This is the Pre-Published Version.

Electronic version of an article published as International Journal of Software Engineering and Knowledge Engineering, Vol. 29, No. 06, pp. 873-895, https://doi.org/10.1142/S021819401950030X © World Scientific Publishing Company https://www.worldscientific.com/worldscinet/ijseke

BINARY INDEPENDENCE LANGUAGE MODEL IN A RELEVANCE FEEDBACK ENVIRONMENT

H.C. Wu*, R.W.P. Luk†

Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
*cshcwu@comp.polyu.edu.hk
†csrluk@comp.polyu.edu.hk

K.F. Wong

Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
kfwong@se.cuhk.edu.hk

J.Y. Nie

Département d'Informatique et Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ Centre-Ville, Montréal, Québec, Canada
nie@iro.umontreal.ca

Received (27 July 2018)
Revised (30 September 2018)
Accepted (Day Month Year)

Model construction is a kind of knowledge engineering, and building retrieval models is critical to the success of search engines. This article proposes a new (retrieval) language model, called binary independence language model (BILM). It integrates two document-context based language models together into one by the log-odds ratio where these two are language models applied to describe document-contexts of query terms. One model is based on relevance information while the other is based on the non-relevance information. Each model incorporates link dependencies and multiple query term dependencies. The probabilities are interpolated between the relative frequency and the background probabilities. In a simulated relevance feedback environment of top 20 judged documents, our BILM performed statistically significantly better than the other highly effective retrieval models at 95% confidence level across four TREC collections using fixed parameter values for the mean average precision. For the less stable performance measure (i.e., precision at the top 10), no statistical significance is shown between the different models for the individual test collections although numerically our BILM is better than two other models with a confidence level of 95% based on a paired sign test across the test collections of both relevance feedback and retrospective experiments.

Keywords: Information retrieval; Language model; Proximity Matching.

1. Introduction

Language model (e.g., Ponte and Croft, 1998; Hiemstra, 1998) is applied widely and successfully not just in information retrieval (e.g., Zhao et al., 2012; Peng and Liu, 2015) but in natural language processing tasks like automatic speech recognition (e.g., Jelinek and Mercer, 1980). It is one of the effective retrieval models. In relevance feedback (that

feedback judged top documents in the initial retrieval for the second retrieval by modifying the query with terms from the top judged documents), most language models only make use of the relevant documents only. We believe that the judged nonrelevant documents in relevance feedback have information that may improve retrieval effectiveness. Our research problem is to extend language model in a principle manner so that it can make use of both judged relevant and nonrelevant documents in relevance feedback for better retrieval effectiveness. Such model construction is a kind of knowledge engineering which is of critical importance to many search engines as the ranking algorithms are specified by the retrieval models.

Our contribution is in extending the language modeling approach by using two document-context based language models which are language models applied to model the document-contexts of query terms (i.e., terms surrounding the query term within a specified distance from the query term in the document). One document-context based language model captures the relevance information and the other captures the non-relevance information. Apart from capturing both relevance and non-relevance information, our other contribution is in the formulation of the document-context based language models. In a novel way, we embedded the document-contexts in the probabilities of the document-context based language models, and estimated them similar to other language models. Our model has more flexibility in terms of modeling. For example, it can be extended to model link dependencies (e.g., Gao et al., 2004) which are the linkages of two possibly non-consecutive dependent terms within a query context (i.e., words surrounding a query term) and multiple query term dependencies within a single framework. It can also be integrated with the log-odds ratio (Robertson and Sparck Jones, 1976) which is the logarithm of the probability of relevance divided by the probability of non-relevance. It has the potential to improve retrieval effectiveness which is hard to do so as the state-of-the-art retrieval models are already highly effective.

We evaluate our document-context based language models using a simulated relevance feedback environment. The mean average precision (MAP) of our proposed model is statistically significantly better than the MAPs of the highly effective baseline models at 95% confidence level, using top 20 documents to feedback the judged documents. The retrospective experiment shows that our model is comparable to support vector machine in terms of MAP, and the MAP of our model is statistically significantly better than the MAPs of other highly effective baseline models.

The significance of our contribution is the creation of a highly effective retrieval model that integrates two document-context based language models. These two language models are instantiated from the log-odds ratio which originally forms the basis for ranking of the binary independence model (Robertson and Sparck Jones, 1976) (i.e., a classical probabilistic retrieval model). This suggests that the language modeling approach can be extended in a novel way by combining it with the classical probabilistic retrieval. The creation of the effective model here is also relevant to relevance feedback for spoken document retrieval which can now leverage both the relevance and non-relevance information for more effective retrieval by using the proposed retrieval model.

The rest of the article is organized as follows. The next section is the related work. The main section following the related work presents the binary independence language model in details, including how to estimate the probabilities. The experiment section presents the evaluation of the model against other highly effective ones (e.g., language model and Markov random field model). One experiment is carried out in a simulated relevance feedback environment, and the other is a retrospective study. The last section makes the concluding remarks and mentions the future work.

2. Related Work

We first discuss the related work about language model since our novel model is a kind of language model. In the next subsection, we discuss the proximity-matching models. Finally, we discuss the context-dependent term frequency adjustment methods.

2.1. *Language Model*

Ponte and Croft proposed the use of language model (Ponte and Croft, 1998) in information retrieval as well as Hiemstra (1998). As language model received much attention, it was investigated by many (e.g., Zhai and Lafferty, 2003). Its strength lies in the principle method to estimate the probabilities as well as the application of smoothing techniques. However, its retrieval effectiveness performance was sometimes comparable to the performance of other state-of-the-art retrieval models (e.g., BM25 in Robertson and Walker, 1994). There had been extension to language models using bigrams (e.g., Song and Croft, 1999), which are two consecutive terms occurring one after the other, but usually they did not advance much the retrieval effectiveness of the language model for heterogeneous collection.

There were concerns over the theoretical basis of language models because it did not have the relevance (random) variable in its probability. Lafferty and Zhai (2003) showed that the language model could be made to be rank equivalent to the log-odds ratio by making two general assumptions. Later, Azzopardi (2007) and Luk (2008) showed that these two assumptions could eliminate the relevance (random) variable making the log-odds ratio rank equivalent to the probability (i.e., $p(d \mid q)$ where $p(\cdot)$ is the probability, d refers to a document and q refers to a query) of the language model of Ponte and Croft.

