Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A novel dependency language model for information retrieval*

CAI Ke-ke†, BU Jia-jun†‡, CHEN Chun, QIU Guang

(*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

†E-mail: caikeke@zju.edu.cn; bjj@zju.edu.cn

Received Nov. 10, 2006; revision accepted Mar. 6, 2007

Abstract: This paper explores the application of term dependency in information retrieval (IR) and proposes a novel dependency retrieval model. This retrieval model suggests an extension to the existing language modeling (LM) approach to IR by introducing dependency models for both query and document. Relevance between document and query is then evaluated by reference to the Kullback-Leibler divergence between their dependency models. This paper introduces a novel hybrid dependency structure, which allows integration of various forms of dependency within a single framework. A pseudo relevance feedback based method is also introduced for constructing query dependency model. The basic idea is to use query-relevant top-ranking sentences extracted from the top documents at retrieval time as the augmented representation of query, from which the relationships between query terms are identified. A Markov Random Field (MRF) based approach is presented to ensure the relevance of the extracted sentences, which utilizes the association features between query terms within a sentence to evaluate the relevance of each sentence. This dependency retrieval model was compared with other traditional retrieval models. Experiments indicated that it produces significant improvements in retrieval effectiveness.

Key words: Term dependency, Language modeling (LM), Retrieval model, Sentence retrieval
doi:10.1631/jzus.2007.A0871 **Document code:** A **CLC number:** TP391

INTRODUCTION

Language modeling (LM) is a new retrieval approach that has been used in many recent information retrieval (IR) studies. Experiments have proved its promising retrieval effectiveness compared with traditional retrieval models. However, most existing LM approaches usually assume independence between terms, i.e., terms are statistically independent from each other. Although this assumption makes the retrieval models easier to be implemented, it is not the truth in actual textual data.

Term dependencies exist when the relationships between terms are such that the presence/absence of one term provides information about the probability of presence/absence of another term (Losee, 1994). The term "dependency" can be understood from two perspectives: (1) dependency between terms within a query or within a document; (2) dependency between query terms and document terms (Cao *et al.*, 2005). This paper will concentrate on the first kind of dependency. Retrieval system incorporated with this type of term dependency evaluates not only consistency between terms in query and document but also the consistency between term dependencies in query and documents. It is believed that such approaches will bring significant improvement in retrieval effectiveness.

Some researches have been carried out to study term dependency in LM approach for IR. To realize the effect of term dependency in the practice of LM retrieval approach, two problems should be taken into account primarily: how to define the term dependency and how to apply term dependency into retrieval.

(1) Different descriptions are found in many studies of term dependency, e.g., (Song and Croft, 1999; Nallapati and Allan, 2003; Alvarez *et al.*, 2004; Gao *et al.*, 2004; Lee *et al.*, 2006). A question one

‡ Corresponding author

* Project (No. 2006CB303000) supported in part by the National Basic Research Program (973) of China

may raise is: can we integrate the various forms of dependency together in IR architecture?

(2) In most LM retrieval approaches, only document model is considered to describe information distribution associated with each document. It is believed that some important information concerning query, such as term dependencies in query, also needs to be considered. Query model, which is used to describe information probability distribution associated with the user's information need, provides a natural and intuitive means of encoding query term dependencies. A question then arises that if it is better to introduce query model into retrieval.

To solve the two problems above, which outline the main tasks of this paper, a novel dependency retrieval approach is proposed. This retrieval approach extends the Kullback-Leibler divergence (KL-divergence) based approach to IR (Lafferty and Zhai, 2001) by introducing a hybrid dependency model for both query and document. Consequently, the query model associated with each query consists of two kinds of models, i.e., query term model and query dependency model, which are used to describe term distribution and term dependency distribution in query, respectively. Similarly, document model associated with each document includes document term model and document dependency model. The relevance status of each document is then evaluated by synthetically considering the KL-divergence (Song and Croft, 1999) between term models of query and document, and the term dependency models of query and document.

The expanded retrieval approach introduces a hybrid dependency structure to describe relationships between terms within a sentence. This hybrid dependency structure considers various forms of dependency. It guarantees not only the comprehensiveness but also accuracy of the identified dependencies.

A pseudo relevance feedback based method is proposed to construct query dependency model. The basic idea is to assign each query a sentence-based context description by using query-relevant top-ranking sentences extracted from the top documents at retrieval time. Dependencies between query terms are then statistically identified from the context. To realize the identification of query relevant sentences, a Markov Random Field (MRF) (Dobrushin, 1968) based approach is employed. Relevance of each sentence is measured by the association features between query terms in the sentence.

This paper aims to construct an efficient retrieval model, which can maximize the efforts of various dependencies between terms to bring significant improvement in retrieval effectiveness. An intensive investigation of related researches is made and then a feasible solution is presented. The contributions of our work include:

(1) Propose a novel hybrid dependency structure for describing the dependencies between terms. The hybrid dependency structure allows the integration of various forms of term dependencies and therefore provides a real and comprehensive depiction of the dependencies between terms.

(2) Extend the KL-divergence retrieval model by introducing the dependency models for both query and document. The relevance between query and document is then evaluated by considering the divergence of their dependency models.

(3) Propose a pseudo relevance feedback based approach for constructing the dependency model for query.

(4) Propose an MRF based approach for identifying query relevant sentences.

The rest of this paper is organized as follows. Section 2 introduces the related LM approaches to IR and reviews various dependence models investigated in previous research. Section 3 presents our proposed dependence model and the methods of parameter estimation. In Section 4, a set of experiments is presented. Experimental results showed that our model achieves significant improvements compared with previous retrieval models. Section 5 concludes the paper.

