*Computational Intelligence*, Volume 59, Number 000, 2016

# Modelling Multiple Interactions with Markov Random Field in Query Expansion for Session Search

JINGFEI LI[1,3], XIAOZHAO ZHAO[1], PENG ZHANG[1,*] AND DAWEI SONG[1,2,*]

[1]*Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, China*

[2]*The Computing and Communications Department, the Open University, UK*

[3]*National Computer Network Emergency Response Technical Team/Coordination Center of China*

How to automatically understand and answer users' questions (e.g., queries issued to a search engine) expressed with natural language has become an important yet difficult problem across the research fields of Information Retrieval (IR) and Artificial Intelligence (AI). In a typical interactive Web search scenario, namely Session Search. To obtain relevant information, user usually interacts with the search engine for several rounds in the forms of, e.g., query reformulations, clicks and skips. These interactions are usually mixed and interwined with each other in a complex way. For the ideal goal, an intelligent search engine can be seen as an AI agent that is able to infer the user's information need from these interactions. However, there still exists a big gap between the current state of the art and this goal. In this paper, in order to bridge the gap, we propose a Markov Random Field based approach to capture dependencies among interactions, queries and clicked documents for automatic query expansion (as a way of inferring the user's information need). Extensive empirical evaluation is conducted on large scale web search datasets, and the results demonstrate the effectiveness of our proposed models.

*Key words:* Session search, query expansion, Markov Random Field, multiple interactions.

## 1. INTRODUCTION

An ideal artificial intelligence system (e.g., one that passes the turing test) is expected to respond to users' questions (or queries) naturally like a real human does. To achieve this ultimate goal involves obtaining, interacting and reasoning about information that is relevant to the questions, thus requiring a synergy across the research fields of information retrieval (IR), artificial intelligence (AI) and human computer interaction (HCI). A large number of IR models are proposed to retrieve relevant information from large scale web or local data repositories. An intelligent IR system can be seen as an AI agent (Guan et al., 2013; Zhang et al., 2013), which allows the user to continuously interact with the system, and is able to automatically infer the user's hidden information need from these interactions, e.g., through query reformulations.

In interactive Web search scenario, users usually interact with the search engine many times in order to accomplish a complex search task. This typical interaction process can be seen as a search session (Guan et al., 2013; Zhang et al., 2013; Huang et al., 2004). In contrary to the traditional Ad Hoc search, session search allows IR models to retrieve documents by utilizing the historical interaction information within the same session. As illustrated in Figure 1, in a user's search session, there exists a sequence of interactions in multiple forms (e.g., query reformulations, clicks and skips), which can be viewed as the user's implicit relevance feedback. For example, the current query reflects the current information need directly (which may be insufficient and need to be refined); clicks may

[1] Corresponding Authors: Dawei Song (dwsong@tju.edu.cn) *and* Peng Zhang (pzhang@tju.edu.cn)

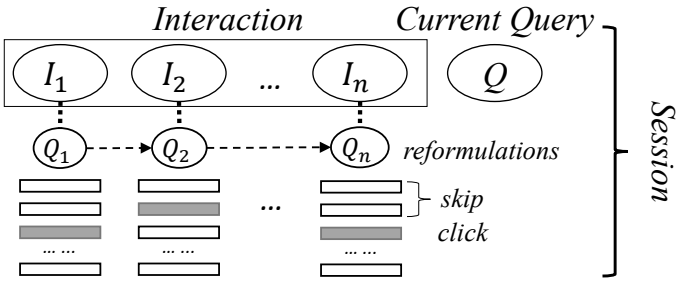


FIGURE 1. Session search allows IR models to retrieve documents by utilizing an entire session, denoted as $S = \{I_1, I_2, ..., I_n, Q\}$, where $I_i$ is a interaction unit which contains a series of behaviors, e.g., skip (irrelevant results), click (relevant results) and query reformulations, etc., and $Q$ is the current query of a session.

imply that users are interested in the viewed content (Li et al., 2014); the skip over some results may be a strong signal reflecting that the corresponding documents are irrelevant; reformulations of queries in a session imply that the user wants to search for some novel topics (reflecting an evolving information need) or enhance the queries' representation power for the current information need; and the whole session reflects the evolution of user's information need in response to the interactions; and so on. It is important to note that these interactions are usually mixed and intertwined with each other in a complex way (e.g., clicking on a document may lead to skipping another document or a query modification), making the session search a difficult task. Thus it is crucial to model and exploit such complex interactions and their interdependencies, in order to predict the user's hidden search intent.

Query expansion, as an important means to represent users' hidden search intent (Zhang et al., 2016; Carpineto and Romano, 2012), expands the user's original query by selecting relevant terms from a series of feedback documents (e.g., through pseudo-relevance feedback that simply assumes the top-ranked documents as relevant, or estimated based on the user interactions as implicit relevance feedback as described above). A representative query expansion method is the Relevance Model (Lavrenko and Croft, 2001), which selects expansion words by considering the likelihood of relevance between original query and feedback documents. Later, a positional relevance model (PRM) (Lv and Zhai, 2010) selects expansion words that are focused on the query topic, based on their positions and proximities to the query terms in feedback documents. Existing query expansion models do not explicitly consider two types of dependency relationships when assigning weights to expansion terms. They are (1) between query terms and multiple interaction behaviors; (2) between feedback documents and interaction behaviors.

In this paper, we propose to utilize the Markov Random Field (MRF) as a unified framework to model the dependencies among different interaction events in a search session, based on which three query expansion models are derived. Different from traditional query expansion models which only consider the relationship between feedback documents and original query, we introduce an additional variable to represent the multiple interactions. Specifically, we explicitly consider three types of information in the process of query expansion. They are the current query $Q$, the feedback documents $D$ (including implicit-relevance feedback documents and pseudo-relevance feedback documents), and the interaction behaviors $I$ (including skips, clicks and reformulations, etc.).

Before building an expanded query model, we first need to obtain a candidate set of words for expansion. For each candidate expansion word, we compute its weight based on the MRF model constructed from the interaction data. A series of feature functions

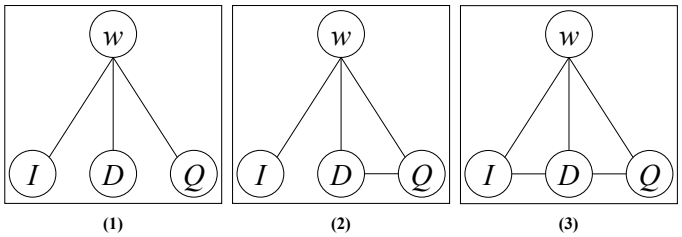


FIGURE 2. Dependency graphs for three weighting functions based on different dependence assumptions between the components of the mixed-feedback. $w$ is a word, $Q$ is current query, $I$ is the interaction behaviors, and $D$ is a feedback document (a clicked document or pseudo-feedback document).

are proposed for the MRF framework, so that different interaction information in search sessions are captured. We systematically investigate three different MRF models that are respectively underpinned by three different dependency assumptions (see Figure 2). The Full Independence Model (FIM) considers that $Q$, $D$ and $I$ determine the weight of an expansion word independently, and assumes that the importance of all feedback documents $D$ is uniform (see Figure 2-(1)). The Query-Document Dependence Model (also called as Partial Dependence Model, PDM) assumes that the selection of feedback documents should be dependent on the original query (Figure 2-(2)). Finally, in the Full Dependence Model (FDM), we assume that the selection of feedback documents are also dependent on the dynamic interaction behaviors (Figure 2-(3)).