Language models were applied in the relevance feedback environment. Specifically, Zhai and Lafferty (2001) looked at model-based (relevance) feedback where a generative model of feedback documents was used or the average divergence of the feedback document was optimized. In this article, we compared one of the model-based feedback by Zhai and Lafferty (see Relevance Feedback Experiments Subsection). Regularization (Diaz, 2008) was applied to improve retrieval effectiveness in the relevance feedback environment. However, the measured performance of regularization included the judged documents whereas our evaluation was based on the residue collection without the judged documents. Also, our proposed method could be added with regularization, so regularization was not compared with our proposed method. Similarly, local concept expansion (Metzler and Croft, 2007) has been applied to MRF. We do not compare the

use of such method as it can be applied to MRF and to our model as well. In general, there are many additional methods that can be applied to language model, MRF and our model, and it is impractical to compare with all of the additional methods, as well as whether it is valid and whether it has value to compare the different retrieval models with such additional methods.

We combine the relevance and non-relevance information in the log-odds ratio similar to the work of Robertson and Sparck Jones (1976). Lv and Zhai (2015) also combined the use of relevance and non-relevance information in the query likelihood log-odds ratio which is rank equivalent to the log-odds ratio of Robertson and Sparck Jones (1976). In general, our work here and the work of Lv and Zhai (2015) are both derived from the log-odds ratio, but they derive the query likelihood probabilities which are used in ranking. Unfortunately, Lv and Zhai (2015) only reported marginal improvement (i.e., 0.0009 points better than 0.2521) using both positive and negative information (i.e., the query likelihood log-odds) for short queries in the Robust04 test collection, the document collection of which is the same as the document collection of TREC-7 (i.e., the 7th annual conference of TREC which stands for Text Retrieval Conference) and TREC-8 as used in this work. In addition to marginal improvement, they did not report any statistical significance. As the work of Lv and Zhai (2015) only has marginal improvement over the language model for short queries (as in here) for Robust04 track, we believe that it is sufficient to compare our proposed model with the language model in this paper.

Wang et al. (2007, 2008) also made use of negative information (or non-relevance information) in (relevance) feedback but (1) their methods only operate with negative information without positive information, (2) they have not tried their method to use both with relevance information and non-relevance information to perform feedback and retrieval, so we are unable to compare with their results, and (3) they admitted that their proposed method is only a(n) (effective) heuristic rather than a principled method. Therefore, we do not compare our work with theirs.

2.2. *Context-dependent Model*

Document-context language models are a special type of document-context based models which are proposed and developed by various researchers. Vechtomova and Robertson (2000) proposed a document-context based model using collocates or word co-occurrences filtered by mutual information or Z statistics (i.e., the score minus the mean and then divided by the standard deivation). Our document-context based language model did not have any collocates or word co-occurrences to be filtered. Pickens and MacFarlane (2006) also made use of document contexts for ranking. The MAP of their model was similar to the state-of-the-art model, and their ideas were independently developed from ours. Earlier, we (Wu et al., 2006) developed a number of document-context based models. However, the MAPs of these models in a retrospective study were not as effective as our proposed model here (see Retrospective Experiments Subsection). Therefore, we focused on developing our proposed model for relevance feedback.

There are various attempts that try to generalize the term-independent model with proximity matching within document-contexts. For example, the BM25 (Robertson and Walker, 1994) is generalized to BM25 with proximity matching (Rasolofo and Savoy, 2003). Huston and Croft (2014) compared a number of proximity matching models and found that the Markov Random Field model (Metzler and Croft, 2005) is among the best. Therefore, we will use the Markov Random Field model as our baseline for comparison.

2.3. *Context-dependent Term Frequency Adjustment Method*

Another approach is to formulate the document-context dependent term weight proposed by Dang et al. (2010, 2014 and 2016). The main idea is to use terms in the context (called context terms) to boost or discount the term frequency factor (i.e., the occurrence frequency component) of the term weights (i.e., a measure of the importance of the term) of the query terms. In a relevance feedback environment, context terms that occur only in relevant documents will be used to boost the term frequency factor. Likewise, context terms that occur only in non-relevant documents will be used to discount the term frequency factor. The amount of boost and discount is shaped by the logistic regression function. Our work here is different from Dang et al. (2010, 2014 and 2016) who seek to modify existing term weights with context dependencies because our work derives and instantiates the document-context language model from the log-odds ratio instead of modifying existing term weights. Also, the MAP of our novel retrieval model is better than the MAP of MRF with statistical significance for all the tested collections whereas the modified term weights of Dang et al. with context dependencies sometimes may not perform better than MRF for some tested collections. Therefore, we believe that our model here is better than the modified term weights of Dang et al. (2010, 2014 and 2016).

Similar to but different from Dang et al (2010), Lv & Zhai (2009, 2010) modified the counting of the language model and the relevance model based on proximity matching, respectively, to rank documents. For the positional language model, Lv & Zhai (2009) shows only marginal improvement using the smaller TREC collections which are known to be difficult to show performance enhancement by proximity matching due to their small size (Buttcher et al., 2006) similar to our test collections used here. Therefore, Lv & Zhai (2010) shows better performance improvement for the positional relevance model after using the larger test collections (i.e., GOV2 and Clueweb09) which can show better performance for proximity matching as there may be more noise terms in the larger collection (Buttcher et al., 2006). Unfortunately, Lv and Zhai (2009, 2010) did not compare their proximity matching language model with other proximity matching model like MRF, so there is no direct comparison available. However, if we dig into the GOV2 terabyte 2006 TREC archive, then we can find the MAP of MRF is 0.3670 by Li and Yan (2006), which is much higher than 0.3322 that is the best MAP of the positional relevance model of Lv and Zhai (2010) so that their difference is unlikely to be due to some preprocessing. Therefore, we believe that it is sufficient for us to compare the performance of our work with MRF. Similarly, there are other proximity matching models (e.g., Rasolofo and Savoy, 2003) such as those based on BM25 (Robertson and

Walker, 1994). However, Huston and Croft (2014) showed that MRF was able perform among the best compared with various proximity matching models. Therefore, we will focus on comparing with MRF.