RELATED WORK

Basic LM retrieval models

The first LM approach to IR is proposed by Ponte and Croft (1998). Its basic idea is to compute the conditional probability of generating a query Q given the LM that is trained for each document D. Documents are then sorted in decreasing order of this conditional probability. In this approach, documents will be ranked according to the conditional probability of their LM. Assuming terms in the query are independent, the general unigram model is formulated as

$$P(Q\,|\,D)=\prod_{q_i\in Q}P(q_i\,|\,D)\,. \qquad (1)$$

Recently, another popular retrieval model based on KL-divergence is widely proposed. In this retrieval model, language models are associated with both query and document. The closeness of these models is taken as the evidence of the document's relevance to the given query (Lafferty and Zhai, 2001; Zhai and Lafferty, 2001a). More specifically, the divergence between document and query is determined by

$$D(\theta_Q\,||\,\theta_D)=\sum_{w\in V}P(w\,|\,\theta_Q)\log\frac{P(w\,|\,\theta_Q)}{P(w\,|\,\theta_D)}, \qquad (2)$$

where V is a vocabulary, θ_Q and θ_D are language models for query Q and document D respectively. Since $D(\theta_Q||\theta_D)$ is always non-negative and equals zero if and only if θ_Q=θ_D, the relevance score of D with respect to Q can be measured by $-D(\theta_Q||\theta_D)$.

In this paper, the proposed dependency retrieval model estimates the underlying dependency models for both query and document by exploiting all the valued dependency information between terms. KL-divergence retrieval model is obviously the most proper retrieval model for our task. Query model in KL-divergence allows the description of various forms of term dependency within query and therefore makes it more convenient to integrate information of term dependency into retrieval.

Studies of term dependency

Studies in IR often assume that terms in both query and document are statistically independent. This assumption is obviously improper but widely accepted for its robust estimation and ease of computation (Srikanth and Srihari, 2003). With the explosion of Internet information, it is difficult for these traditional retrieval approaches to provide users the most exact information required. Recent studies showed that term dependency, which can efficiently describe the implicit or explicit relationships between terms, provides the important hints for document relevance. Many researches have been dedicated to this field, hoping to utilize the term dependency information to produce more improvements in retrieval effectiveness.

In traditional probabilistic retrieval models, phrase is one of the most popular forms of term dependency. According to different definitions, two kinds of phrases have been popularly discussed. They are statistical phrases (van Rijsbergen, 1977; 1979; Fagan, 1987) and syntactic phrases (Dillon and Gray, 1983; Fagan, 1987; Smeaton and van Rijsbergen, 1988). Statistical phrase considers the information about the co-occurrences of words in a document. Comparatively, identification of syntactic phrase requires the satisfaction of certain syntactic relationships among the component words. Many studies have evaluated these forms of phrases in traditional retrieval models. Fagan (1987) showed that significant improvements can be achieved by using statistical phrases. However, it is not the case for syntactic phrases. Croft *et al.*(1991) proved the effectiveness of phrases modeled as terms co-occurring in a document. Their experiments also showed that other forms of statistical phrases, which are identified by considering either term proximity or frequency of individual terms and phrase, cannot perform as well as expected. For syntactic phrases, Croft *et al.*(1991) showed that although the efficiency of these syntactic phrases is not significant, better filtering and parsing techniques would facilitate the application of syntactic phrases.

Most of the traditional dependence models did not produce consistent improvements when applied to large-scale document set. There are mainly two reasons for this. Firstly, term dependencies are usually identified on a large scale of documents. It generates a large number of dependency candidates, many of which, however, do not describe term associations exactly. The noise information about term dependencies heavily degrades retrieval performance. Secondly, there is no uniform strategy, by which term dependencies can be incorporated into the retrieval models perfectly (Spark Jones *et al.*, 1998).

Term dependency also has been examined in LM approaches to IR. Early works on term dependencies in LM retrieval models tried to capture term dependency by bi-grams (Song and Croft, 1999) or tri-grams model (Katz, 1987). These approaches assume the dependency between adjacent terms. As a consequence, terms that appear contiguously are considered related to each other. Bi-terms model (Srikanth and Srihari, 2002) expands traditional bi-gram model and considers the dependency between adjacent terms in any order. Although the approaches above can capture certain dependencies between terms, they present marginal improvements in retrieval performance. The reason is twofold: (1) Too "strict". Dependencies between terms that do not satisfy the adjacent constraints are ignored; (2) Too "coarse". Dependencies between any adjacent terms are assumed without any filtering. These limitations heavily weaken the retrieval performance.

Instead of defining term dependencies simply according to the positional information between terms, many statistical and linguistic techniques are applied to LM retrieval models to retain the most probable dependencies between terms. Nallapati and Allan (2002) represented term dependencies in a sentence using a maximum spanning tree. The dependencies between terms in a sentence are measured by using the Jaccard coefficient. Based on the analysis of the most probable linkages between terms in each sentence of the training data, term dependencies within a query are recognized to construct the linkage structure of the query (Gao *et al*., 2004). Documents with the capability to generate similar term linkage are considered relevant. In (Alvarez *et al*., 2004), term dependency was modeled through "word pair". Based on statistical analysis, any pair of words co-occurring within five words is considered to be a word pair. In addition to statistical approaches, syntactic analysis is also applied to determine the dependencies existing in the natural language query. In (Lee *et al*., 2006), term dependencies were set up between terms with syntactic dependency relationships. Experiments affirmed the achievements of the LM approaches above. One of the common features of these approaches is that they assume the dependency structure on the scale of a sentence since most significant dependencies between terms always occur within a sentence (Nallapati and Allan, 2003). This paper makes a similar assumption and explores dependencies between terms within a sentence. Considering the significance of the identified dependencies, better retrieval performance is expected.

