (IDF) when there is no relevant information.

$$\log \frac{\Pr(rel \mid d)}{\Pr(\overline{rel} \mid d)} = \log \frac{\Pr(d \mid rel)}{\Pr(d \mid \overline{rel})} + \log \frac{\Pr(rel)}{\Pr(\overline{rel})} \quad (1)$$

where d is a document description and rel designates the relevance set.

Among the probabilistic retrieval model, the 2-Poisson model proposed by Bookstein and Swanson [22] assumes that a content word plays two different roles in documents. In documents with a low average number of word occurrences, the occurrences are accidental and should therefore not be used as index terms. On the other hand, in documents with a high average number of word occurrences, the word is a central content word and therefore a good index term. Robertson and S. Walker presented an IR model approximating this 2-Poisson model, well known as Okapi BM series [13], 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 [18]:

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

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 [6]. 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 [16]. Later, he and Lewis presented an algorithm to generate dependent term groups from their own developed representations [17]. In the same paper, they showed the performance improvements were gained by re-ranking the top 100 documents based on the dependent term groups.

Losee also proposed a probabilistic model integrating boolean query in CNF (Conjunctive Normal Form), where most of the dependences exist between the disjunctions of the terms [8]. Other probabilistic models incorporating the term dependences using the maximum entropy techniques [5][19] were also proposed. However, these models require high cost of computing, so the parameter estimation cannot be performed in real time.

Previously, Losee incorporated term dependence information in estimating *Pr(d|rel)* using the Bahadur-Lazarsfeld Expansion (BLE) [10], and documents were ranked by Expected Precision (EP) of the documents [24] 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) [23] 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.

Losee also proposed that Expected Mutual Information Measure (EMIM) is superior to Inverse Document Frequency (IDF) for the weighting function [11]. In fact, the two measures are actually similar to each other on the

theoretical ground based on Luhn [4] and Zipf [20] model, but it is meaningful to have some empirical experiment results.

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 [9]. 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 Programme 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 [7]. 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. CHOW EXPANSION THEORY AND DEPENDENCY PARSE TREE

When the components of the vector $d=\{d_1,d_2,\ldots,d_n\}$ are binary values, the problem of estimating a density becomes the problem of estimating the probability $\Pr(x=d)$. Since there are 2^n possible vectors d, we must estimate 2^n probabilities, which is an enormous task.

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

$$\Pr(d) = \prod_{i=1}^{n} \Pr(d_i) = \prod_{i=1}^{n} p_i^{d_i} (1-p_i)^{1-d_i}$$

where $p_i = \Pr(d_i=1)$ and $1-p_i = \Pr(d_i=0)$.

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 $\Pr(d)$ and approximating $\Pr(d)$ by partial sum, e.g. the Rademacher-Walsh Expansion and the Bahadur-Lazarsfeld Expansion [21].

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

$$\begin{aligned}\Pr(d) &= \Pr(d_1, \cdots, d_n) \\ &= \Pr(d_1)\Pr(d_2 \mid d_1)\Pr(d_3 \mid d_2, d_1)\cdots\Pr(d_n \mid d_{n-1}, \cdots, d_1).\end{aligned}$$

Suppose the variables are not independent, but we can number the variables so that $\Pr(d_i|d_{i-1}, \ldots, d_1)$ is solely dependent on some preceding variable $d_{j(i)}$. Then we obtain the product expansion

$$\Pr(d) = \Pr(d_1)\Pr(d_2 \mid d_{j(2)})\Pr(d_3 \mid d_{j(3)})\cdots\Pr(d_n \mid d_{j(n)}). \tag{2.1}$$

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

$$\Pr(d_i \mid d_{j(i)}) = \left[p_{i|j(i)}^{d_i}(1-p_{i|j(i)})^{1-d_i}\right]^{d_{j(i)}} \left[p_i^{d_i}(1-p_i)^{1-d_i}\right]^{1-d_{j(i)}} \tag{2.2}$$

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

By letting $p_i = \Pr(d_i=1)$, substituting Eq. (2.2) in Eq. (2.1), taking the logarithm, and collecting terms, we obtain the Chow Expansion [27].

$$\begin{aligned}\log\Pr(d) = &\sum_{i=1}^{n}\log(1-p_i) + \sum_{i=1}^{n} d_i \log\frac{p_i}{1-p_i} + \sum_{i=2}^{n} d_{j(i)} \log\frac{1-p_{i|j(i)}}{1-p_i} \\ &+ \sum_{i=2}^{n} d_i d_{j(i)} \log\frac{p_{i|j(i)}(1-p_i)}{(1-p_{i|j(i)})p_i}\end{aligned} \tag{3}$$

Similar results for higher-order dependence can be obtained in an obvious way.