After obtaining the expansion terms for each session, we run the expanded Indri[1] query language in the Indri search engine to obtained the final search results. We have conducted extensive empirical evaluation on the Session Track data in TREC (Text REtrieval Conference[2]) 2013 and 2014. The evaluation results demonstrate the effectiveness of the proposed models in comparison with a number of state-of-the-art baselines.

In a nutshell, the main contributions of this paper can be summarized as follows:

- We proposed a framework based on Markov Random Field (MRF) which can model the dependency relationships between interaction behaviors, current query and feedback documents in query expansion for session search;
- We proposed a series of feature functions for the Markov Random Field models, so that diversified interaction information can be captured within a unified framework;
- We conducted extensive comparative experiments on large scale session search benchmarking datasets, and demonstrated the effectiveness of our proposed models.

The rest of this paper is organized as follows. Section 2 reviews the related work. The proposed query expansion models are described in Section 3. Extensive evaluations are conducted in Section 4. In Section 5, we will draw conclusions about this paper and discuss on some possible directions for the future work.

## 2. RELATED WORK

Typical query expansion models can be based on explicit and implicit relevance feedback (Buscher et al., 2008; Chirita et al., 2007; Cui et al., 2002, 2003), pseudo relevance feedback (Bai et al., 2005; Lavrenko and Croft, 2001; Lee et al., 2008; Lv and Zhai, 2010; Rocchio,

[1]https://sourceforge.net/projects/lemur/
[2]http://trec.nist.gov/

1971; White et al., 2002, 2005) and external knowledge (Arguello et al., 2008; Bodner and Song, 1996; Gottipati and Jiang, 2011; Navigli and Velardi, 2003; Pan et al., 2013). In this paper, we focus on implicit and pseudo relevance feedback based query expansion methods.

The implicit-feedback methods can be categorized into two directions: query log based and Human-Computer Interaction (HCI) based. For example, Cui et al. proposed query expansion models based on user interactions recorded in user logs (Cui et al., 2002, 2003). They selected high-quality expansion terms according to the correlations between query terms and document terms extracted by analyzing query logs. Chirita et al. expanded the short Web search queries with terms collected from each user's Personal Information Repository, in order to resolve the ambiguity of the Web search queries and personalize the search results (Chirita et al., 2007). Gao et al. proposed a unified query expansion framework based on query logs using the Path-Constrained Random Walks (Gao et al., 2013). Joachims et al. (Joachims et al., 2007) examines the reliability of implicit feedback generated from clickthrough data and query reformulations in WWW search, and concludes that clicks are informative but biased. In addition to the log-based implicit feedback, there are attempts to utilize HCI information to enhance the query expansion models. For example, Buscher et al. employed eye tracking data to keep track of document parts that the user reads, and then the information on the subdocument level is used as implicit feedback for query expansion and document re-ranking (Buscher et al., 2008). More recently, Chen et al. (Chen et al., 2015) proposed a query expansion model based on the real-time reading content captured by eye tracker.

Pseudo-relevance feedback based query expansion assumes that the top-ranked documents from search engine are relevant. Rocchio proposed a classical query expansion model based on pseudo-relevant documents for the SMART retrieval system (Rocchio, 1971). After that, a series of pseudo-relevance feedback based models emerged. For example, Lavrenko and Croft proposed the well known relevance model (RM) to estimate a language model from feedback documents (Lavrenko and Croft, 2001), which can be used to estimate the weights of expanded terms. RM3, a further variant of RM, interpolates the term weights in RM with that in the original query language model (Lv and Zhai, 2009a). The traditional pseudo-feedback approaches utilize the whole feedback documents to extract words for query expansion, which may contain considerable irrelevant information (Zhang et al., 2009). To solve this problem, retrieval models based on subdocument (e.g., passages (Liu and Croft, 2002)) or term positions (Lv and Zhai, 2010) have been proposed. Similarly, Miao et al. exploit the proximity between candidate expansion terms and query terms in the process of query expansion (Miao et al., 2012). Another directions to improve the performance of pseudo-feedback models is to select more reliable pseudo-documents. For example, Lee et al. presented a cluster-based resampling method to select better pseudo-relevant documents based on the relevance model (Lee et al., 2008). Miao et al. integrate the topic space into pseudo-relevance feedback in order to measure the reliability of the feedback documents (Miao et al., 2016). Ye and Huang evaluate the quality level of pseudo-feedback documents with Learning-to-Rank approach in pseudo relevance feedback (Ye and Huang, 2016).

In this paper, we select expanded terms from both clicked documents (implicit relevance feedback documents) and top-ranked documents in initial retrieval results (pseudo-relevant feedback documents). The clicked documents in previous queries can reveal the information need in current search session, and the pseudo feedback documents can reflect the information need for current query to some extent. The weighting functions for expanded terms will take into account both interaction behaviors in the same session and the relevance of pseudo feedback documents. In our model, the position information is also considered, inspired by the positional language model (Lv and Zhai, 2009b) and positional relevance model(Lv and Zhai, 2010). The dependencies among mixed and multiple types of interactions are modeled in the principled framework of MRF, which has been applied in IR successfully.

For example, Metzler and Croft modeled the term dependencies in queries when ranking documents (Metzler and Croft, 2005) with MRF, and then utilize it to model term dependencies in query expansion (Metzler and Croft, 2007). In addition, a previous work on Markov Random Field models for session search have been proposed(Gao and Zhang, 2012) in order to capture the relationship between document and interactions. The MRF models in this paper are proposed in order to capture the relationship between expansion terms and multiple interaction behaviors.

## 3. INCORPORATING MULTIPLE INTERACTIONS IN MRF FOR QUERY EXPANSION

In this section, we first present a framework, based on the Markov Random Field, to estimate a query expansion model for session search, namely a session language model. Then, we formalize a series of feature functions based on three dependence assumptions. More details for the parameters estimation are also given.

### 3.1. A MRF Based Session Language Model Framework

We propose to estimate a Session Language Model (SLM) $\theta_S$ with the Markov Random Field, based on which we can generate needed expanded terms $w$. According to Figure 1, we can estimate the SLM with current query and the interaction behaviors. The estimation framework of the SLM is formalized as follows:

$$P(w|\theta_S) = P(w|Q,I) \propto P(w,Q,I) = \sum_{D \in \mathcal{F}} P(w,Q,I,D) \tag{1}$$

where $Q$ and $I$ are the current query and the interaction behaviors respectively, and $D \in \mathcal{F}$ is a feedback document (clicked document or pseudo-feedback document). Now, the estimation of SLM becomes a problem of estimating the joint probability $\mathcal{P} = P(w,Q,I,D)$. Note that in this paper we distinguish the interaction behaviors $I$ from the feedback documents $D$. $I$ is particularly focused on the general "behaviors" (e.g., skip the irrelevant results, click on the possibly relevant documents, and query reformulations, etc.), rather than a specific text document.