EXPANDED RETRIEVAL MODEL

Traditional KL-divergence retrieval models measure the distance between the probabilistic distributions established for query and document over the same set of terms (Lafferty and Zhai, 2001). This paper extends KL-divergence based approach to IR and evaluates the relevance of document to a given query through two components.

The first component is to measure the similarity between query and document from the perspective of term distributions. Let $\theta_{Q,\mathrm{T}}$ and $\theta_{D,\mathrm{T}}$ be the term models for query Q and document D respectively, the relevance score between Q and D can be measured by

$$score_{w}(D,Q)=\sum_{w_i\in Q}P(w_i\,|\,\theta_{Q,\mathrm{T}})\log P(w_i\,|\,\theta_{D,\mathrm{T}})-\sum_{w_i\in Q}P(w_i\,|\,\theta_{Q,\mathrm{T}})\log P(w_i\,|\,\theta_{Q,\mathrm{T}}). \tag{3}$$

Since the second item on the right side of Eq.(3) is identical for all documents, it can be ignored in the computing.

The second component is to measure the relevance between document and query according to the underlying dependency models of query and document. The dependencies detected in the query are also expected to be detected in the relevant document. Let $\theta_{Q,\mathrm{P}}$ and $\theta_{D,\mathrm{P}}$ be the dependency models of query Q and document D respectively, the relevance score between Q and D can be formulated as

$$score_{\mathrm{DP}}(D,Q)=\sum_{w_i,w_j\in Q}P(w_i,w_j\,|\,\theta_{Q,\mathrm{P}})\log P(w_i,w_j\,|\,\theta_{D,\mathrm{P}})-\sum_{w_i,w_j\in Q}P(w_i,w_j\,|\,\theta_{Q,\mathrm{P}})\log P(w_i,w_j\,|\,\theta_{Q,\mathrm{P}}). \tag{4}$$

Similarly, the second item is also document independent and can be dropped.

The above two components observe the relevance between query and document from different perspectives. It is beneficial to include them into a unified retrieval framework. From a viewpoint of effectiveness, the simplest strategy, i.e. linear interpolation method is adopted. Thus, the expanded KL-divergence retrieval model is formulated as Eq.(5), in which the coefficient λ is to control the influence of each component.

$$score(D,Q)=(1-\lambda)\cdot score_{w}(D,Q)+\lambda\cdot score_{\mathrm{DP}}(D,Q). \tag{5}$$

Hybrid dependency structure

This paper attempts to incorporate the most significant term dependencies within a sentence in the retrieval model. Term dependencies within a sentence are various and can be described from different perspectives, such as direct or indirect syntactic relation, proximity relation. In this paper, we define term dependency by using a more intuitive way. According to the association degree between terms in a sentence, three forms of dependency are formed: syntax based, proximity based and co-occurrence based.

Syntax based dependency describes the direct syntactic relation between terms; proximity based dependency describes the relationship between terms within a certain syntactic distance; co-occurrence based dependency relaxes the requirement of syntactic relationship between terms and defines the dependency by using co-occurrence information of terms within a sentences.

Incorporating the above three kinds of dependency into the retrieval, Eq.(4) is then reformulated as

$$score_{\mathrm{DP}}(D,Q)=\sum_{R_k}\alpha_k\left[\sum_{w_i,w_j:R_k}P(w_i,w_j\mid\theta_{Q,\mathrm{P}})\log P(w_i,w_j\mid\theta_{D,\mathrm{P}})\right],\quad(6)$$

where R_k represents the kth form of dependency and the coefficient α_k is to control the influence of each form of dependency to relevance judgment, $k\in[1, 3]$.

Dependency parse tree

As mentioned above, syntax based and proximity based dependencies are constructed based on the syntactic relationship between terms. In this paper, a special type of syntactic relationship, namely dependency syntactic relation, is concerned. A dependency syntactic relation is an asymmetric relationship between a word called governor and another word called modifier (Hays, 1964). The dependency syntactic relations linguistically represent the relations between terms and therefore can help to capture the underlying semantics of a query. In this paper, we use MINIPAR (Lin, 1994) as the dependency parser and correspondingly describe the syntactic relationships with dependency tree. An example of dependency parse tree parsed by MINIPAR is shown in Fig.1, in which nodes are labelled by part of speeches and edges are labelled by relation types. As shown in Fig.1, any dependency tree has a unique root node and any other node has a unique parent node.

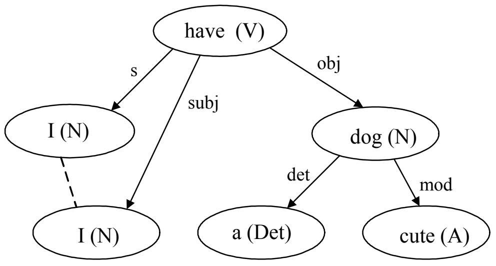

Fig.1 Dependency parse tree of sentence "I have a cute dog"

Before proceeding, let us give some definitions on dependency parse tree.

Definition 1 The distance between two nodes A and B is denoted by $dist(A, B)$ and defined as the number of linkages between A and B in the tree with no consideration of the direction.

Definition 2 The height of node A is denoted by $height(A)$ and defined as the distance between A and the root node.

Definition 3 A set of nodes NS is assumed to be syntactically connected, if they satisfy the following conditions: Among NS, (1) there exists one node, which is the common ancestor node of other nodes; (2) except for the common ancestor node, each node is the child of one other node.

Definition 4 A set of nodes is assumed to be syntactic proximity nodes if the distance between any two nodes in the set is no more than the predefined threshold ζ.