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 [27]. Given two points on the tree such that the *i*th point is directly and immediately above the *j*th point, a Maximum Spanning Tree (MST) may be defined as maximizing the sum:

$$\sum_{i,j} I(Node_i, Node_j)$$

where *I(i,j)* represents the expected mutual information provided by *i* about *j*,

$$I(i,j) = \sum_{i,j} \Pr(i,j) \log \frac{\Pr(i,j)}{\Pr(i)\Pr(j)}.$$

Syntactical phrases can be used to obtain the phrases in linguistic information retrieval. 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.

A dependency relationship [28] 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.

For example, figure (1) is a dependency structure of a sentence. The head of the sentence is “have”. There are four pairs of dependency relationships, depicted by four arcs from heads to modifiers.

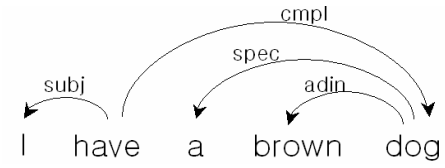

Figure -1. A dependency structure of a sentence.

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 [29] (e.g. Non-crossing condition, Constraint under surface information and Nearest modifiee principle). Our dependency parser shows about 70% precision[1].

4. ADAPTING TO THE PROBABILISTIC IR MODEL

We propose a method of incorporating the term dependence into probabilistic models, in particular 2-Poisson models, using the Chow Expansion [27] and a structural 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 suggest using 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 phrases as one way of relaxing the independence assumption of the terms. So, we use a structural indexing system which consists of the dependency parse tree and Chow Expansion to relax the independence assumption.

4.1 Adapting the Chow Expansion

The $\Pr(d \mid rel)$ and $\Pr(d \mid \overline{rel})$ of Eq. (1) are transformed by adapting the Chow Expansion (Eq. 3) as follows:

$$\log \Pr(d \mid rel) = \sum_{i=1}^{n} d_i \log \frac{p_i}{1-p_i} + \sum_{i=2}^{n} d_{j(i)} \log \frac{1-p_{i|j(i)}}{1-p_i} + \sum_{i=2}^{n} d_i d_{j(i)} \log \frac{p_{i|j(i)}(1-p_i)}{(1-p_{i|j(i)})p_i} + \text{constant}$$

$$\log \Pr(d \mid \overline{rel}) = \sum_{i=1}^{n} d_i \log \frac{q_i}{1-q_i} + \sum_{i=2}^{n} d_{j(i)} \log \frac{1-q_{i|j(i)}}{1-q_i} + \sum_{i=2}^{n} d_i d_{j(i)} \log \frac{q_{i|j(i)}(1-q_i)}{(1-q_{i|j(i)})q_i} + \text{constant}$$

[1] The percentage of dependency relationships in the output of a dependency parser that are also found in the answer

where $p_{i|j(i)} = \Pr(d_i = 1 \mid d_{j(i)} = 1, rel)$, $p_i = \Pr(d_i = 1 \mid rel)$, $q_{i|j(i)} = \Pr(d_i = 1 \mid d_{j(i)} = 1, \overline{rel})$ and $q_i = \Pr(d_i = 1 \mid \overline{rel})$.

In the equations above, we note that if the variables are indeed independent, $p_i = p_{i|j(i)}$ and the last two sums in the expansion disappear, leaving the familiar expansion for the independent case. When dependence exists, we obtain additional linear and quadratic terms.

From the equations above, we can adapt the Chow Expansion to the probabilistic IR model [14] as follows:

$$\begin{aligned}
&\log \frac{\Pr(d \mid rel)}{\Pr(d \mid \overline{rel})} \\
&= \log \Pr(d \mid rel) - \log \Pr(d \mid \overline{rel}) \\
&= \sum_{i=1}^{n} d_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=2}^{n} d_{j(i)} \left(\log \frac{1-p_{i|j(i)}}{1-p_i} - \log \frac{1-q_{i|j(i)}}{1-q_i} \right) \\
&\quad + \sum_{i=2}^{n} d_i d_{j(i)} \left(\log \frac{p_{i|j(i)}(1-q_{i|j(i)})}{q_{i|j(i)}(1-p_{i|j(i)})} - \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \right) + \text{constant}
\end{aligned} \tag{4}$$