In order to estimate the joint probability $P(w,Q,I,D)$, we construct a Markov Random Field (MRF) based on each of the dependence graphs $G$ in Figure 2. Each node in the graph represents a random variable. Particularly, the random variables are mutually independent if there is no edge between them. Therefore, we can make different dependence assumptions by deploying corresponding edge configurations in the MRF graph, which will be presented in next subsection in more detail. In this framework, the MRF graph $G$ contains 4 nodes, i.e., $w$, $I$, $D$ and $Q$. The joint probability distribution over the 4 random variables is defined as follows, in a similar way to (Metzler and Croft, 2005):

$$\mathcal{P}_\Lambda = P_\Lambda(w,Q,I,D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \varphi(c;\Lambda) \tag{2}$$

where $C(G)$ is the set of cliques in the MRF graph $G$, $\varphi(c;\Lambda) \geqslant 0$ is a potential function over a clique $c$ ($c$ includes a series of nodes which are fully connected by edges between each other, see Figure 3), $\Lambda$ is a series of parameters which are need to be estimated, $Z_\Lambda = \sum_{w,Q,I,D} \prod_{c \in C(G)} \varphi(c;\Lambda)$ is a normalization factor. However, it is generally infeasible to compute $Z_\Lambda$, since the number of terms in the summation is extremely large. To address

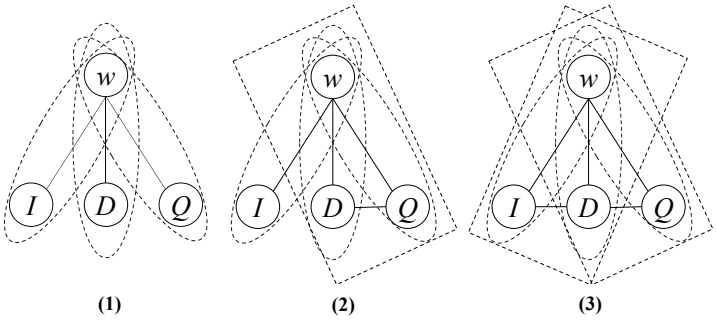


FIGURE 3. All cliques in Markov Random Fields circled by elliptical and squared dotted lines.

this issue, we utilize an exponential function to guarantee the non-negative property of the function (as in (Metzler and Croft, 2005)), formalized as follows:

$$\varphi(c;\Lambda) = \exp\left[\lambda_c f_c(c)\right] \tag{3}$$

where $f_c(c)$ is a real-valued *feature function* over the clique $c$, $\lambda_c$ is the importance weight for a specific feature function. Substituting this potential function back into the Equation 2, we obtain the joint probability distribution:

$$\mathcal{P}_\Lambda \propto \exp\left[\sum_{c\in C(G)} \lambda_c f_c(c)\right] \tag{4}$$

### 3.2. Variants of MRF

In this subsection, we describe and analyze three variants of MRF model underpinned by three different dependence assumptions respectively. The *Full Independence* (FI) assumption considers $Q$, $I$ and $D$ are independent of each other, so is the generation of an expanded term $w$ from them respectively. The *Partial Dependence* (PD) assumption (namely the Query-Document Dependence assumption) assumes that the selection of a feedback document should depend on the current query. This is similar to the underlying idea for the classic Relevance Model (RM1) (Lv and Zhai, 2009a), which assigns each feedback document a weight with the query likelihood. The *Full Dependence* (FD) assumption assumes that the selection of feedback document is dependent on both the current query and the previous interaction behaviors. These three assumptions lead to three variants of MRF, which are detailed next.

3.2.1. *Full Independence Model (FIM).* Underpinned by the full independence assumption, we can construct the MRF model $FIM$ (Figure 3-(1)), in which the three nodes, corresponding to the current query $Q$, interaction behaviors $I$ and feedback document $D$, are independent of each other when $w$ is known. It contains three cliques, i.e., $\{I, w\}$, $\{D, w\}$ and $\{Q, w\}$. Their corresponding feature functions are defined as follows:

$$\begin{aligned} f_Q(Q,w) &= \log P(w|Q)P(Q) \propto \log P(w|Q) \\ f_D(D,w) &= \log P(w|D)P(D) \propto \log P(w|D) \\ f_I(I,w) &= \log P(w|I)P(I) \propto \log P(w|I) \end{aligned} \tag{5}$$

where $P(Q)$, $P(I)$ and $P(D)$ are removed from the feature functions, since they can be regarded as certain events that have occurred (thus $P(Q) = 1$, $P(I) = 1$ and $P(D) = 1$).

In the proposed feature functions, $f_Q(Q, w)$ quantifies how likely the word $w$ and query terms co-occur in the same documents, $f_D(D, w)$ measures the probability of $w$ occurring in the feedback document $D$, and $f_I(I, w)$ models the possibility that the user utilizes $w$ to represent the current information need in mind. Substituting this three feature functions back into Equation 4, we obtain the joint probability distribution

$$\mathcal{P}_\Lambda = \exp\left[\lambda_Q f_Q(Q, w) + \lambda_D f_D(D, w) + \lambda_I f_I(I, w)\right] \tag{6}$$

where $\lambda_Q$, $\lambda_D$ and $\lambda_I$ are three positive free parameters which satisfy condition $\lambda_Q + \lambda_D + \lambda_I = 1$. In Section 3.3, we will describe the computation details for these feature functions.

3.2.2. *Partial Dependence Model (PDM).* The full independence assumption in previous model has an obvious limitation, since intuitively the selection of feedback documents should depend on the original query. Therefore, we can add an edge between the current query node $Q$ and the feedback document node $D$. The motivation is to reward the expanded terms from the feedback documents that are more relevant to original query. Compared with $FIM$, $PDM$ contains one more clique $\{Q, D, w\}$. We define its feature function as below:

$$f_{Q,D}(Q, D, w) = \log P(w|Q, D)P(D|Q)P(Q) \propto \log P(w|Q, D)P(D|Q) \tag{7}$$

In this feature function, $P(D|Q)$ is the relevance probability of feedback document $D$ with respect to the original query $Q$. $P(w|Q, D)$ is the probability that the document and query jointly generate the expansion terms $w$, which rewards the terms that are close to the query terms in the documents. Overall, the joint probability distribution estimated with this Partial Dependence Model (PDM) can be formalized as follows:

$$\mathcal{P}_\Lambda = \exp[\lambda_Q f_Q(Q, w) + \lambda_D f_D(D, w) + \lambda_I f_I(I, w) + \lambda_{Q,D} f_{Q,D}(Q, D, w)] \tag{8}$$

where the positive free parameters should satisfy condition $\lambda_Q + \lambda_D + \lambda_I + \lambda_{Q,D} = 1$. This model can integrate more dependence information than the Full Independence Model. The computation details will be described in Section 3.3.

3.2.3. *Full Dependence Model (FDM).* The Full Dependence Model (FDM) considers that the selection of feedback documents is dependent on both the original query and the previous interaction behaviors. Some feedback documents are relevant to users' information need for current session, while some others are not. Intuitively, the relevant feedback documents should have an influence on the selection of expanded terms. However, the PDM method only uses the original query to estimate the relevant degree of a feedback document. To estimate the importance of each feedback document more precisely and generate more reliable expansion terms, we further improve the MRF model by adding another edge between $I$ and $D$. In this way, richer interaction information is integrated into the model. Accordingly, a new clique $\{I, D, w\}$ is brought into the MRF graph. For the added clique, we define its feature function as follows:

$$f_{I,D}(I, D, w) = \log P(w|I, D)P(D|I)P(I) \propto \log P(w|I, D)P(D|I) \tag{9}$$

where $P(D|I)$ is the relevance probability of feedback documents estimated based on the interaction information, $P(w|I, D)$ is the generative probability given the interactions and the feedback documents. In this way, we can obtain a new joint probability distribution based on the feature functions over all cliques in $FDM$ graph:

$$\begin{aligned}\mathcal{P}_\Lambda = \exp[&\lambda_Q f_Q(Q, w) + \lambda_D f_D(D, w) + \lambda_I f_I(I, w) \\ &+ \lambda_{Q,D} f_{Q,D}(Q, D, w) + \lambda_{I,D} f_{I,D}(I, D, w)]\end{aligned} \tag{10}$$

Similarly, the positive free parameters should satisfy condition $\lambda_Q + \lambda_D + \lambda_I + \lambda_{Q,D} + \lambda_{I,D} = 1$. The computation of the feature functions will be detailed in Section 3.3.