Construction of query dependency model

Query dependency model reflects the dependency information between query terms. However, it is clear that the query itself is always too simple to provide sufficient information on term dependency. To solve this problem, the most intuitive idea is to assign query with some additional descriptive information. Pseudo relevance feedback method (Buckley *et al.*, 1995; Robertson *et al.*, 1995) seems to be a good choice, which facilitates the feedback process and identifies the top-ranking sentences extracted from the top documents at retrieval time as query relevant sentences. To achieve it, an MRF based sentence retrieval approach is proposed. The main objective is to extract the most potentially useful sentences from the top-ranked documents initially retrieved for a given query. Here, the traditional

KL-divergence based approach is chosen for the initial document retrieval.

Sentence relevance estimate based on MRF

MRF is an indirect graphical model and often used in the statistical machine learning domain to model joint distributions. It is constructed based on a graph *G*, in which the nodes represent random variables and the edges present the dependence between the random variables.

Different edge configurations split random variables of *G* into different set of cliques. All pairs of random variables in a clique are neighbors of each other. Neighborhood means the direct dependence between variables. A random object *X* is an MRF if it satisfies

$$P(x_i \mid x_j, \forall j, j \neq i) = P(x_i \mid x_j, \forall j, j \in N(i)), \quad (7)$$

where x_i is the value of the *i*th random variable and $N(i)$ means the neighbors of the *i*th random variable.

Given an MRF, the dependencies among the random variables are determined. In our case, *G* consists of a set of query term variables and a sentence variable. Accordingly, the MRF randomly decides the dependencies among variables of query terms and sentence. By referring to (Metzler and Croft, 2005)'s definition of joint distribution over random variables in *G*, the relevance estimate of a sentence *S* given a query *Q* can be formally defined as

$$P(S \mid Q) \overset{\text{rank}}{=} \sum_{c \in C(G)} \lambda_c f(c), \quad (8)$$

where $C(G)$ is a set of cliques in *G*, $f(c)$ is defined as the real-valued feature function over cliques, λ_c is the weight of each particular feature function. As shown by Eq.(8), two factors are crucial for the relevance estimate: the configuration of *G*, which determines the model of query term dependencies, and the set of feature functions defined over cliques of *G*.

(1) Configuration. By referring to (Metzler and Croft, 2005), we consider three forms of configurations in this paper, i.e., full independence, sequential dependence and full dependence. These three configurations represent different dependence assumptions of query terms. Full independence configuration assumes the independence between query terms, sequential dependence configuration considers the dependence among contiguous query terms and full dependence configuration assumes that all query terms are dependent with each other [For more details, please refer to (Metzler and Croft, 2005)].

(2) Feature functions. Configurations above provide the structure basis for relevance estimates, but the definition of each feature function determines the accuracy of the estimates. Feature function illustrates the state of variables in each clique. Three feature functions were formulated in (Metzler and Croft, 2005) attempting to measure cliques by capturing term occurrences in the context of the document. However, these functions are not applicable to our work. Because compared with documents, sentences always have considerably fewer words, which may be too short to accurately estimate the probability distributions of words co-occurrence. New potential feature functions should be explored in our case.

As we know, any two terms within a sentence can be described by certain syntactic relationship (direct or indirect). Moreover, different relationships describe different degrees of associations between terms. Given a query, the relevance of each sentence is considered different if query terms contained in it present different forms of syntactic relationships. This paper investigates the syntactic relationships between terms and then defines three potential functions.

(1) The first potential function involves cliques that contain a query term and a sentence. Such cliques assume the independence between query terms. The potential functional is therefore defined as

$$f_1(c) = P(q_i \mid S) = \frac{freq(q_i, S)}{|S|} \sqrt{\sigma}^{height(q_i)}, \quad (9)$$

where $freq(q_i, S)$ is the number of times query term q_i occurs in sentence *S*, $|S|$ is the total number of terms in *S*, parameter $\sigma \in [0, 1]$. As shown in Eq.(9), this function considers sentence relevance from two perspectives: the amount of occurrence of query term in sentence and the importance of each query term to sentence.

(2) Given a clique containing two or more query terms, the fact that these query terms syntactically connected in sentence provides more evidence of the sentence relevance. The syntactic relationship among

terms is defined as Definition 3. The second potential function considers such syntactic relationship and for every clique that contains two or more terms q_u, ..., q_v and the sentence S, the following feature function is applied:

$$\begin{aligned} f_{\mathrm{SR}}(c) &= P(q_u,...,q_v \mid S) \\ &= BoolSynR(q_u,...,q_v)\sqrt{\omega}^{d}\sqrt{\sigma}^{h'}, \end{aligned} \tag{10}$$

where ω, $\sigma\in[0, 1]$; $BoolSynR(q_u, ..., q_v)$ is set to 1 if q_u, ..., q_v are syntactically connected in the dependency parse tree of S, otherwise it is set to 0; d and h' denote the distance and height of terms q_u, ..., q_v in the dependency parse tree of S respectively and are defined as

$$d = \sum_{i,j\in[u,v]} dist(q_i,q_j) \Big/ N, \tag{11}$$

$$h' = \min_i(height(q_i)), \quad i \in [u,v], \tag{12}$$

where N is the number of term pairs of q_u, ..., q_v.