3. Binary Independence Language Model Initial Formulation

Our binary independence language model (BILM) is formulated initially as follows. First, we introduce the event space of the binary independence language model. Next, we start the derivation from the log-odds ratio which is originally used by the binary independence model. This derivation leads to the use of document context models in the ranking formula.

3.1. *Event Space*

We follow the notation and interpretation of probabilities for relevance decisions in Section 2.2 of Wu et al. (2008). For convenience, we mention the following. First, probabilities are defined by the axioms of Kolmogorov (1950). A probability is a measure in some event space Ω (Robertson, 2005) and it is denoted by $p_\Omega(\cdot)$. We define the local relevance as the relevancy of the information (to the query) at or near a particular position in the document. The event space of the local relevance is $\Omega_\partial = D \times N \times Q \times R$ where D is the set of documents, N is the set of nonnegative integers (indicating the position of the relevance in the document), Q is the set of queries, and R is the set of local relevance values (i.e., for binary relevance, $R = \{r, \bar{r}\}$). We will show how the local relevance is related to the document-wide relevance later.

To make the local relevance to depend on the information in the document instead of the position in the document, we formulate the event space of the (local) relevance decisions which are independent of the position in the document. The event space, $\Omega_{\partial,n}$, of the relevance decisions is determined by the context-based local relevance decision (CLRD) assumption in (Wu et al., 2008). It states that a local relevance decision at any location k in any document d for any query q is made on the basis of the information in the context that is centered at k in d for some maximal context size n (i.e., the context has $2n$+1 terms). We denote the following in order to specify the event space of the relevance decisions:

- the set of terms in the collection D be $V(D)$;
- the set of possible strings of length $2n$+1 over $V(D)$ be $V(D)^{2n+1}$ where $V(D)^{2n+1}$ is the cross product of $V(D)$ itself by $2n$+1 times.

According to the CLRD assumption, the event space, $\Omega_{\partial,n}$, of the local relevance decisions is a subset of the following cross-product space:

$$\Omega_{\partial,n} \subseteq V(D)^{2n+1} \times Q \times R,$$

since some events in the cross-product space are undefined in $\Omega_{\partial,n}$. The context string, $c(d, k)$, of length $2n$+1, is in $V(D)^{2n+1}$. Near the beginning or the end of the document, the context string is padded with a unique character so that the length of the context string is $2n + 1$. We assume that this special character is already in $V(D)$. The previous cross-

product space includes the set N of positive integers because events are specified by the document locations.

In this article, the document-wide relevance values are r_g and $\bar{r}_g$, and the set of these two values is denoted by H. Similar to local relevance, the event space, Ω_∇, of the document-wide relevance is

$$\Omega_\nabla = D \times Q \times H ,$$

which is the same as the event space of the evaluation model of Dang et al. (2009). Note that all events in the above cross-product space are defined in Ω_∇. Both D and Q in Ω_∇ are the same as those in Ω_∂. We denote the probability of document-wide relevance as $p_\nabla(R_{d,q})$ where $R_{d,q}$ is the global relevance (random) variable for document d and query q. Similarly, we denote the probability of local relevance as $p_\partial(R_{d,k,q})$, where ∂ identifies the underlying event space as Ω_∂ and $R_{d,k,q}$ is the local relevance (random) variable for document d at location k for query q.

3.2. *Derivations from the Log-Odds Ratio from BIM*

Sections, sub-sections and sub-subsections are numbered in Arabic. Use double spacing before all section headings, and single spacing after section headings. Flush left all paragraphs that follow after section headings.

Using our notation, the log-odds ratio (1992) of the binary independence model (BIM) by Robertson and Sparck-Jones (1976) is

$$O_\nabla(r_g \mid d, q) = \log \frac{p_\nabla(R_{d,q} = r_g)}{p_\nabla(R_{d,q} = \bar{r}_g)}, \qquad (1)$$

which is defined over the event space Ω_∇. Effectively, the log-odds ratio is pooling the logarithm of the relevance probability and the logarithm of the non-relevance probability. The log-odds ratio is used because it can incorporate nonrelevant information and it is rank equivalent to the probability of relevance given the query and the document according to Manning et al. (2008), so that the ranking based on the log-odds ratio complies with the probability ranking principle (Robertson, 1977). This principle is followed because Dang et al. (2009) shown that if the probability of relevance is estimated accurately, then the retrieval will yield optimal effectiveness in a range of measures, such as Mean Average Precision (MAP), R-precision, etc.

Our model aggregates the evidence found in events at each location in the document. These pieces of evidence can be grouped into two types according to the generalized query-centric assumption. This assumption states that for any topic q and any document d relevant to q, the relevant information for q locates only in the contexts $c(d, k)$ where $|d|$ is the city-block length of the document d, $k \in [1, |d|]$, $G(q)$ returns the set of terms related to q and $d[k] \in G(q)$. One type, $E_1(d, q)$, contains events, $\{(R_{d,k,q} = r) : d[k] \in G(q)\}$, of query terms or query-related term occurrences in the document, and these events are expected to be locally relevant to the query q. Another type, $E_2(d, q)$, consists of events, $\{(R_{d,k,q} = \bar{r}) : d[k] \in V(d) - G(q)\}$, of non-query-

related term occurrences in the document, and these events are expected to be locally non-relevant to q according to the generalized query-centric assumption. Using these two

$$p_{\nabla}(R_{d,q}=r_g) \propto p_{\partial}(E_1(d,q),E_2(d,q))$$
$$= p_{\partial}\left(\left[\bigwedge_{t\in G(q)}\bigwedge_{k\in Loc(t,d)}(R_{d,k,q}=r)\right]\wedge\left[\bigwedge_{t\in V(d)-G(q)}\bigwedge_{k\in Loc(t,d)}(R_{d,k,q}=\bar{r})\right]\right).$$

types of events, the probability of relevance in Eq. (1) is as follows
where $\propto$ stands for rank equivalence (Lafferty and Zhai, 2003), $\wedge$ stands for conjunction and $Loc(t, d)$ is the set of locations where term t has occurred in document d. We assume that the events in the previous equation are all mutually (binary) independent (hence the name binary independence language model) in order to simplify that equation as follows:

After making the CLRD assumption, we can assign $p_{\partial,n}(r \mid c(d,k),t,q)$ to $p_{\partial}(R_{d,k,q}=r)$ in the previous equation which is re-arranged as follows:
where $p_{\partial,n}(\cdot)$ is the probability defined over the event space $\Omega_{\partial,n}$. Similarly, the same is

$$p_{\nabla}(R_{d,q}=r_g) \propto \prod_{t\in G(q)\cap V(d)}\prod_{k\in Loc(t,d)} p_{\partial}\left(R_{d,k,q}=r\right)\times \prod_{t\in V(d)-G(q)}\prod_{k\in Loc(t,d)} p_{\partial}\left(R_{d,k,q}=\bar{r}\right).$$
$$p_{\nabla}(R_{d,q}=r_g) \propto \prod_{t\in G(q)\cap V(d)}\prod_{k\in Loc(t,d)} p_{\partial,n}\left(r \mid c(d,k),t,q\right) \prod_{t\in V(d)-G(q)}\prod_{k\in Loc(t,d)} p_{\partial,n}\left(\bar{r} \mid c(d,k),t,q\right), \quad (2)$$

done for non-relevance in Eq. (2).

For the irrelevance decision component model, the probability of non-relevance in Eq. (1) is derived according to the Disjunctive Relevance Decision (DRD) principle in (Kong et al., 2004), which is formulated according to TREC ad hoc evaluation policy (Harman, 2004). Basically, the DRD principle states that if any part of the document is relevant, then the entire document will be relevant. Effectively, the logical form of the DRD principle is $\bigvee_{k=1}^{|d|} R_{d,k,q}$ which is $\overline{\bigwedge_{k=1}^{|d|} \bar{R}_{d,k,q}}$, where $\vee$ stands for disjunction. Its probabilistic version is rank equivalent to the following, $-\sum_{k=1}^{|d|}\log p_{\partial}(R_{d,k,q}=\bar{r})$, where each $\bar{R}_{d,k,q}$ maps to $p_{\partial}(R_{d,k,q}=\bar{r})$. These probabilities of local non-relevance are partitioned into two groups as follows by the generalized query centric assumption: one group for terms in $G(q)$ and the other group for terms not in $G(q)$, as follows.

$$-\log p_{\nabla}\left(R_{d,q}=\bar{r}_g\right) = -\sum_{t\in G(q)\cap V(d)}\sum_{k\in Loc(t,d)}\log p_{\partial}\left(R_{d,k,q}=\bar{r}\right) - \sum_{t\in V(d)-G(q)}\sum_{k\in Loc(t,d)}\log p_{\partial}\left(R_{d,k,q}=\bar{r}\right).$$

The above equation has a negative sign because the probability on the left of the equation is the denominator of the log-odds ratio in Eq. (1).

4. Context Modeling of Binary Independence Language Model

Section 4.1 discusses the context dependent modeling. As there are different types of terms related to the query, several types of contexts are introduced in Section 4.2 and

applied to the ranking formula. Extrapolating from the bigram terms for contexts, linked dependencies are introduced in Section 4.3 to determine the probability of the contexts.

4.1. *Context Dependent Model Derivation*

Assuming that the CLRD assumption is true, we assign $p_{\partial,n}(\bar{r} \mid c(d,k),t,q)$ to $p_{\partial}(R_{d,k,q} = \bar{r})$ in the last equation of Section 3.2, which becomes

$$-\log p_{\nabla}(R_{d,q} = \bar{r}_g) = -\sum_{t \in G(q) \cap V(d)} \sum_{k \in Loc(t,d)} \log p_{\partial,n}(\bar{r} \mid c(d,k),t,q,) - \sum_{t \in V(d) - G(q)} \sum_{k \in Loc(t,d)} \log p_{\partial,n}(\bar{r} \mid c(d,k),t,q,). \quad (3)$$

where $p_{\partial,n}(\cdot)$ is a probability defined in the event space, $\Omega_{\partial,n}$. Substituting Eqs. (2) and (3) into the log-odds ratio in Eq. (1), this ratio is rank equivalent to

$$\sum_{t \in G(q) \cap V(d)} f(t,d) \cdot \log \frac{p_{\partial,n}(r \mid t,q)}{p_{\partial,n}(\bar{r} \mid t,q)} + \sum_{t \in G(q) \cap V(d)} \sum_{k \in Loc(t,d)} \log \frac{p_{\partial,n}(c(d,k) \mid t,q,r)}{p_{\partial,n}(c(d,k) \mid t,q,\bar{r})}. \quad (4)$$

where $f(t, d)$ is the occurrence frequency of term t in document d. Eq. (4) consists of two major components. The left component is a by-product of the re-arrangement of the conditional probabilities (i.e., from $p_{\partial,n}(r \mid c(d,k),t,q)$ to $p_{\partial,n}(c(d,k) \mid t,q,r)$, and similarly for $p_{\partial,n}(\bar{r} \mid c(d,k),t,q)$ to become $p_{\partial,n}(c(d,k) \mid t,q,\bar{r})$). The left component may be considered as the product of the term frequency and the log-odds that is similar to w_4 in (Robertson and Sparck Jones, 1976). In here, we assign the probability of a half to both $p_{\partial,n}(r \mid t, q)$ and $p_{\partial,n}(\bar{r} \mid t,q)$ since we are uncertain of the relevance given only the term t and the query q. In this case, the left component in Eq. (4) vanishes after taking the logarithm. The right component is similar to the log-odds ratio of the document-context decision that appears in (Wu et al., 2006). This ratio is the ratio between the probability of document-contexts being relevant against it being irrelevant similar to the right component. The probabilities of this component are computed similar those in the language models, where they are the product of the probabilities of the individual term occurrences. Therefore, we call our model the Binary Independence Language Model (BILM).

In this article, the query terms and their related terms (i.e., $G(q)$) are the union of (1) single query terms (i.e., $S(q)$), (2) coverage terms (i.e., $C(q)$) and (3) expansion terms (i.e., $E(q)$). That is, $G(q) = S(q) \cup C(q) \cup E(q)$. The single query term (i.e., $S(q)$) refers to the original individual query terms of the topic. The coverage term (i.e., $C(q)$) refers to the set of selected terms according to their number of occurrences with the single query terms. For each topic, the coverage terms are selected by the number of occurrences of the term in the contexts of the original query terms in the relevant documents from the top X ranked documents. In other words, the *coverage* of a term means the number of contexts of query terms containing the term. After the *coverage* of all terms occurred in the relevant documents from the top X ranked documents are calculated, top k_{cov} terms are selected. We believe that the higher the *coverage* of a term, the higher is the correlation between the term and the query terms. Lastly, the expansion query term (i.e., $E(q)$) are the terms obtained from the relevant documents from the top X ranked documents

according to the relevance model (RM) (Lavernko and Croft, 2001) for query expansion. Top k_{exp} expansion terms are selected.