### 3.3. Computation for Feature Functions

To compute all feature functions defined in previous sections, we need to estimate following probabilities: $P(w|Q)$, $P(w|D)$, $P(w|I)$, $P(D|Q)$, $P(w|Q,D)$, $P(D|I)$ and $P(w|D,I)$.

3.3.1. *Estimating $P(w|Q)$, $P(w|D)$ and $P(D|Q)$.* The conditional probability of term $w$ conditioned on the query $Q$ is computed as follows:

$$P(w|Q) = \lambda \frac{co\_df(w, q_1, ..., q_n)}{co\_df(q_1, ..., q_n)} + (1 - \lambda) \sum_{i=1}^{n} \delta_i \frac{co\_df(w, q_i)}{df(q_i)} \tag{11}$$

where $q_1, .., q_n$ are query terms, $co\_df(t_1, ..., t_n)$ is the co-occurrence frequency of terms $t_1, ..., t_n$, $df(t)$ is the document frequency of term $t$. Moreover, in the equation, the first item is the main part, the second item is the smoothing part, $\lambda$ is the smoothing parameter (in this paper, $\lambda = 0.8$), $\delta_i$ is the importance weight of a query term in the query, $\delta_i = \frac{tfidf(q_i)}{\sum_{j=1}^{n} tfidf(q_j)}$, $tfidf(t) = tf(t) \times log\frac{N_C}{df(t)}$, $tf(t)$ is the term frequency of term $t$ in the collection, $N_C$ is the total number of documents in the collection.

The probability of $w$ occurring in a feedback document $D$ is estimated as follows:

$$P(w|D) = \frac{tf(w, D) + \mu P(w|C)}{|D| + \mu} \tag{12}$$

where $tf(w, D)$ is the term frequency of $w$ in $D$, $P(w|C)$ is the prior probability of $w$ occurring in the collection, and $\mu = 2500$ is the smoothing parameter.

Based on the Bayes Rule and the law of total probability, the relevance probability of $D$ to the original query $Q$ can be estimated as follows:

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} = \frac{P(Q|D)P(D)}{\sum_{d \in \mathcal{F}} P(Q, d)} = \frac{P(Q|D)P(D)}{\sum_{d \in \mathcal{F}} P(Q|d)P(d)} \propto \frac{P(Q|D)}{\sum_{d \in \mathcal{F}} P(Q|d)} \tag{13}$$

where each prior probability of feedback documents $P(d)$ or $P(D)$ are assumed to be uniform, $P(Q|D) = \prod_{i=1}^{n} P(q_i|D) \propto \sum_{i=1}^{n} log\ P(q_i|D)$, $n$ is the number of words in the current query $Q$. The final formula can be regarded as the normalized query likelihood of each feedback document based on all feedback documents.

3.3.2. *Estimating $P(w|Q, D)$.* Inspired by the Positional Language Model (PLM) (Lv and Zhai, 2009b) and Positional Relevance Model (Lv and Zhai, 2010), we develop a positional estimation method for the probability of a word $w$ conditioned on the joint of the original query $Q$ and a feedback document $D$. This model rewards the expanded terms that are within a closer proximity to the query terms in feedback documents. The estimation method is shown as follows:

$$P(w|Q, D) = \frac{\sum_{i=1}^{|D|} c(w, i) \cdot \sum_{j=1}^{|Q|} \delta_j \cdot \exp\left[\frac{-(i-p_j)^2}{2\sigma^2}\right] + \mu P(w|C)}{|D| + \mu} \tag{14}$$

where $i$ is a absolute position in the document $D$, $c(w, i) \in \{0, 1\}$ is the occurrence of term $w$ in position $i$, $\delta_j$ is the importance weight of the $j_{th}$ query term in original query $Q$ which is also defined in Equation 11, $p_j$ is the nearest position of query term $q_j$ to the expanded term $w$ in the document. Note that, if $q_j$ is not in the document, then $p_j = i$. We follow the setting of $\sigma = 200$ as used in (Lv and Zhai, 2010). $P(w|C)$ and $\mu$ are the same as in Equation 12.

3.3.3. *Estimating* $P(w|I)$. The conditional probability $P(w|I)$ models how likely a word $w$ is generated conditioned on a user's rich interaction behaviors in the session. To compute the probability, we define three frequently observed behaviors in users' interaction histories, i.e., skips, clicks and query reformulations, formalized as $I = \{skip, click, QR\}$. $QR$ is abbreviation of Query Reformulations. Users usually skip some irrelevant results before deciding to click a document which looks like relevant to current information need after a series of query reformulations. Therefore, we regard "skip" as a kind of negative feedback which indicates the irrelevance of the skipped results, while we regard "click" as positive feedback indicating the relevance of clicked results. Different from the feedback information of "skip" and "click" at the document level, "query reformulations" can signal the relevance or irrelevance at query terms level. Based on these intuitions and inspired by the Query Change Model (QCM) (Guan et al., 2013; Zhang et al., 2013), we propose to model the conditional probability of $w$ conditioned on the interaction behaviors as follows:

$$P(w|I) = \sum_{i=1}^{n} \omega_i P(w|T_i) \tag{15}$$

where $T_i$ is a transition unit from the $i^{th}$ interaction unit to $(i+1)^{th}$ interaction unit. Each $T_i$ can stand for transition between different interaction behaviors including "skip", "clicks" in $I_i$ and the "query reformulations" between $Q_{i+1}$ and $Q_i$, corresponding to $I_{i+1}$ and $I_i$. $T_n$ is the transition from last interaction unit $I_n$ to current query $Q$, $n$ is the total number of the interaction units in the session. $P(w|T_i)$ is the conditional probability conditioned on the transition unit. $\omega_i = Z_n \log(1+i)$ is the discount factor for $i_{th}$ transition unit which penalizes the distant transition units to current query, where $Z_n = 1/\sum_{i=1}^{n} \log(1+i)$ is the normalization factor. We further model the interaction behaviors in each transition unit to estimate the probability $P(w|T_i)$ as follows:

$$\begin{aligned} P(w|T_i) = & \frac{1}{1+\exp\left[P(w|\mathcal{D}_{skip}) - P(w|\mathcal{D}_{click})\right]} \\ & \times \left[\alpha P(w|T_i^{rmv}) + \beta P(w|T_i^{com}) + \gamma P(w|T_i^{add})\right] \end{aligned} \tag{16}$$

where $T_i^{rmv}$, $T_i^{com}$ and $T_i^{add}$ respectively denote the removed query terms, common query terms and the added query terms, compared between $Q_{i+1}$ and $Q_i$, i.e., the query reformulation information.

For example, suppose $Q_{i+1}$ = "$abd$" and $Q_i$ = "$abc$", then the corresponding reformulations are $T_i^{rmv}$ = "$c$", $T_i^{com}$ = "$ab$" and $T_i^{add}$ = "$d$". The conditional probability of $w$ conditioned on reformulation terms (i.e., $T_i^{rmv}$, $T_i^{com}$ and $T_i^{add}$) can be estimated with Equation 11. $\alpha$, $\beta$ and $\gamma$ are the importance factors corresponding to three categories of reformulation terms. We should penalize the removed terms, reward the common terms and the added terms in different degrees when generating expanded terms, since query reformulations can reflect the trend of how a user's search intent changes in the session (Guan et al., 2013; Zhang et al., 2013). To this end, we must guarantee that the conditions $\alpha < \beta < \gamma$ and $\alpha + \beta + \gamma = 1$ are satisfied.