(3) Although the occurrence of the syntactically connected query terms provides strong evidence of relevance, the occurrence of syntactic proximity query terms can also provide valuable evidence. The syntactic proximity relationship among terms is defined as Definition 4. Such kind of relationship is considered in this paper and for every clique that contains two or more terms q_u, ..., q_v and the sentence S, the third feature function is defined as

$$\begin{aligned} f_{\mathrm{PR}}(c) &= P(q_u,...,q_v \mid S) \\ &= BoolProxR(q_u,...,q_v)\sqrt{\omega}^{d}\sqrt{\sigma}^{h''}, \end{aligned} \tag{13}$$

where $BoolProxR(q_u, ..., q_v)$ is set to 1 if q_u, ..., q_v are proximity terms in the dependency parse tree of S, otherwise it is set to 0; d, ω and σ are defined similarly as those in Eq.(10); h'' also represents the depth of q_u, …, q_v in the dependency parse tree of S, but is defined as

$$h'' = \frac{1}{v-u+1}\sum_{i=u}^{v} height(q_i). \tag{14}$$

Using the potential functions above we derive the following ranking function for sentence retrieval:

$$\begin{aligned} P(S \mid Q) &\overset{\text{rank}}{=} \sum_{c\in C(G)} \lambda_{\mathrm{c}} f(c) = \sum_{c\in I} \lambda_{\mathrm{I}} f_{\mathrm{I}}(c) + \\ &\sum_{c\in CD\cup UD} \lambda_{\mathrm{SR}} f_{\mathrm{SR}}(c) + \sum_{c\in CD\cup UD} \lambda_{\mathrm{PR}} f_{\mathrm{PR}}(c), \end{aligned} \tag{15}$$

where, I is a set of cliques containing a query term and a sentence; CD represents a set of cliques involving a sentence and multi-query terms that appear contiguously within the query; UD denotes a set of cliques, with the terms appearing non-contiguously with each other in the query; λ_{I}, λ_{SR} and λ_{PR} are valued between 0 and 1, and are used to control the influence of each feature function on the relevance estimate.

Query dependency model

Based on the process above, top-ranking sentences relevant to query can be identified by using pseudo relevance feedback approach. Dependency probabilities between query terms are then generalized by statistical method. In this paper, the backoff scheme (Gao *et al.*, 2004) is adopted. The probability of w_i, w_j being generated from $\theta_{Q,\mathrm{P}}$ with the form of R_k is evaluated by

$$\begin{aligned} P(w_i, w_j, R_k \mid \theta_{Q,\mathrm{P}}) &= \lambda_1 E_1(w_i, w_j, R_k) + (1-\lambda_1)\cdot \\ &[\lambda_2 E_2(w_i, w_j, R_k) + (1-\lambda_2)E_3(w_i, w_j, R_k)], \end{aligned} \tag{16}$$

where

$$\lambda_1 = \frac{\delta_1}{\delta_1+1}, \quad \lambda_2 = \frac{\delta_2+\delta_3}{\delta_2+\delta_3+1},$$

$$E_1(w_i, w_j, R_k) = \frac{\eta_1}{\delta_1}, \quad E_2(w_i, w_j, R_k) = \frac{\eta_{2+}\eta_3}{\delta_2+\delta_3},$$

$$E_3(w_i, w_j, R_k) = \eta_4/\delta_4,$$

$$\eta_1 = c(w_i, w_j, R_k), \quad \delta_1 = c(w_i, w_j),$$

$$\eta_2 = c(w_i, *, R_k), \quad \delta_2 = c(w_i, *),$$

$$\eta_3 = c(*, w_j, R_k), \quad \delta_3 = c(*, w_j),$$

$$\eta_4 = c(*, *, R_k), \quad \delta_4 = c(*, *).$$

Here, $c(w_i, w_j, R_k)$ denotes the number of times that the dependency R_k is discovered between w_i and w_j in the feedback sentences; $c(w_i, w_j)$ is the number of times that w_i and w_j appear in the same feedback sentences.

Estimation of other parameters

In Eq.(5), except the distributions of $P(w_i$, w_j,

$R_k|\theta_{Q,\mathrm{P}}$), other three parameters, i.e. $P(w_i|\theta_{Q,\mathrm{T}})$, $P(w_i|\theta_{D,\mathrm{T}})$ and $P(w_i, w_j, R_k|\theta_{D,\mathrm{P}})$ also should be estimated.

(1) $P(w_i, w_j, R_k|\theta_{D,\mathrm{P}})$

The linear interpolation smoothing approach is used to estimate the value of $P(w_i, w_j, R_k|\theta_{D,\mathrm{P}})$. Based on the dependency models of document D and the entire document collection C, $P(w_i, w_j, R_k|\theta_{D,\mathrm{P}})$ is finally defined as

$$P(w_i,\ w_j,\ R_k\mid\theta_{D,\mathrm{P}})=(1-\gamma)\cdot P(w_i,\ w_j,\ R_k\mid D)+\gamma\cdot P(w_i,\ w_j, R_k\mid C),\tag{17}$$

where, $P(w_i, w_j, R_k|D)$ and $P(w_i, w_j, R_k|C)$ can be evaluated by using the same approach of $P(w_i, w_j, R_k|\theta_{Q,\mathrm{P}})$. The only difference is that the computation of $c(w_i, w_j, R_k)$, etc. are respectively implemented in the context of document D and collection C.

(2) $P(w_i|\theta_{Q,\mathrm{T}})$

The generation probability of term from the query Q is evaluated by the maximum likelihood estimate. Formally, $P(w_i|\theta_{Q,\mathrm{T}})$ is evaluated by

$$P(w_i|\theta_{Q,\mathrm{T}})=c(w_i,Q)\Big/\sum_{w_i}c(w_i,Q),\tag{18}$$

where $c(w_i, Q)$ denotes the number of times that w_i occurs in Q.

(3) $P(w_i|\theta_{D,\mathrm{T}})$

Zhai and Lafferty (2001b) showed that the smoothing methods of Jelinek-Mercer and Dirichlet clearly have a better average precision than absolute discounting. We select Dirichlet smoothing method to compute $P(w_i|\theta_{D,\mathrm{T}})$. The model is given by

$$P(w_i\mid\theta_{D,\mathrm{T}})=\frac{c(w_i,D)+\mu P(w_i\mid C)}{u+\sum_{w_i\in D}c(w_i,D)},\tag{19}$$

where μ is the parameter of the Dirichlet distribution. The experiments also demonstrated that the optimal value of μ appears to have a wide range 1500~10000 and is around 2000. $P(w_i|C)$ denotes the probability of w_i occurring in the collection C and is computed by the maximum likelihood estimate:

$$P(w_i\mid C)=c(w_i,C)\Big/\sum_{w_i\in C}c(w_i,C),\tag{20}$$

EXPERIMENTS

This section reports results from experiments we conducted to evaluate the effect of the suggested retrieval models.