4.2. *Different Context Types Derivation*

Given the three sets of terms which are believed to be highly related to the topic, we define five types of contexts according to their middle term of the context; they are (1) contexts with a query term $t \in S(q)$ in the middle, (2) contexts with a query term $t \in S(q)$ in the middle and there is another query term $s \in S(q)$ where $s \neq t$ occurs within a window size W with t, (3) contexts with a query term $t \in S(q)$ in the middle and immediately followed by another query term $s \in S(q)$ where $s \neq t$, (4) contexts with a coverage term $t \in C(q)$ in the middle and (5) contexts with an expansion term $t \in E(q)$ in the middle.

The first three types of contexts have an original query term (i.e., $S(q)$) as the middle term of the context. The second type allows two different original query terms occur within a distance W while the third type requires the two different original query terms to occur as a phrase. To define the second and third types of contexts, we define the locations where such contexts occur as follows. Let $Loc_p(t, q, d)$ returns the set of locations of term t in document d such that there is another term $s \in S(q)$ where $s \neq t$ immediately follows t. That is a 2-term phrase $t \cdot s$ occurred in the following locations:

$$Loc_p(t,q,d) = \{k : 1 \le k \le |d|, d[k] = t, d[k+1] \in S(q), d[k+1] \neq t\}.$$

Let $Loc_w(t, q, d)$ returns the set of locations of term t in document d such that there is another term $s \in S(q)$ where $s \neq t$ occurs with the term t within a distance of W:

$$Loc_w(t,q,d) = \{k : 1 \le k \le |d|, d[k] = t, d[k \pm x] \in S(q), d[k \pm x] \neq t, x \le W\}.$$

From Eq. (4), the right component used in the rank function of BILM is the log-odds ratio of the document-context decision. In practice, we can only obtain an estimate of these probabilities, and we make a weaker assumption that the estimates are only rank equivalent to the actual probabilities as follows:

$$\begin{aligned}\sum_{t \in G(q) \cap V(d)} \sum_{k \in Loc(t,d)} \log \frac{p_{\partial,n}(c(d,k)|t,q,r)}{p_{\partial,n}(c(d,k)|t,q,\bar{r})} \propto & \sum_{t \in S(q) \cap V(d)} \sum_{k \in Loc(t,d)} \log \frac{\hat{p}_{\partial,n,S}(c(d,k)|t,q,r)}{\hat{p}_{\partial,n,S}(c(d,k)|t,q,\bar{r})} + \\ & \sum_{t \in C(q) \cap V(d)} \sum_{k \in Loc(t,d)} \log \frac{\hat{p}_{\partial,n,C}(c(d,k)|t,q,r)}{\hat{p}_{\partial,n,C}(c(d,k)|t,q,\bar{r})} + \\ & \sum_{t \in E(q) \cap V(d)} \sum_{k \in Loc(t,d)} \log \frac{\hat{p}_{\partial,n,E}(c(d,k)|t,q,r)}{\hat{p}_{\partial,n,E}(c(d,k)|t,q,\bar{r})} + \\ & \sum_{t \in S(q) \cap V(d)} \sum_{k \in Loc_w(t,q,d)} \log \frac{\hat{p}_{\partial,n,W}(c(d,k)|t,q,r)}{\hat{p}_{\partial,n,W}(c(d,k)|t,q,\bar{r})} + \\ & \sum_{t \in S(q) \cap V(d)} \sum_{k \in Loc_p(t,q,d)} \log \frac{\hat{p}_{\partial,n,P}(c(d,k)|t,q,r)}{\hat{p}_{\partial,n,P}(c(d,k)|t,q,\bar{r})} \quad (5).\end{aligned}$$

The right hand side of the above equation corresponds are summing for all the five different context types.

When the context of an expansion term has less than three different expansion terms nor does it have a query term, this context is ignored because it is assumed to be not related to the query. Eq. (5) is used in retrieval for ranking documents. There are 5 components on the right hand side as $G(q) = S(q) \cup C(q) \cup E(q)$. For $S(q)$, it is further divided into single query term, query terms occurs in proximity and query terms occurs in a phrase.

We generalize $\hat{p}_{\partial,n,S}(\cdot)$, $\hat{p}_{\partial,n,C}(\cdot)$ and $\hat{p}_{\partial,n,E}(\cdot)$ to $\hat{p}_{\partial,n,G}(\cdot)$. If $t \in S(q)$, $\hat{p}_{\partial,n,G}(\cdot) = \hat{p}_{\partial,n,S}(\cdot)$. Similarly, if $t \in C(q)$, $\hat{p}_{\partial,n,G}(\cdot) = \hat{p}_{\partial,n,C}(\cdot)$, and if $t \in E(q)$, $\hat{p}_{\partial,n,G}(\cdot) = \hat{p}_{\partial,n,E}(\cdot)$. In the unigram model, the context probabilities are the multiplication of the probabilities of individual context terms by assuming that they are conditionally independent to each other, i.e., we write in a general way that

$$\hat{p}_{\partial,n,G}\big(c(d,k) \mid t,q,r\big) = \prod_{l=1}^{2n+1} \hat{p}_{\partial,n,G}(c(d,k)[l] \mid c[n+1] = t,q,r),$$

$$\hat{p}_{\partial,n,G}\big(c(d,k) \mid t,q,\bar{r}\big) = \prod_{l=1}^{2n+1} \hat{p}_{\partial,n,G}(c(d,k)[l] \mid c[n+1] = t,q,\bar{r}).$$

The other two types of context probabilities (i.e., $\hat{p}_{\partial,n,W}(\cdot)$ and $\hat{p}_{\partial,n,P}(\cdot)$) in Eq. (5) are determined similarly.