In order to quantify the importance factors ($\alpha$, $\beta$ and $\gamma$), we first formalize their non-normalized formulas, $\alpha' = 1 - \sum_{t \in T_i^{rmv}} P(t|\mathcal{D}_i)$, $\beta' = 1 + \sum_{t \in T_i^{com}} P(t|\mathcal{D}_i)$ and $\gamma' = \max\left[\sum_{t \in T_i^{add}} \log \frac{N_C}{df(t)}, \beta'\right]$, where $\mathcal{D}_i$ is the concatenation (can be seen as a special document) of all snippets for the viewed search results in $i^{th}$ interaction unit, $P(t|\mathcal{D}_i)$ is estimated with Equation 12. Then, they will be normalized, e.g., $\alpha = \alpha'/(\alpha'+\beta'+\gamma')$. The coefficient term before "$\times$" in the Equation 16 models the positive and negative feedbacks indicated by "click" and "skip" behaviors. We will reward the expanded terms occurring in clicked

snippets frequently and penalize those terms occurring in the skipped snippets frequently. $\mathcal{D}_{click}$ and $\mathcal{D}_{skip}$ denote the concatenations of all snippets for clicked results and skipped results. Note that, if the user has not clicked any result, we will select the snippets for all non-clicked results to form $\mathcal{D}_{skip}$. $P(w|\mathcal{D}_{click})$ and $P(w|\mathcal{D}_{skip})$ can be estimated with Equation 12.

3.3.4. *Estimating* $P(D|I)$. $P(D|I)$ measures the relevance probability of a feedback document $D$ given the interaction behaviors, under the assumption that a more relevant feedback document will have a greater influence on generating expanded query terms. For simplicity, we propose an approximation method for estimating the conditional probability, since we are concerned about the relative relevance between feedback documents. The formula is presented as follows:

$$P(D|I) = \frac{Score(D, I)}{\sum_{d \in \mathcal{F}} Score(d, I)} \tag{17}$$

where $Score(\bullet, I)$ is the relevance score of a feedback document (represented with $\bullet$) given a series of interaction behaviors, $\mathcal{F}$ is the set of all feedback documents. Inspired by the idea utilizing the whole session to score the retrieved documents in (Guan et al., 2013), we develop a novel scoring function for modeling the complex interaction behaviors (i.e., skip, click and query reformulations etc.) in the whole session as follows (which is similar to Equations 15 and 16):

$$\begin{aligned} Score(d, I) = \sum_{i=1}^{n} \omega_i Score(d, T_i) = \sum_{i=1}^{n} \omega_i \times \frac{1}{1 + \exp\left[sim(d, \mathcal{D}_{skip}) - sim(d, \mathcal{D}_{click})\right]} \\ \times \left[\alpha QL(d, T_i^{rmv}) + \beta QL(d, T_i^{com}) + \gamma QL(d, T_i^{add})\right] \end{aligned} \tag{18}$$

where the meanings and computation approaches for $T_i$, $\omega_i$, $\mathcal{D}_{click}$, $\mathcal{D}_{skip}$, $T_i^{rmv}$, $T_i^{com}$, $T_i^{add}$, $\alpha$, $\beta$ and $\gamma$ are the same as those in Equations 15 and 16. $sim(d, \bullet)$ is the Cosine similarity between a feedback document $d$ and the special document $\mathcal{D}_{click}$ or $\mathcal{D}_{skip}$, in which all documents are represented with $tf \times idf$ vectors (Salton et al., 1975). The function $1/\left[1 + \exp(\bullet)\right]$ maps the reward values for "click" and the penalty values for "skip" into the interval of $(0, 1)$. We utilize the query likelihood function to compute $QL(d, \bullet)$: specifically, $QL(d, Q) = \prod_{t \in Q} P(t|d) \propto \sum_{t \in Q} \log P(t|d)$, where $P(t|d)$ is estimated by the maximization likelihood approximation with Dirichlet smoothing (Zhai, 2008), and the smoothing parameter is set as $\mu = 1500$ empirically.

3.3.5. *Estimating* $P(w|D, I)$. Similar to $P(w|I)$ and $P(D|I)$, we also estimate $P(w|D, I)$ by utilizing the whole session and considering the positive feedbacks, negative feedbacks and all query reformulations, formalized as follows:

$$\begin{aligned} P(w|D, I) = \sum_{i=1}^{n} \omega_i P(w|D, T_i) = \sum_{i=1}^{n} \omega_i \times \frac{1}{1 + \exp\left[P(w|\mathcal{D}_{skip}) - P(w|\mathcal{D}_{click})\right]} \\ \times \left[\alpha P(w|T_i^{rmv}, D) + \beta P(w|T_i^{com}, D) + \gamma P(w|T_i^{add}, D)\right] \end{aligned} \tag{19}$$

where most parameters have appeared in previous equations (i.e., Equations 15, 16, 17 and 18). The conditional probability of $w$ conditioned on query reformulations and a feedback document $P(w|\bullet, D)$ can be estimated with Equation 14.

**Algorithm 1** : Parameter Tuning Algorithm.

$\mathcal{C}_{FIM} \leftarrow \phi$; // the parameter space for FIM
$\mathcal{C}_{PDM} \leftarrow \phi$; // the parameter space for PDM
$\mathcal{C}_{FDM} \leftarrow \phi$; // the parameter space for FDM
**for** $\lambda_Q = 0$; $\lambda_Q \leqslant 1$; $\lambda_Q$ += $step$ **do**
  **for** $\lambda_D = 0$; $\lambda_D \leqslant 1 - \lambda_Q$; $\lambda_D$ += $step$ **do**
    $\lambda_I \leftarrow 1 - \lambda_Q - \lambda_D$;
    Add configuration $c = \{\lambda_Q, \lambda_D, \lambda_I\}$ into $\mathcal{C}_{FIM}$;
  **end for**
**end for**
Retrieving with each configuration $c \in \mathcal{C}_{FIM}$ and get the best configuration $c' = \{\lambda'_Q, \lambda'_D, \lambda'_I\}$ for FIM;
**for** $\lambda_{Q,D} = 0$; $\lambda_{Q,D} \leqslant 1$; $\lambda_{Q,D}$ += $step$ **do**
  $t \leftarrow 1 - \lambda_{Q,D}$; // $t$ is a temp value
  $\lambda_Q \leftarrow \lambda'_Q \times t$; $\lambda_D \leftarrow \lambda'_D \times t$; $\lambda_I \leftarrow \lambda'_I \times t$;
  Add $c = \{\lambda_Q, \lambda_D, \lambda_I, \lambda_{Q,D}\}$ into $\mathcal{C}_{PDM}$;
**end for**
Retrieving with each $c \in \mathcal{C}_{PDM}$ and get the best configuration $c'' = \{\lambda''_Q, \lambda''_D, \lambda''_I, \lambda''_{Q,D}\}$ for PDM;
**for** $\lambda_{I,D} = 0$; $\lambda_{I,D} \leqslant 1$; $\lambda_{I,D}$ += $step$ **do**
  $t \leftarrow 1 - \lambda_{I,D}$;
  $\lambda_Q \leftarrow \lambda''_Q \times t$; $\lambda_D \leftarrow \lambda''_D \times t$;
  $\lambda_I \leftarrow \lambda''_I \times t$; $\lambda_{Q,D} \leftarrow \lambda''_{Q,D} \times t$;
  Add $c = \{\lambda_Q, \lambda_D, \lambda_I, \lambda_{Q,D}, \lambda_{I,D}\}$ into $\mathcal{C}_{FDM}$;
**end for**
Retrieving with each $c \in \mathcal{C}_{FDM}$ and get the best configuration $c''' = \{\lambda'''_Q, \lambda'''_D, \lambda'''_I, \lambda'''_{Q,D}, \lambda'''_{I,D}\}$ for FDM;

FIGURE 4. Parameter Tuning Algorithm.