Setting

We use TREC disks 4 and 5 for experiments and evaluate our methods on the ad hoc tasks for TREC 6, with topics 301~350, TREC 7 with 351~400, and TREC 8 with 401~450, respectively. Here, only the title portions of these topics are used to construct our experimental queries. Relevance of retrieved documents is assessed by using the relevance assessments provided by NIST for the ad hoc task. All documents and queries are processed as standard; terms are stemmed by using the Porter stemmer and stop words are removed by referring to the stoplist.

In our proposed retrieval model, several parameters have to be determined, such as λ in Eq.(5) and other smoothing parameters. These parameters should be tuned to obtain the best effectiveness. In many systems, such parameters are tuned automatically, e.g., using maximum likelihood and maximum a posteriori estimation (Lo, 1988; Gauvain and Lee, 1994). However, for simplicity, our experiments determine these parameters empirically. This paper will narrate the parameter tuning process of our proposed models in detail and just report the best effectiveness of other retrieval models.

Results

As introduced above, this paper introduces three variants of MRF model for sentence retrieval. Accordingly, the proposed dependency retrieval model is evaluated by using these three variants respectively. In the following description, EKLM_I, EKLM_SD and EKLM_FD represent the proposed retrieval model with different assumptions of query term dependency in sentence retrieval. These assumptions are respectively full independence between query terms, dependence among the contiguous query terms and dependence among any combination of query terms.

To determine the benefits of our proposed retrieval model to document retrieval, a set of experiments are implemented. In the first set of experiments, we compare the proposed dependency retrieval models with three traditional retrieval models, including the TFIDF model (TFIDF), the probabilistic retrieval model OKAPI (OKAPI) and the KL-divergence LM with Dirichlet smoothing (KL_DIR). All these three retrieval models were implemented by the Lemur toolkit (http://www.lemurproject.org) and the result of KL-divergence provided the baseline from which to compare other retrieval models.

Table 1 shows the comparison results implemented on different sets of topics. The best effectiveness of our proposed retrieval approach is reported. The percentage in parentheses indicates the performance change of the current retrieval model compared with the baseline approach. MAP is defined as the non-interpolated average precision averaged over all queries. It reflects the efficiency of each retrieval model by synthetically considering their performances in precision and recall. MRR is defined as the reciprocal of the first relevant document's rank in the ranked list retrieved for a given query. As shown in Table 1, improvements by using EKLM_SD and EKLM_FD over traditional retrieval models are statistically significant. Compared with the KL_DIR model, they provide improvements of 11%~20% in MAP and 4%~21% in MRR. These results showed that EKLM_SD and EKLM_FD can largely enhance retrieval precision and not weaken retrieval recall. As for EKLM_I, its performance is comparable to those of KL_DIR, but the difference is statistically insignificant. Compared with EKLM_SD and ELKM_FD, EKLM_I realizes sentence retrieval with no consideration of term dependency. The lower performance of EKLM_I further proves the importance of term dependency in IR.

Among the three variants of the proposed retrieval model, EKLM_SD performs better than the other two variants. It on one hand makes sure the correctness of the assumption of term dependency and on the other hand shows that dependencies estimated among continuous query terms are more significant than those among any combination of query terms. It is mostly attributed to the fact that our experimental queries, which are taken from the titles of TREC topics, are well formed. Thus, dependencies often occur among those continuous query terms. Comparatively, the dependencies assumed among any combination of query terms introduce lots of noises and influence the accuracy of the retrieved sentence to some extent.

To further validate the effect of term dependency on retrieval, Fig.2 illustrates precision-recall curve of each retrieval model. As shown in Fig.2, in any recall level, EKLM_SD and EKLM_FD perform better than other retrieval models.

In the proposed retrieval approach, several parameters should be determined. Thus, another set of experiments are implemented to tune the value of each parameter to obtain the best retrieval effectiveness. Table 2 illustrates the non-interpolated average precision (MAP) averaged over all queries 301~450 of EKLM_I, EKLM_SD and EKLM_FD with the change of parameters in Eq.(6), i.e., α_1, α_2 and α_3. As can be seen from Table 2, better performance can be achieved when α_1, α_2 and α_3 are respectively set to around 0.1, 0.3 and 0.6. Results showed the dominance of looser relationship between terms in the experimental queries. Low values of α_3 will decrease retrieval performance, because in these cases the "strict" dependencies, i.e., syntax based dependency, proximity based dependency, are main dependency information to be used. Such dependency information, however, are relatively limited because of data sparseness. Similarly, higher values of α_3 also will limit the retrieval performance, because in these cases the effects of "strict" dependencies are totally ignored.