4.3. *Link Dependency Modeling*

One way of relaxing the unigram model is the bigram model (Srikanth and Srihari, 2002) which assumes a term is related to its previous term. In 2004, Gao et al. (2004) proposed a new dependence language modeling approach which extended the unigram model by relaxing the independence assumption. They introduced a dependency structure called *linkage* which models the relationship of any two terms. Later, Maisonnasse et al. (2007) refined the dependence model by representing the terms as a graph $J = (C, E)$ where C is the set of terms and E is a binary relation from $C \times C$ in {0, 1}. That is, $E(c_i, c_j) = 1$ if c_i and c_j are related, and 0 otherwise. The dependence model generalizes the bigram model (Srikanth and Srihari, 2002). In our proposed model, we use the linkage relationship between a term and its previous (four) terms for calculating the probabilities:

$$\hat{p}_{\partial,n,G}(c(d,k)\,|\,t,q,r) \approx \prod_{l=1}^{2n+1} \begin{pmatrix} \alpha_{G,0} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c[n+1]=t,q) + \\ \alpha_{G,1} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c[n+1]=t,q,r) + \\ \alpha_{G,2} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-1],c[n+1]=t,q,r) + \\ \alpha_{G,3} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-2],c[n+1]=t,q,r) + \\ \alpha_{G,4} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-3],c[n+1]=t,q,r) + \\ \alpha_{G,5} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-4],c[n+1]=t,q,r) \end{pmatrix}, \quad (6)$$

,

$$\hat{p}_{\partial,n,G}(c(d,k)\,|\,t,q,\bar{r}) \approx \prod_{l=1}^{2n+1} \begin{pmatrix} \alpha_{G,0} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c[n+1]=t,q) + \\ \alpha_{G,1} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c[n+1]=t,q,\bar{r}) + \\ \alpha_{G,2} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-1],c[n+1]=t,q,\bar{r}) + \\ \alpha_{G,3} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-2],c[n+1]=t,q,\bar{r}) + \\ \alpha_{G,4} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-3],c[n+1]=t,q,\bar{r}) + \\ \alpha_{G,5} \times \hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c(d,k)[l-4],c[n+1]=t,q,\bar{r}) \end{pmatrix}, \quad (7)$$

where in general

$$\sum_{j=0}^{5} \alpha_{G,j} = 1.$$

Specifically, when $t \in S(q)$, $\alpha_{G,j} = \alpha_{S,j}$ such that $\sum_{j=0}^{5} \alpha_{S,j} = 1$. Likewise, when $t \in C(q)$, $\alpha_{G,j} = \alpha_{C,j}$ such that $\sum_{j=0}^{5} \alpha_{C,j} = 1$, and when $t \in E(q)$, $\alpha_{G,j} = \alpha_{E,j}$ such that $\sum_{j=0}^{5} \alpha_{E,j} = 1$. The other two types of context probabilities (i.e., $\hat{p}_{\partial,n,W}(\cdot)$ and $\hat{p}_{\partial,n,P}(\cdot)$) are determined similarly.

5. Binary Independence Language Model Estimation

The estimation of the probabilities of the contexts is described in Section 5.1. Finally, to increase the amount of data for probability estimation, a novel bootstrap method is applied to discover more contexts for probability estimation of the context terms.

5.1. *Probability Estimation*

The probability $\hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c[n+1]=t,q)$ in Eqs. (6) and (7) is said to be the collection probability which is used to avoid zero values in the (relative frequency) estimation. This may happen when the conditional relative frequency estimates of a term u is zero. That is the term u does not occur in the contexts of relevant or irrelevant documents during re-ranking. The zero values will propagate to the context probabilities which can cause anomalies in ranking of the documents during retrieval. This is the problem of zero probability similarly found in the language modeling approach (Ponte and Croft, 1998), and smoothing (Chen and Goodman, 1996; Zhai and Lafferty, 2004) of the distribution of terms is a solution to this problem. The basic idea of smoothing is to adjust the distribution of terms so that zero probability will not be assigned to unseen terms. In (Wu et al., 2006), we have tested a similar model using three interpolation-

$$\hat{p}_{\partial,n,G}(c(d,k)[l]\,|\,c[n+1]=t,q) = \left(\frac{\sum_{d \in D} \sum_{k \in Loc(t,d)} f(c(d,k),u)}{\sum_{d \in D} \sum_{k \in Loc(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \right)$$

based smoothing techniques namely additive smoothing (e.g., Lidstone, 1920), Jelinek-Mercer smoothing (Ponte and Croft, 1998; Zhai and Lafferty, 2004) and absolute discounting (Zhai and Lafferty, 2004; Ney et al., 1994), and we found that the performance of the three smoothing techniques is close to each other when the parameters are set appropriately. In this article, we used Jelinek-Mercer smoothing that linearly interpolates with the background probability, the relative frequency estimate of which is:

Using the simplifying notation that u refers to some context term $c(d, k)[l]$, let $f(c(d, k), u)$ be the raw frequency of the term u in the context $c(d, k)$. Let $U = R_X$ and I_X be the set of relevant and irrelevant documents from the top X ranked documents of the initial retrieval list, respectively. The conditional relative frequency estimates of u are:

$$\hat{p}_{freq,P}(u \mid t,q,r) = \frac{\sum_{d \in U} \sum_{k \in Loc_P(t,q,d)} f(c(d,k),u)}{\sum_{d \in U} \sum_{k \in Loc_P(t,q,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \quad (12)$$

$$\hat{p}_{freq,P}(u \mid t,q,\bar{r}) = \frac{\sum_{d \in I_X} \sum_{k \in Loc_P(t,d)} f(c(d,k),u)}{\sum_{d \in I_X} \sum_{k \in Loc_P(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \quad (13)$$

$$\hat{p}_{freq,W}(u \mid t,q,r) = \frac{\sum_{d \in U} \sum_{k \in Loc_W(t,q,d)} f(c(d,k),u)}{\sum_{d \in U} \sum_{k \in Loc_W(t,q,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \quad (10)$$

$$\hat{p}_{freq,W}(u \mid t,q,\bar{r}) = \frac{\sum_{d \in I_X} \sum_{k \in Loc_W(t,d)} f(c(d,k),u)}{\sum_{d \in I_X} \sum_{k \in Loc_W(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \quad (11)$$

$$\hat{p}_{freq,G}(u \mid t,q,r) = \frac{\sum_{d \in U} \sum_{k \in Loc(t,d)} f(c(d,k),u)}{\sum_{d \in U} \sum_{k \in Loc(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \quad (8)$$

$$\hat{p}_{freq,G}(u \mid t,q,\bar{r}) = \frac{\sum_{d \in I_X} \sum_{k \in Loc(t,d)} f(c(d,k),u)}{\sum_{d \in I_X} \sum_{k \in Loc(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v)} \quad (9)$$