### 3.4. Strategies of Parameter Tuning

Given the formalized joint distribution and a set of feature functions, we should further tune the free parameters for each model, i.e., $\lambda_Q$, $\lambda_D$, $\lambda_I$, $\lambda_{Q,D}$ and $\lambda_{I,D}$. It is infeasible to obtain a globally optimized parameter configuration. To address this challenge, we develop an approximation algorithm to find the best parameter configurations in relatively small parameter spaces, which is shown in Figure 4 (Algorithm 1). This algorithm can reduce the parameter space greatly (the variable $step$ controls the actual size of the parameter space), and the tuning speed will depend on the retrieval speed and the number of queries in training set. Specifically, we first search the optimized parameter configuration ($\{\lambda_Q, \lambda_D, \lambda_I\}$) for FIM. Then, control the relative ratio for parameters of FIM and search $\lambda_{Q,D}$ for PDM. Finally, we control the relative ratio for parameters of FDM and search $\lambda_{I,D}$.

## 4. EMPIRICAL EVALUATION

We have developed three query expansion approaches for session search by modeling mixed interactions based on Markov Random Field. To verify the effectiveness of the proposed models, we conduct extensive experiments on the Session Track data of TREC (Text REtrieval Conference) 2013 and 2014 with the Clueweb12 Full corpus.

TABLE 1. The distributions of session number on current query Length (CQLen) and sessions' interaction unit count ($\#I$) for TREC 2013 and 14.

| CQLen | 2013 | 2014 | #I | 2013 | 2014 |
|---|---|---|---|---|---|
| 1 | 0 | 3 | 1 | 20 | 19 |
| 2 | 17 | 17 | 2 | 18 | 12 |
| 3 | 13 | 34 | 3 | 14 | 18 |
| 4 | 20 | 22 | 4 | 6 | 26 |
| 5 | 16 | 9 | 5 | 7 | 12 |
| 6 | 9 | 8 | 6 | 6 | 4 |
| 7+ | 12 | 7 | 7+ | 16 | 9 |
| #ALL | 87 | 100 | - | - | - |

### 4.1. Experimental Setup

The evaluation datasets are from the TREC 2013 and TREC 2014. TREC released 87 session search tasks (sessions) in 2013 and 1021 tasks in 2014. However, given that the TREC 2014's official ground truth only contains the first 100 sessions, we only select 100 sessions for TREC 2014 in our evaluations. In the ground truth, documents are labeled with graded relevance degrees (i.e., -2, 0, 1, 2, 3 and 4, where -2 indicates the document is a spam document, 0 stands for irrelevant document, 1-4 represents different relevance degrees of the document) with respect to the current query. Each search session includes a current query and a series of interaction units (see Figure 1), where each interaction unit records a historical query, the corresponding search results, and some interaction information (e.g., skip, click and dwell time etc.). Session search allows to utilize the whole session to retrieve documents for the current query. We classify all search sessions into several classes according to the lengths of the current queries and the number of interaction units in sessions as shown in Table 1. From the table, we can find that the current-query-lengths of most sessions fall in the interval between 2 and 5. Most sessions have 1 to 5 interactions units.

The document collection used in retrieval is Clueweb12 Full corpus[3] which consists of 733,019,372 English web pages, collected between February 10, 2012 and May 10, 2012. We clean the Clueweb12 corpus by filtering out the spam documents whose Waterloo Spam Ranking scores are less than 70 (Cormack et al., 2011). The corpus is indexed by Indri[4] 5.6. In the indexing process, the stop words are removed and all words are stemmed by porter Stemmer (Porter, 1980). Furthermore, we compare seven retrieval models in our evaluations:

- *LM* (Baseline): The classical language model with Dirichlet smoothing. Negative KL dervergence between language models of query and document is used as ranking function.
- *RM-PF*: Traditional Relevance Model based on pseudo-feedback documents only. We re-implement it to expand the original query based on pseudo-relevance feedback documents.
- *RM-MF*: Relevance Model based on mixed-feedback documents including pseudo feedback documents and clicked documents in interaction histories.
- *PRM-PF*: Positional relevance model (Lv and Zhai, 2010) only based on pseudo-feedback documents, we re-implement it to expand the original query from pseudo relevance feedback documents.
- *PRM-MF*: Positional relevance Model based on mixed-feedback documents including pseudo feedback documents and clicked documents in interaction histories.
- *QCM*: Query Change Model for session search proposed in (Guan et al., 2013). We re-implement it as a re-ranking approach.
- *FIM*: Full Independence Model.
- *PDM*: Partial Dependence Model.

[3] http://www.lemurproject.org/clueweb12/index.php
[4] https://sourceforge.net/projects/lemur/

TABLE 2. Optimal parameter configurations for different retrieval models.

| Models | Optimal Parameter Configurations |
|---|---|
| PRM-PF | $\lambda = 0.1, \sigma = 200$ (Lv and Zhai, 2010) |
| PRM-MF | $\lambda = 0.1, \sigma = 200$ (Lv and Zhai, 2010) |
| QCM | $\alpha = 2.2, \beta = 1.8, \epsilon = 0.07, \delta = 0.4, \gamma = 0.92$ (Guan et al., 2013) |
| FIM | TREC 2013 Part A: $\lambda_Q = 0.2, \lambda_D = 0.16, \lambda_I = 0.64$;<br>TREC 2013 Part B: $\lambda_Q = 0.4, \lambda_D = 0.36, \lambda_I = 0.24$;<br>TREC 2014 Part A: $\lambda_Q = 0.2, \lambda_D = 0.64, \lambda_I = 0.16$;<br>TREC 2014 Part B: $\lambda_Q = 0.2, \lambda_D = 0.16, \lambda_I = 0.64$ |
| PDM | TREC 2013 Part A: $\lambda_Q = 0.16, \lambda_D = 0.13, \lambda_I = 0.51, \lambda_{Q,D} = 0.2$;<br>TREC 2013 Part B: $\lambda_Q = 0.32, \lambda_D = 0.29, \lambda_I = 0.19, \lambda_{Q,D} = 0.2$;<br>TREC 2014 Part A: $\lambda_Q = 0.16, \lambda_D = 0.51, \lambda_I = 0.13, \lambda_{Q,D} = 0.2$;<br>TREC 2014 Part B: $\lambda_Q = 0.04, \lambda_D = 0.03, \lambda_I = 0.13, \lambda_{Q,D} = 0.8$; |
| FDM | TREC 2013 Part A: $\lambda_Q = 0.26, \lambda_D = 0.23, \lambda_I = 0.15, \lambda_{Q,D} = 0.16, \lambda_{I,D} = 0.2$;<br>TREC 2013 Part B: $\lambda_Q = 0.26, \lambda_D = 0.23, \lambda_I = 0.15, \lambda_{Q,D} = 0.16, \lambda_{I,D} = 0.2$;<br>TREC 2014 Part A: $\lambda_Q = 0.26, \lambda_D = 0.23, \lambda_I = 0.15, \lambda_{Q,D} = 0.16, \lambda_{I,D} = 0.2$;<br>TREC 2014 Part B: $\lambda_Q = 0.26, \lambda_D = 0.23, \lambda_I = 0.15, \lambda_{Q,D} = 0.16, \lambda_{I,D} = 0.2$; |

- *FDM*: Full Dependence Model.