Table 1 MAP and MRR of each retrieval model on different topics*

Model	MAP			MRR		
	301~350	351~400	401~450	301~350	351~400	401~450
TFIDF	0.184 (−23.9%)	0.186 (−34.5%)	0.160 (−25.9%)	0.389 (−25.3%)	0.420 (−24.5%)	0.373 (−20.3%)
OKAPI	0.241 (−0.4%)	0.258 (−9.2%)	0.194 (−10.2%)	0.528 (−0.9%)	0.533 (−4.1%)	0.374 (−20.1%)
KL_DIR	0.242	0.284	0.216	0.533	0.556	0.468
EKLM_I	0.239 (−1.2%)	0.287 (+1.0%)	0.212 (−1.9%)	0.538 (+0.9%)	0.563 (+1.3%)	0.451 (−3.6%)
EKLM_SD	0.288 (+19.0%)	0.332 (+16.9%)	0.261 (+20.8%)	0.646 (+21.2%)	0.597 (+5.5%)	0.528 (+12.8%)
EKLM_FD	0.271 (+12.0%)	0.317 (+11.6%)	0.244 (+12.9%)	0.598 (+12.2%)	0.586(+5.3%)	0.487 (+4.1%)

*: The percentages in parentheses indicate the performance change of the current retrieval model compared with the baseline approach

Our experiments also evaluated how λ, as shown in Eq.(5), influences the effectiveness of the retrieval model. We changed λ value in the series of experiments. Table 3 shows P@k of models EKLM_I, EKLM_SD and EKLM_FD with different values of λ. The metrics P@k for k=5, 10, 20 are defined as the precision at the top k sentences averaged over all queries 301~450. As shown in Table 3, the performance of EKLM_I is optimal when λ is set to 0.9, but for EKLM_SD and EKLM_FD, the performance is best when λ is set between 0.7 and 0.8. It is indicated that term dependencies identified in EKLM_SD and EKLM_FD are more significant than those identified in EKLM_I and therefore should be paid more attention to.

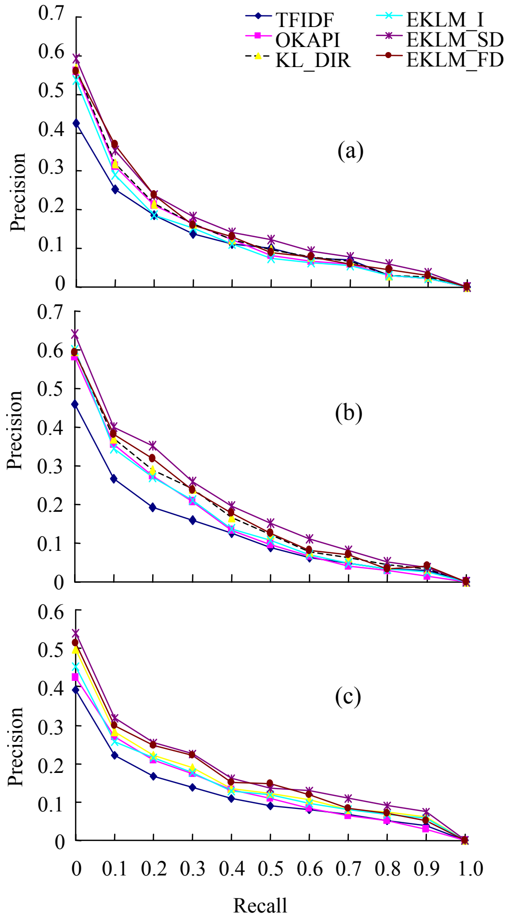

Fig.2 Precision-Recall curve of each retrieval model on the ad hoc task for TREC-6 (a), TREC-7 (b), TREC-8 (c)

Table 2 MAP of EKLM_I, EKLM_SD and EKLM_FD with the change of α_1, α_2 and α_3

α_1, α_2, α_3	EKLM_I	EKLM_SD	EKLM_FD
1, 0, 0	0.237	0.278	0.265
0, 1, 0	0.239	0.280	0.269
0, 0, 1	0.242	0.285	0.273
0, 0.1, 0.9	0.240	0.289	0.275
0.1, 0.1, 0.8	0.240	0.289	0.274
0.1, 0.2, 0.7	0.243	0.291	0.277
0.1, 0.3, 0.6	**0.246**	**0.294**	**0.277**
0.1, 0.4, 0.5	0.243	0.293	0.276
0.1, 0.5, 0.4	0.237	0.293	0.271
0.4, 0.3, 0.3	0.233	0.284	0.269
0.5, 0.3, 0.2	0.236	0.286	0.267
0.6, 0.3, 0.1	0.235	0.281	0.266

The hightlighted values represent better perfermance

As shown in Eq.(15), parameters involved in the ranking function include λ_{I}, λ_{SR} and λ_{PR}. We explored various settings of these parameters to study their impact on retrieval effectiveness. Since EKLM_I dose not use the parameters λ_{SR}, λ_{PR}, its parameter λ_{I} is then set to 1. For EKLM_SD and EKLM_FD, the results showed that the performance is optimal when λ_{I}, λ_{SR}, λ_{PR} are respectively set to 0.7, 0.1, 0.2. The results showed that in the service of sentence retrieval, common terms contained in query and sentence are the most important factors in determining sentence relevance. However, dependency information about terms is also important. They can provide further validations for sentence relevance.

Table 3 Impact of parameter λ

λ	Model	P@5	P@10	P@20
0.9	EKLM_I	0.293	0.240	0.217
	EKLM_SD	0.404	0.374	0.318
	EKLM_FD	0.340	0.346	0.283
0.8	EKLM_I	0.241	0.228	0.178
	EKLM_SD	0.477	0.394	0.351
	EKLM_FD	0.432	0.377	0.303
0.7	EKLM_I	0.233	0.219	0.198
	EKLM_SD	0.449	0.428	0.331
	EKLM_FD	0.416	0.384	0.358
0.5	EKLM_I	0.243	0.227	0.146
	EKLM_SD	0.370	0.316	0.263
	EKLM_FD	0.366	0.323	0.294
0.3	EKLM_I	0.237	0.212	0.125
	EKLM_SD	0.329	0.304	0.260
	EKLM_FD	0.341	0.319	0.265
0.1	EKLM_I	0.203	0.196	0.170
	EKLM_SD	0.317	0.298	0.254
	EKLM_FD	0.324	0.318	0.224

CONCLUSION

In this paper, a novel dependency language modeling approach for information retrieval is proposed. It extends the state-of-the-art KL-divergence approach to IR by introducing dependency models for both query and document. Query dependency model is constructed based on the relevant sentences captured from feedback documents. An MRF based estimate approach is introduced, which utilizes the relationships between query terms in sentences to measure the relevance of sentences for the given query. A hybrid dependency structure is introduced as well, which views term dependency from three perspectives, including syntax based dependency, proximity based dependency, and co-occurrence based dependency. Experiments showed that the proposed retrieval model outperforms the traditional models without consideration of term dependency.