Since $\Pr(d_i = 1 \mid d_{j(i)} = 1, rel) = \Pr(d_i = 1, d_{j(i)} = 1 \mid rel) / \Pr(d_{j(i)} = 1 \mid rel)$, Eq. (4) are transformed as follows:

$$\begin{aligned}
&\log \frac{\Pr(d \mid rel)}{\Pr(d \mid \overline{rel})} \\
&= \sum_{i=1}^{n} d_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=2}^{n} d_{j(i)} \left(\log \frac{p_{j(i)} - p_{i,j(i)}}{p_{j(i)}(1-p_i)} - \log \frac{q_{j(i)} - q_{i,j(i)}}{q_{j(i)}(1-q_i)} \right) \\
&\quad + \sum_{i=2}^{n} d_i d_{j(i)} \left(\log \frac{p_{i,j(i)}(1-q_{i,j(i)})}{q_{i,j(i)}(1-p_{i,j(i)})} - \log \frac{p_i(1-q_i)}{q_i(1-p_i)} - \log \frac{p_{j(i)}(1-q_{j(i)})}{q_{j(i)}(1-p_{j(i)})} \right) \\
&\quad + \text{constant}
\end{aligned} \tag{5}$$

where $p_{i,j(i)} = \Pr(d_i = 1, d_{j(i)} = 1 \mid rel)$, $p_i = \Pr(d_i = 1 \mid rel)$, $q_{i,j(i)} = \Pr(d_i = 1, d_{j(i)} = 1 \mid \overline{rel})$ and $q_i = \Pr(d_i = 1 \mid \overline{rel})$. $p_{i,j(i)}$ can be regarded as a probability of a phrase which consists of term i and term $j(i)$ given that it is relevant, and $q_{i,j(i)}$ is a counter part probability for non-relevant document.

We define $MS_{prob\text{-}Chow(R)}$, a query-document scoring function which adapts the Chow Expansion to the probabilistic model as follows:

$$
\begin{aligned}
&MS_{prob-Chow(R)}(d) \\
&= \sum_{i=1}^{n} d_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=2}^{n} d_{j(i)} \left(\log \frac{p_{j(i)} - p_{i,j(i)}}{p_{j(i)}(1-p_i)} - \log \frac{q_{j(i)} - q_{i,j(i)}}{q_{j(i)}(1-q_i)} \right) \qquad (6) \\
&+ \sum_{i=2}^{n} d_i d_{j(i)} \left(\log \frac{p_{i,j(i)}(1-q_{i,j(i)})}{q_{i,j(i)}(1-p_{i,j(i)})} - \log \frac{p_i(1-q_i)}{q_i(1-p_i)} - \log \frac{p_{j(i)}(1-q_{j(i)})}{q_{j(i)}(1-p_{j(i)})} \right)
\end{aligned}
$$

This model consists of linear and quadratic terms. If $p_{i,j(i)}=p_i*p_j$, that is, term i and term $j(i)$ are statistically independent, the model is same to the independence model which consists of only linear terms. And the model is similar to the models which use the weight of phrases as a single term weight, if we ignore the second sum, that is $d_{j(i)}$ in Eq. (6). However our model is more efficient to prevent the problem of over-scored phrase weight, because our model subtracts the weight of single terms from the weight of phrases. So we can overcome the well-known anomaly that any document containing the phrase is likely to be scored too highly in the system which uses the same weighting function for phrase and single term [7].

4.2 Using no relevance information

In the probabilistic retrieval model, if the relevant information is not available, we can generally assume that $R<<N$. Therefore, we assume the following approximation:

$$
\Pr(d \mid \overline{rel}) \approx \Pr(d), \; q_i \approx \frac{n_i}{N}
$$

where N is the size of (number of documents in) the collection and n_i is the number of documents in which the term occurs.

We also assume that $p_i \propto q_i$, especially $p_i = 1/(2-q_i)$ [15]. Then we can write the coefficient of d_i in Eq. (6) as follows:

$$
\log \frac{p_i(1-q_i)}{q_i(1-p_i)} = \log \frac{1}{q_i} = \log \frac{N}{n_i}
$$

that is, the original Spark Jones inverse collection frequency weight.