For all expansion models ***RM-PF, RM-MF, PRM-PF, PRM-MF, FIM, PDM and FDM***, the common free parameters are set to the same values. Specifically, when retrieving for a query, we apply corresponding model to select 50 weighted terms from the feedback documents and expand the representation of original query. Top 10 retrieved documents in the first round search results with language model are selected as pseudo-feedback documents, since existing work has indicated that pseudo-feedback models often gain the best performance when selecting about 10 pseudo relevance feedback documents (Lv and Zhai, 2009a). The second round search results are obtained by running the expanded queries with the Indri search engine. The TREC's official evaluation metrics, NDCG and MAP, are adopted to evaluate the performance of aforementioned retrieval models. Note that, we compute the MAP based on top N retrieved documents rather than all, namely, $MAP@N = \frac{\sum_{q=1}^{Q} AP_q@N}{Q}$, where $Q$ is the number of tested queries and $AP_q@N = \frac{\sum_{k=1}^{n}(P_q(k) \times rel(k))}{N}$, where $P_q(k) = |\{relevant\ documents\} \cap \{top\ k\ retrieved\ documents\}|/k$, $rel(k)$ is an indicator function equaling 1 if the item at rank $k$ is a relevant document, zero otherwise. The difference between the definition of $AP$ in this paper and the standard $AP$ is that we use the number of top N retrieved documents as the denominator rather than the number of all retrieved documents. In this paper, we report the MAP values based on different N values.

## 4.2. Evaluation Results

In this subsection, we test different retrieval models with corresponding optimal parameters (tuned in previous section) on large scale data, i.e., all Session Track Tasks in TREC 2013 and 2014 as reported in Table 1. The best parameter configurations for all tested models are summarized in Table 2. We tune the parameters for the proposed models (i.e., FIM, PDM and FDM) on different subsets of data. Specifically, we separate TREC 2013 (and 2014) data into two parts (i.e., Part A and B) randomly. For TREC 2013, there are 44 sessions in Part A and 43 sessions in Part B. For TREC 2014, there are 50 sessions in two parts. We use one part as training set (tuning parameters) and another as the testing set (using the trained parameter for testing). We apply Algorithm 1 described in Figure 4 to tune parameters. In Table 2, we report the used parameters for each testing part. The overall average performance are analyzed respectively.

Table 3 reports the performance of retrieval models on TREC 2013 and 2014 evaluated

with NDCG and MAP. The table shows that all query expansion models and query change model outperform the baseline (LM), which demonstrates that the exploitation of session interaction information can significantly benefit for the search performance.

Moreover, our models outperform other query expansion models (i.e., RM and PRM) with respect to all evaluation metrics. Specifically, our models outperform the RM and PRM models on both data sets. From the table, we can also find that our models are competitive with the state-of-the-art session search model, i.e., QCM, with respect to most evaluation metrics (except for NDCG@10). According to the reported results, we find that, for TREC 2013, our proposed models are similar to QCM in terms of different evaluation metrics. For TREC 2014, our proposed models can significantly outperform QCM. This phenomena may be resulted from the different features (e.g., query length and interaction unit count) of TREC 2013 and TREC 2014. Specifically, from the Table 1, we find that the proportion of long sessions (i.e., $\#I \geqslant 3$) in TREC 2014 (69%) is larger than that in TREC 2013 (56.3%), which shows that our proposed models can better handle the long sessions than QCM by effectively modeling the dependency among different interactions.

Comparing between RM-PF and RM-MF, we find that the retrieval performances of RM-MF are better than RM-PF on both TREC 2013 and 2014 with respect to all evaluation metrics. This demonstrates that utilizing clicked documents as implicit feedback when expanding the original query can improve the quality of expansion terms significantly. This also shows that "click" is one of the positive feedback interaction behaviors. An unexpected phenomenon is that PRM-MF fails to outperform PRM-PF consistently. The possible reason my be that Positional Relevance Model (PRM) assigns weights for expanded terms considering the distance between current query and expanded terms in the feedback documents. In historical clicked documents, the occurrence frequency of current query terms is small, which leads to that the weights of expanded terms in clicked documents are very small.

From Table 3, we find that the dependence models (PDM and FDM) are often superior to the independence model (FIM), which shows that modeling dependencies among mixed interactions is effective for improving the retrieval performance. However, FDM fails to outperform the PDM. The possible reason is that FDM rewards or penalizes some wrong documents when selecting feedback documents compared with PDM.

| (a) Overall Performance for TREC 2013 | | | | |
|---|---|---|---|---|
| Models | NDCG@10 | NDCG@100 | MAP@10 | MAP@100 |
| LM | 0.0570 | 0.0600 | 0.0071 | 0.0147 |
| RM-PF | 0.0951 | 0.0982 | 0.0178 | 0.0354 |
| RM-MF | 0.1042‡ | 0.113‡ | 0.0201† | 0.0426‡ |
| PRM-PF | 0.1103‡ | 0.1137‡ | 0.0214† | 0.046‡ |
| PRM-MF | 0.1114‡ | 0.1141‡ | 0.0211† | 0.0457‡ |
| QCM | **0.1425**‡ | 0.1299‡ | 0.0247† | 0.0495‡ |
| FIM | 0.1418‡ | 0.1264‡ | 0.0268‡ | 0.0518‡ |
| PDM | 0.1396‡ | **0.1306**‡ | **0.0268**‡ | **0.0521**‡ |
| FDM | 0.1348‡ | 0.1302‡ | 0.0265‡ | 0.051‡ |
| (b) Overall Performance for TREC 2014 | | | | |
| Models | NDCG@10 | NDCG@100 | MAP@10 | MAP@100 |
| LM | 0.1084 | 0.114 | 0.0202 | 0.0404 |
| RM-PF | 0.1247 | 0.1305 | 0.0230 | 0.0522 |
| RM-MF | 0.1246 | 0.1356‡ | 0.0231 | 0.0539† |
| PRM-PF | 0.1407† | 0.1517‡ | 0.0247 | 0.0629‡ |
| PRM-MF | 0.1392† | 0.1541‡ | 0.0244 | 0.0637‡ |
| QCM | 0.1321 | 0.1450 | 0.0200 | 0.0553† |
| FIM | 0.1658† | 0.1590‡ | 0.0281† | 0.0674‡ |
| PDM | **0.1661**‡ | **0.1614**‡ | **0.0286**‡ | **0.0679**‡ |
| FDM | 0.1638† | 0.1605‡ | 0.0284‡ | 0.0676‡ |

TABLE 3. Overall performances for TREC 2013 and 2014 with respect to NDCG and MAP. Significance Test has been done for different retrieval models compared with RM3-PF model, where the symbol ‡ means $p < 0.01$ with paired t-test, † means $p < 0.05$.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a unified framework based on Markov Random Fields (MRF), for modelling and incorporating complex dependencies between mixed interaction feedbacks, e.g., skip, click and query reformulations in search sessions, to estimate a session language model and then use it to expand the current query in session search. Based on MRF, we presented three dependency assumptions and correspondingly three MRF variants are derived. Rich interaction information is captured by computing the feature functions of MRF variants. Extensive experiments have been carried out on two large scale standard data sets. The results demonstrate that our models outperform two strong baselines (RM and PRM) significantly on the most sessions. Moreover, the dependency models outperform the independence model significantly.

The experimental results have validated the importance of modeling the mixed interactions and their complex dependencies in information retrieval. In the future, more dependency cases could be considered in the MRF. The proposed models may be further improved by reducing free parameters and exploiting automatic parameters tuning methods, e.g., machine learning. Additionally, we consider that efficiency is another important performance issue for our model, especially when the method is applied to real Web search settings. Basically, before the proposed models can be applied to Web search settings, we should improve the efficiency of the computation of required probabilities. For example, we can perform some complex probability computation offline. This is left as a key direction of our future work.