Using the simplifying notation that u refers to some context term $c(d, k)[l]$, and $u(h)$ refers to some context term $c(d, k)[l\text{-}h]$, let $f(c(d, k), u, u(h))$ be the raw frequency of the term u in the context $c(d, k)$ such that the term $u(h)$ occurs at h number of terms before u. The probabilities of seeing term u given $u(h)$, t, q and r are:

$$\hat{p}_{\partial,n,G}(u \mid u(h),t,q,r) = \frac{\sum_{d \in R_{20}X} \sum_{k \in Loc(t,d)} f(c(d,k),u,u(h))}{\sum_{d \in R_X} \sum_{k \in Loc(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v,v(h)=u(h))}$$

$$\hat{p}_{\partial,n,G}(u \mid u(h),t,q,\bar{r}) = \frac{\sum_{d \in I_{20}X} \sum_{k \in Loc(t,d)} f(c(d,k),u,u(h))}{\sum_{d \in I_X} \sum_{k \in Loc(t,d)} \sum_{v \in c(d,k)} f(c(d,k),v,v(h)=u(h))}$$

$$\hat{p}_{\partial,n,W}(u \mid u(h),t,q,r) = \frac{\sum_{d \in R_X} \sum_{k \in Loc_W(t,q,d)} f(c(d,k),u,u(h))}{\sum_{d \in R_X} \sum_{k \in Loc_W(t,q,d)} \sum_{v \in c(d,k)} f(c(d,k),v,v(h)=u(h))}$$

$$\hat{p}_{\partial,n,W}(u \mid u(h),t,q,\bar{r}) = \frac{\sum_{d \in I_X} \sum_{k \in Loc_W(t,q,d)} f(c(d,k),u,u(h))}{\sum_{d \in I_X} \sum_{k \in Loc_W(t,q,d)} \sum_{v \in c(d,k)} f(c(d,k),v,v(h)=u(h))}$$

$$\hat{p}_{\partial,n,P}(u \mid u(h),t,q,r) = \frac{\sum_{d \in R_X} \sum_{k \in Loc_P(t,q,d)} f(c(d,k),u,u(h))}{\sum_{d \in R_X} \sum_{k \in Loc_P(t,q,d)} \sum_{v \in c(d,k)} f(c(d,k),v,v(h)=u(h))}$$

$$\hat{p}_{\partial,n,P}(u \mid u(h),t,q,\bar{r}) = \frac{\sum_{d \in I_X} \sum_{k \in Loc_W(t,q,d)} f(c(d,k),u,u(h))}{\sum_{d \in I_X} \sum_{k \in Loc_W(t,q,d)} \sum_{v \in c(d,k)} f(c(d,k),v,v(h)=u(h))}$$

When estimating the irrelevance probability, we make use of the bottom end documents. The *IrlBotStart* parameter controls the number of bottom end documents used. For documents ranked below *IrlBotStart*, the contexts of these documents are treated as irrelevant and its terms are added to the irrelevance model. Since the number of contexts in the bottom end documents is greater than the number of contexts in the judged irrelevant documents from the top X ranked documents, we weight the frequency

count of terms in the contexts of the bottom end documents by $IrlBotWeight \in [0,1]$ for irrelevance probability estimation so that these frequency counts will not overshadow the frequency counts from the contexts in the judged irrelevant documents from the top X ranked documents. As a result, the number of training data for the irrelevance model will not be too small.

5.2. *Increasing Amount of Data for Estimation by Bootstrapping*

When the number of relevant contexts of a term $t \in G(q)$ is too small, the relative frequency estimate, $\hat{p}_{freq,G}(u \mid t,q,r)$, will be inaccurate. In order to solve this problem, we bootstrap using the contexts of the term t in the unjudged documents where such contexts are similar to the contexts of t in the judged relevant documents. The similarity of contexts is calculated using log-odds:

$$\log p(c(d,k) \mid t,q,r) - \log p(c(d,k) \mid t,q,\bar{r}).$$

These contexts in the unjudged documents are ranked by this log-odd score, and their top T% is also considered as relevant contexts of t for raw frequency counting (i.e., $f(c(d, k), u)$ and $f(c(d, k), v)$) when the number of relevant contexts of a term $t \in G(q)$ is below a threshold, *relCon*. That is $U \neq R_X$ in Eqs. (8), (10) and (12) since U is now the set of documents which is the union of the judged relevant documents and the unjudged documents with the highest T% log-odds scores.

When there is no relevant document in the top X ranked documents, the best performing parameter values are quite different from the ones when there are relevant documents. Therefore, we use two sets of parameter values: one set calibrated when there is at least a relevant document in the top X ranked documents and another set calibrated when there is no relevant document in the top X.

Table 1 lists the parameters used in our proposed model and their range of values in the experiments during calibration. The calibration involves setting some of the parameters with different values and observing which parameter value gives the best results. Then, the next iteration will set the other parameter values to observe for better performance. Each setting involves stepping through different values in the given range. For example, the context size, C, may be set to 31, 51, 71 or 91 to observe the different performance within the given range between 11 and 101 (in Table 1). Although there is no guarantee that the global optimal setting is discovered, it is hope that a good enough local optimal setting is found as it is very difficult to find the global optimal setting.

Table 1: Parameters used in the model

Parameters	Description	Range
X	Number of documents used for RF	20
C	Context size	11 - 101
W	Window size for query terms occur in proximity	5 - 15
k_{cov}	Number of coverage terms	10 - 150
k_{exp}	Number of expansion terms	10 - 150

$\alpha_{S,j}$ (j =0, 1, 2, 3, 4, 5)	Mixture parameters for $S(q)$	0.1 – 0.9
$\alpha_{C,j}$ (j =0, 1, 2, 3, 4, 5)	Mixture parameters for $C(q)$	0.1 – 0.9
$\alpha_{E,j}$ (j =0, 1, 2, 3, 4, 5)	Mixture parameters for $E(q)$	0.1 – 0.9
IrlBotStart	Documents ranked below the parameter are considered to be non-relevant	1,000 – 100,000
IrlBotWeight	Frequency count of terms which occur in documents ranked below *IrlBotStart*	0.1 – 0.9
T (in percentage)	Parameter used in bootstrapping	10% – 50%
relCon	Threshold for the number of contexts in order to perform bootstrapping	10 – 1000

6. Experiments

We performed two sets of experiments. One set is relevance feedback (RF) experiments which use the top 20, judged documents (i.e., X = 20) from the initial retrieval for training. Another set is a retrospective experiment which uses all the judged documents for training. This is done to get an upper bound of the performance of the model for comparison.