Furthermore, we assume that $p_{i,j(i)} \propto q_{i,j(i)}$. Then we can approximate the coefficients of $d_{j(i)}$ in Eq. (6) respectively as follows:

$$\log\frac{p_{j(i)}-p_{i,j(i)}}{p_{j(i)}(1-p_i)}-\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}$$
$$\approx c\cdot\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}-\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}=k_7\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}$$

where k_7 is an unknown constant parameter.

We can also approximate the coefficient of $d_i d_{j(i)}$ in Eq. (6) as follows:

$$\log\frac{p_{i,j(i)}(1-q_{i,j(i)})}{q_{i,j(i)}(1-p_{i,j(i)})}-\log\frac{p_i(1-q_i)}{q_i(1-p_i)}-\log\frac{p_{j(i)}(1-q_{j(i)})}{q_{j(i)}(1-p_{j(i)})}$$
$$\approx k_8\left(\log\frac{1}{q_{i,j(i)}}-\log\frac{1}{q_i}-\log\frac{1}{q_{j(i)}}\right)=k_8\log\frac{q_i q_{j(i)}}{q_{i,j(i)}}$$

where k_8 is another unknown constant parameter.

From the above assumption and approximation, we can adapt the Chow Expansion into the probabilistic retrieval model without relevance information as follows:

$$\log\frac{\Pr(d\mid rel)}{\Pr(d\mid \overline{rel})}$$
$$\approx\sum_{i=1}^{n}d_i\log\frac{1}{q_i}+k_7\sum_{i=2}^{n}d_{j(i)}\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}+k_8\sum_{i=2}^{n}d_i d_{j(i)}\log\frac{q_i q_{j(i)}}{q_{i,j(i)}}+\text{constant} \quad (7)$$

We define $MS_{prob\text{-}Chow(N)}$, a query-document scoring function which adapts the Chow Expansion to the probabilistic retrieval model without relevance information as follows:

$$MS_{prob-Chow(N)}(d)$$
$$=\sum_{i=1}^{n}d_i\log\frac{1}{q_i}+k_7\sum_{i=2}^{n}d_{j(i)}\log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}+k_8\sum_{i=2}^{n}d_i d_{j(i)}\log\frac{q_i q_{j(i)}}{q_{i,j(i)}} \quad (8)$$

where k_7 and k_8 are constant parameters.

From Eq. (8), we can define the weight $w^{(CE)}$ for a term i as follows:

$$w^{(CE)}=\begin{cases}d_i\log\dfrac{N}{n_i} & \text{if } i=1\\ d_i\log\dfrac{N}{n_i}+d_{j(i)}k_7\log\dfrac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)}+d_i d_{j(i)}k_8\log\dfrac{q_i q_{j(i)}}{q_{i,j(i)}} & \text{otherwise}\end{cases} \quad (9)$$

4.3 Extending to the 2-Poisson Model

Our method can be extended to incorporate the term dependence into the state-of-the-art 2-Poisson model [13], in particular Okapi BM25 [12], using the Chow Expansion.

The weight of a term *t* in a 2-Poisson model is represented as in the following Eq. (10) [13]:

$$w = \log \frac{(p'+(1-p')(\mu/\lambda)^{tf} e^{\lambda-\mu})(q' e^{\mu-\lambda} + (1-q'))}{(q'+(1-q')(\mu/\lambda)^{tf} e^{\lambda-\mu})(p' e^{\mu-\lambda} + (1-p'))} \tag{10}$$

where *tf* is the frequency of term *t*, λ and μ are the Poisson means for *tf* in the elite and non-elite sets for *t* respectively, *p'* = *P* (elite document for *t* | *rel*), and *q'* is the same probability for non-relevant document. Normally, μ is smaller than λ, so as *tf* → ∞ (to give the asymptotic maximum), $(\mu/\lambda)^{tf}$ goes to zero. Then we can safely remove the components, and $e^{\mu-\lambda}$ will be small, so the approximation will be:

$$w \approx \log \frac{p'(1-q')}{q'(1-p')} \tag{11}$$

Since we cannot estimate (10) directly, the alternative is $w^{(CE)}$ in Eq. (9).

From the above results, Eq. (10) can be transformed into a simple formulation, such as BM25 [12], as given below.

$$w = qtf \frac{tf}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf} w^{(CE)} \tag{12}$$

where *qtf* is query term frequency, *tf* is the frequency of term *t*, *dl* is document length, *avdl* is average length of documents, and k_1 and *b* are constant parameters.