References

Alvarez, C., Langlais, P., Nie, J., 2004. Word Pairs in Language Modeling for Information Retrieval. Proc. 7th International Conference on Computer Assisted Information Retrieval. Avignon, France, p.686-705.

Buckley, C., Salton, G., Allan, J., Singhal, A., 1995. Automatic Query Expansion Using SMART: TREC-3. Proc. 3rd Text Retrieval Conference. Maryland, USA, p.65-80.

Cao, G., Nie, J., Bai, J., 2005. Integrating Word Relationships into Language Models. Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Brazil, p.298-305.

Croft, W.B., Turtle, H.R., Lewis, D.D., 1991. The Use of Phrases and Structured Queries in Information Retrieval. Proc. 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, USA, p.32-45. [doi:10.1145/122860.122864]

Dillon, M., Gray, A.S., 1983. FASIT: a fully automatic syntactically based indexing system. *J. Am. Soc. Inf. Sci.*, **34**(2):99-108.

Dobrushin, P.L., 1968. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and Its Applications*, **13**(2):197-224. [doi:10.1137/1113026]

Fagan, J.L, 1987. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-syntactic Methods. Proc. 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Louisiana, USA, p.91-101. [doi:10.1145/42005.42016]

Gao, J., Nie, J., Wu, G., Cao, G., 2004. Dependence Language Model for Information Retrieval. Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK, p.170-177.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Processing*, **2**(2):291-298. [doi:10.1109/89.279278]

Hays, D.G., 1964. Dependency theory: a formalism and some observations. *Language*, **40**(4):511-525. [doi:10.2307/411934]

Katz, S.M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. on Acoustics, Speech and Signal Processing*, **35**(3):400-401. [doi:10.1109/TASSP.1987.1165125]

Lafferty, J., Zhai, C., 2001. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Louisiana, USA, p.111-119. [doi:10.1145/383952.383970]

Lee, C., Lee, G., Jang, M., 2006. Dependency structure language model for information retrieval. *ETRI*, **28**(3):337-346.

Lin, D., 1994. Principar—An Efficient, Broad-coverage, Principle-based Parser. Proc. 15th International Conference on Computational Linguistics. Kyoto, Japan, p.482-488.

Lo, A.W., 1988. Maximum likelihood estimation of generalized Ito processes with discretely sampled data. *Econ. Theory*, **4**:231-247.

Losee, R.M.Jr, 1994. Term dependence: truncating the Bahadur Lazarsfeld expansion. *Inf. Process. Manage.*, **30**(2):293-303. [doi:10.1016/0306-4573(94)90071-X]

Metzler, D., Croft, W.B., 2005. A Markov Random Field Model for Term Dependencies. Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Brazil, p.472-479. [doi:10.1145/1076034.1076115]

Nallapati, R., Allan, J., 2002. Capturing Term Dependencies Using a Language Model Based on Sentence Trees. Proc. 11th ACM CIKM International Conference on Information and Knowledge Management. Virginia, USA, p.383-390.

Nallapati, R., Allan, J., 2003. An Adaptive Local Dependency Language Model: Relaxing the Naive Bayes' Assumption. Proc. Workshop on Mathematical and Formal Models in Information Retrival, the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada.

Ponte, J.M., Croft, W.B., 1998. A Language Modeling Approach to Information Retrieval. Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, p.275-281. [doi:10.1145/290941.291008]

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M., 1995. Okapi at TREC-3. Proc. 3rd Text Retrieval Conference. Maryland, USA, p.109-216.

Smeaton, A.F., van Rijsbergen, C.J., 1988. Experiments on Incorporating Syntactic Processing of User Queries into a Document Retrieval Strategy. Proc. 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Grenoble, France, p.31-51. [doi:10.1145/62437.62439]

Song, F., Croft, W.B., 1999. A General Language Model for Information Retrieval. Proc. 8th International Conference on Information and Knowledge Management. Missouri, USA, p.316-321.

Spark Jones, K., Walker, S., Robertson, S.E., 1998. A Probabilistic Model of Information Retrieval: Development and Status. Technical Report 446, University of Cambridge Computer Laboratory.

Srikanth, M., Srihari, R., 2002. Biterm Language Models for Document Retrieval. Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information. Tampere, Finland, p.425-426. [doi:10.1145/564376.564476]

Srikanth, M., Srihari, R., 2003. Exploiting Syntactic Structure of Queries in a Language Modeling Approach to IR. Proc. 12th International Conference on Information and Knowledge Management. LA, USA, p.476-483.

van Rijsbergen, C.J., 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *J. Document.*, **33**(2):106-119.

van Rijsbergen, C.J., 1979. Information Retrieval. Butterworths, London.

Zhai, C., Lafferty, J., 2001a. Model-based Feedback in the Language Modeling Approach to Information Retrieval. Proc. 10th ACM CIKM International Conference on Information and Knowledge Management. Atlanta, Georgia, USA, p.403-410.

Zhai, C., Lafferty, J., 2001b. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Louisiana, USA, p.334-342. [doi:10.1145/383952.384019]