## 6. ACKNOWLEDGEMENTS

This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No. 2014CB744604, 2013CB329304), the Chinese 863 Program (grant No. 2015AA015403), the Natural Science Foundation of China (grant No. U1636203,

61772363, 61272265, 61402324), the Tianjin Research Program of Application Foundation and Advanced Technology (grant no. 15JCQNJC41700), and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721321.

## REFERENCES

ARGUELLO, JAIME, JONATHAN L ELSAS, JAMIE CALLAN, and JAIME G CARBONELL. 2008. Document representation and query expansion models for blog recommendation. ICWSM, **2008**(0):1.

BAI, JING, DAWEI SONG, PETER BRUZA, JIAN-YUN NIE, and GUIHONG CAO. 2005. Query expansion using term relationships in language models for information retrieval. *In* CIKM, ACM, pp. 688–695.

BODNER, RICHARD C, and FEI SONG. 1996. Knowledge-based approaches to query expansion in information retrieval. Springer.

BUSCHER, GEORG, ANDREAS DENGEL, and LUDGER VAN ELST. 2008. Query expansion using gaze-based feedback on the subdocument level. *In* SIGIR, ACM, pp. 387–394.

CARPINETO, CLAUDIO, and GIOVANNI ROMANO. 2012. A Survey of Automatic Query Expansion in Information Retrieval. ACM.

CHEN, YONGQIANG, PENG ZHANG, DAWEI SONG, and BENYOU WANG. 2015. A real-time eye tracking based query expansion approach via latent topic modeling. *In* CIKM, pp. 1719–1722.

CHIRITA, PAUL-ALEXANDRU, CLAUDIU S FIRAN, and WOLFGANG NEJDL. 2007. Personalized query expansion for the web. *In* SIGIR, ACM, pp. 7–14.

CORMACK, GORDON V, MARK D SMUCKER, and CHARLES LA CLARKE. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval, **14**(5):441–465.

CUI, HANG, JI-RONG WEN, JIAN-YUN NIE, and WEI-YING MA. 2002. Probabilistic query expansion using query logs. *In* WWW, ACM, pp. 325–332.

CUI, HANG, J-R WEN, JIAN-YUN NIE, and WEI-YING MA. 2003. Query expansion by mining user logs. Knowledge and Data Engineering, IEEE Transactions on, **15**(4):829–839.

GAO, JIANFENG, GU XU, and JINXI XU. 2013. Query expansion using path-constrained random walks. *In* SIGIR, ACM, pp. 563–572.

GAO, YASI, and C. ZHANG. 2012. A session-oriented retrieval model based on markov random field. *In* Proceedings of the 3rd IEEE International Conference on Network Infrastructure and Digital Content, pp. 641–645.

GOTTIPATI, SWAPNA, and JING JIANG. 2011. Linking entities to a knowledge base with query expansion. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 804–813.

GUAN, DONGYI, SICONG ZHANG, and HUI YANG. 2013. Utilizing query change for session search. *In* SIGIR, ACM, pp. 453–462.

HUANG, XIANGJI, FUCHUN PENG, AIJUN AN, and DALE SCHUURMANS. 2004. Dynamic web log session identification with statistical language models. Journal of the Association for Information Science and Technology, **55**(14):12901303.

JOACHIMS, THORSTEN, LAURA GRANKA, BING PAN, HELENE HEMBROOKE, FILIP RADLINSKI, and GERI GAY. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM TOIS, **25**(2):7.

LAVRENKO, VICTOR, and W BRUCE CROFT. 2001. Relevance based language models. *In* SIGIR, ACM, pp. 120–127.

LEE, KYUNG SOON, W BRUCE CROFT, and JAMES ALLAN. 2008. A cluster-based resampling method for pseudo-relevance feedback. *In* SIGIR, ACM, pp. 235–242.

LI, JINGFEI, DAWEI SONG, PENG ZHANG, JI-RONG WEN, and ZHICHENG DOU. 2014. Personalizing web search results based on subspace projection. *In* Information Retrieval Technology. Springer, pp. 160–171.

LIU, XIAOYONG, and W BRUCE CROFT. 2002. Passage retrieval based on language models. *In* CIKM, ACM, pp. 375–382.

LV, YUANHUA, and CHENGXIANG ZHAI. 2009a. A comparative study of methods for estimating query language models with pseudo feedback. *In* CIKM, ACM, pp. 1895–1898.

LV, YUANHUA, and CHENGXIANG ZHAI. 2009b. Positional language models for information retrieval.

*In* SIGIR, ACM, pp. 299–306.

LV, YUANHUA, and CHENGXIANG ZHAI. 2010. Positional relevance model for pseudo-relevance feedback. *In* SIGIR, ACM, pp. 579–586.

METZLER, DONALD, and W BRUCE CROFT. 2005. A markov random field model for term dependencies. *In* SIGIR, ACM, pp. 472–479.

METZLER, DONALD, and W BRUCE CROFT. 2007. Latent concept expansion using markov random fields. *In* SIGIR, ACM, pp. 311–318.

MIAO, JUN, JIMMY XIANGJI HUANG, and ZHENG YE. 2012. Proximity-based rocchio's model for pseudo relevance. *In* International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 535–544.

MIAO, JUN, JIMMY XIANGJI HUANG, and JIASHU ZHAO. 2016. Topprf: A probabilistic framework for integrating topic space into pseudo relevance feedback. Acm Transactions on Information Systems, **34**(4).

NAVIGLI, ROBERTO, and PAOLA VELARDI. 2003. An analysis of ontology-based query expansion strategies. *In* ECML, pp. 42–49.

PAN, DAZHAO, PENG ZHANG, JINGFEI LI, DAWEI SONG, JI-RONG WEN, YUEXIAN HOU, BIN HU, YUAN JIA, and ANNE DE ROECK. 2013. Using dempster-shafers evidence theory for query expansion based on freebase knowledge. *In* Information Retrieval Technology. Springer, pp. 121–132.

PORTER, MARTIN F. 1980. An algorithm for suffix stripping. Program: electronic library and information systems, **14**(3):130–137.

ROCCHIO, JOSEPH JOHN. 1971. Relevance feedback in information retrieval.

SALTON, G., A. WONG, and C. S. YANG. 1975. A vector space model for automatic indexing. Communications of the Acm, **18**(11):273–280.

WHITE, RYEN W., IAN RUTHVEN, and JOEMON M. JOSE. 2002. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. *In* Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 57–64.

WHITE, RYEN W., IAN RUTHVEN, JOEMON M. JOSE, and C. J. VAN RIJSBERGEN. 2005. Evaluating implicit feedback models using searcher simulations. Acm Transactions on Information Systems, **23**(3):325–361.

YE, ZHENG, and JIMMY XIANGJI HUANG. 2016. A learning to rank approach for quality-aware pseudo-relevance feedback. Journal of the Association for Information Science and Technology, **67**(4):942959.

ZHAI, CHENGXIANG. 2008. Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies, **1**(1):1–141.

ZHANG, PENG, YUEXIAN HOU, and DAWEI SONG. 2009. Approximating true relevance distribution from a mixture model based on irrelevance data. *In* SIGIR, ACM, pp. 107–114.

ZHANG, PENG, JINGFEI LI, BENYOU WANG, XIAOZHAO ZHAO, DAWEI SONG, YUEXIAN HOU, and MASSIMO MELUCCI. 2016. A quantum query expansion approach for session search. Entropy, **18**(4):146.

ZHANG, SICONG, DONGYI GUAN, and HUI YANG. 2013. Query change as relevance feedback in session search. *In* SIGIR, ACM, pp. 821–824.