We define $MS_{BM25\ -Chow1(N)}$, a query-document scoring function which adapts the Chow Expansion to the 2-Poisson model without relevance information as follows:

$$MS_{BM25-Chow1(N)}(d) = \sum_i qtf_i \frac{tf_i}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_i} \left(d_i \log\frac{N}{n_i} + d_{j(i)} k_7 \log\frac{q_{j(i)} - q_{i,j(i)}}{q_{j(i)}(1-q_i)} + d_i d_{j(i)} k_8 \log\frac{q_i q_{j(i)}}{q_{i,j(i)}} \right) \tag{13}$$

where k_1, *b*, k_7 and k_8 are constant parameters.

This model can not reflect the term frequency of phrasal term which is consisted of term i and $j(i)$. So we change the model as follows:

$$MS_{BM25\text{-}Chow2(N)}(d) = \sum_i qtf_i \frac{tf_i}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_i}\left(d_i \log\frac{N}{n_i} + d_{j(i)}k_7 \log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)} + d_i d_{j(i)}k_8 \frac{tf_{i,j(i)}}{k_2(1-b+b\cdot\frac{dl}{avdl})+tf_{i,j(i)}} \log\frac{q_i q_{j(i)}}{q_{i,j(i)}} \right) \quad (14)$$

where tf_i is the frequency of term i, $tf_{i,j(i)}$ is the frequency of term i and $j(i)$, and k_1, b, k_2, k_7 and k_8 are constant parameters.

We define another model which separates the phrase terms from single terms as follows:

$$\begin{aligned} MS_{BM25\text{-}Chow3(N)}(d) &= \sum_{i=1}^{n} qtf_i \frac{tf_i}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_i} \log\frac{N}{n_i} \\ &+ k_7 \sum_{i=2}^{n} qtf_i \frac{tf_{j(i)}}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_{j(i)}} \log\frac{q_{j(i)}-q_{i,j(i)}}{q_{j(i)}(1-q_i)} \\ &+ k_8 \sum_{i=2}^{n} qtf_i \frac{tf_{i,j(i)}}{k_1(1-b+b\cdot\frac{dl}{avdl})+tf_{i,j(i)}} \log\frac{q_i q_{j(i)}}{q_{i,j(i)}} \end{aligned} \quad (15)$$

5. EXPERIMENTS AND PERFORMANCE EVALUATION

In this section, we will empirically demonstrate that the 2-Poisson model incorporating the Chow Expansion and dependency parse tree term dependences (Eq. 13, 14, 15) actually gives a significantly better performance than the 2-Poisson model under the conventional linked dependence assumption.

5.1 Test collection

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 [26]. It is published in six volumes with 500 pages per volume. The text data contains 23113 entries, and its size is about 10 mega-bytes. The content of each entry

describes the concept with other entries or more fundamental words. The test set contains 46 natural language queries and the relevance information of the entry lists related to each query. The average document length of the test set is 56 words. The average number of relevant documents of the test set is 9.

5.2 Experiments and results

The goal of the experiments is to validate the proposed model. We took the 46 ETRI-KEMONG queries as originally written and retrieved the 1000 top-ranked documents.

Figure (2) shows an example of the structural indexing system. Through the dependency structure analysis of the documents, we extract single terms and phrase terms for indexing. In retrieving, keyword extraction from the query is performed in the same manner as the indexing process.

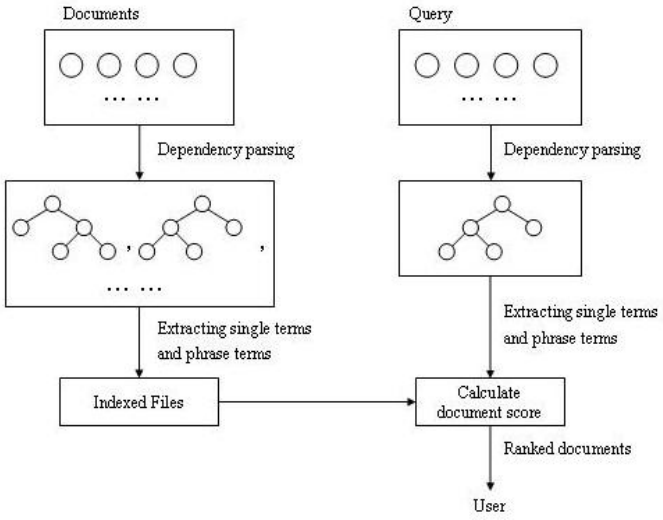

Figure -2. The structural indexing system.

The followings are brief descriptions of the three methods tested and compared.

- BM25: Okapi BM25 model;
- BM25-Chow1: Proposed method (Eq. 13) using the Chow Expansion and dependency parse trees;

The performance was assessed by 'trec_eval', the standard evaluation program of SMART project [25]. The performance measures used in the result table are average precision (non-interpolated) over all relevant

documents (AvgP), Precision at 5, 10, 20 documents (P@5, P@10, P@20, respectively) and R-Precision (R-P).

The results of our experiments for all the cases are listed in the Table 1.

Table -1. **Performace on the ETRI-KEMONG test collection**.

Cases	AvgP	P@5	P@10	P@20	R-P
BM25	0.4137	0.3304	0.2261	0.1457	0.3629
BM25-Chow1	**0.4238**	**0.3455**	**0.2341**	**0.1511**	**0.3707**

The proposed method achieved about a 4.57% improvement for BM25 in the average precision. This result is modestly encouraging.

No experiments have yet been conducted with other models in chapter 4 (Eq. 14 and 15).

6. CONCLUSION AND FUTURE WORK

We presented a new method of incorporating the term dependences in probabilistic retrieval model to compensate the weakness of the linked dependence assumption. We also presented a new weight function for dependency-based phrase terms using the Chow Expansion and dependency parse tree as applied to BM25 function, the widely used 2-Poisson model, for term dependence generation. Co-occurrence between the two terms was obtained from the dependency parse tree.

We carried out some experiments to verity the proposed models. The experiments were performed on the ETRI-KEMONG test collection of Korean to make observations of practicality and usefulness. From the results, an improvement of the performance was obtained on the document collections by incorporating the term dependences using the Chow Expansion and dependency structural indexing system.

We can conclude that incorporating term dependences using the Chow Expansion and structural indexing system into a 2-Poisson model is a viable and appropriate technique to overcome the weakness of the linked dependence assumption model.

The greatest 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 Chow Expansion to Auto Relevance Feedback (ARF). Many researches on query expansions using ARF have verified a significant performance improvement. But it is not yet known whether the Chow Expansion techniques indeed work well on the

ARF expanded queries or not, and this question is another interest point to the 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. Croft W.B. and Harper D.J. Using Probabilistic Models of Document Retrieval Without Relevance Information. *Journal of Documentation*, 35(4):285-295, 1979.

4. Luhn, H. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159-165. 1958.

5. Paul B. Kantor. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology: Research and Developments*, 3(2):88-94, 1984.

6. 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.

7. 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.

8. 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.

9. 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.

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

11. Rorbert M. Losee. Term Dependence: A Basis for Luhn and Zipf Models. *Journal of the American Society for Information Science and Technology*, 52(12):1019-1025, 2001.

12. S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford. Okapi at TREC-3. *In Overview of the Third Text Retrieval Conference(TREC-3).* 109-126, 1995.

13. 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.

14. S.E. Robertson and Sparck Jones K. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3), 129-146, 1976.

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

16. 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.

17. W. Bruce Croft, David D. Lewis. An Approach to Natural Language Processing for Document Retrieval. *In Proceedings of the ACM SIGIR'87*, 26-32, 1987.

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

19. William S. Cooper and P.Huizinga. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Developments*, 1(2):99-112, 1982.

20. Zipf, G.K. Human Behavior and the Principle of Least Effort. *Addision-Wesley, Reading, Mass*. 1949.

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

22. A. Bookstein, D. R. Swanson. A decision theoretic foundation for indexing. *Journal of the American Society for Information Science*, vol 26:45-50, 1975.

23. William M. Shaw, Jr., Judith B. Wood, Robert E. Wood, and Helen R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information science Research*, 13:347-366, 1991.

24. Robert M. Losee. An analytic measure predicting information retrieval system performance. Information Processing and Management, 27(1):1-13, 1991.

25. TREC Eval. The `trec eval' program is available via ftp from the SMART site at Cornell University, ftp://ftp.cs.cornell.edu/pub/smart/. 1992.

26. Kemong. *The Kemong Company new encyclopedia.* Seoul: Kemongsa Publishing Co. 1992.

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

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

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