6.1. *Experimental Set Up*

The proposed model is trained using the TREC-2005 ad-hoc retrieval text collection and we perform experiments on TREC-6, -7, -8 and -2005 (heterogeneous) collections using calibrated, fixed parameter values. Table 2 shows some information of the various collections. TREC-7 and TREC-8 use the same text collection which is a subset of the TREC-6 text collection. In particular, TREC-6 contains articles of the Congressional Records which are relatively long compared with news articles and which do not appear in TREC-7 or TREC-8. We want to use these collections, which are smaller than the terabyte data sets like GOV2 or Clueweb09, because it is harder to show improvement for proximity matching model like ours for smaller collections than for larger collections according to Buttcher et al. (2006). Therefore, we used harder collections for our work to claim better performance than non-proximity matching models like language models. Title queries are used in the initial retrieval which is performed using the query likelihood (QL) model (Lafferty and Zhai, 2001) of the Indri retrieval system (available at http://www.lemurproject.org/indri/ and details of which from Strohman et al., 2004). The mean average precisions (MAPs) of the initial retrievals are shown in Table 2. The MAPs are reported because they are relatively stable measures of performance. Top 20 documents from the initial retrieval list are used for relevance feedback. Top 20 documents are used as implicit feedback can easily provide such information. The relevance judgements are from the TREC judgement files for the corresponding collections.

Table 2: Collections used in the experiments

	TREC-6	TREC-7	TREC-8	TREC-2005
No. of documents	556,077	528,155	528,155	1,033,461
Topics	301-350	351-400	401-450	50 past hard topics
Storage (GB)	3.3	3.0	3.0	5.3
Mean Average Precision	0.247	0.200	0.253	0.263

We compare our results with those produced by the support vector machine (SVM) using the SVM_Light package (Joachims, 1999) which can be downloaded at http://svmlight.joachims.org. SVM is used because it is among the best for learning-to-rank and it is a common baseline. After testing on TREC-2005, we use the radial basis kernel function for SVM because it is found to be more effective compared with other kernel functions. We also compare our results with the combination of query expansion (RM3) algorithm (Lavrenko and Croft, 2001) with Markov random field modeling (MRF) (Metzler and Croft, 2005) as in (Lease, 2008) which produced the best results in the relevance feedback track in TREC-2008. The model-based feedback in language model (LM) approach to information retrieval (Zhai and Lafferty, 2001) which uses the expectation-maximization (EM) algorithm is also compared. We produced the residue result, of which the judged documents are removed from the judgement list when calculating the performance measures such as the mean average precision (MAP).

The calibration of parameters is performed on TREC-2005 and the same parameter values are used for all the test collections. Table 1 shows the range of values of the parameters during calibration. The calibration involves a grid search that explores the retrieval performance (as the vertical dimension of the grid) controlled by a set of parameters (as the other horizontal dimensions of the grid) with values taken at regular intervals forming a grid to find the best set of parameter values. Basically, the grid search steps through the possible parameter values within the given ranges in Table 1. The highest performance for a particular set of parameter values in the grid search is used as the result to further the search for better parameter values using finer steps of smaller regular intervals. Typically, the grid search iterates three times before arriving at the final calibrated parameter values.

6.2. *Relevance Feedback Experiments*

Table 3 shows the results in terms of the mean average precision (MAP) of our model, SVM, the modified MRF and LM. All the four methods use the same amount of relevance information which is the top 20 judged documents from the initial retrieval list. Note that the MAPs are obtained for the residue collection and not the entire collection, so that the retrieval by these algorithms is more difficult than retrieving the entire document collections as the easy-to-identify relevant documents have been retrieved and judged. For our model and SVM, they use both relevant and nonrelevant documents from the top 20 during training. However, for LM and the modified MRF algorithm, only

relevant documents from the top 20 are considered. From the results, our model performed significantly better than the effective SVM, LM and the highly effective, modified MRF model with a 95% confidence interval (C.I). This is achieved for TREC-6, TREC-7 and TREC-8 test collections using fixed parameter values that are calibrated by the TREC-2005 retrieval performance. This demonstrates that our model is effective.

Table 3: MAP of our BILM, SVM, the modified MRF and Language Model (LM)

TREC	LM	MRF	SVM	Ours
6	.224	.229	.216	.252$^{*+\$}$
7	.241	.247*	.236^{+}	.278$^{*+\$}$
8	.239	.248*	.228^{+}	.273$^{*+\$}$
2005	.306	.318	.310	.345$^{*+\$}$

* - The result compared with SVM is statistically significantly different with a 95% C.I.

\+ - The result compared with MRF is statistically significantly different with a 95% C.I.

$ - The result compared with LM is statistically significantly different with a 95% C.I.

Note that in initial retrieval and pseudo-relevance feedback, the MAP performance of TREC-6 is usually higher than TREC-7 for the top TREC performers (e.g., Walker et al, 1998). This is because apart from document length normalization, the top TREC performers also used passage-based retrieval to handle widely varying document lengths. In our experiments, the MAP performance of TREC-6 is lower than TREC-7 because we did not have passage-based retrieval to combat the effect of document length wide variation. Therefore, the extra long documents in TREC-6 which do not appear in TREC-7 document collection have an impact on retrieval effectiveness. So, it is appropriate to treat TREC-6 as a different collection to TREC-7.

The differences of MAPs between our model and the other models in Table 3 are fairly consistent. For example, the MAP of our model is about 3 to 4 percentage points higher than LM across all the used test collections. The test collection used for calibration did not achieve a substantially higher performance when it is compared with the other models. So, the higher MAP performance can be generalized to other test collections. Note that MRF and SVM were able to perform statistically significantly better than LM only for TREC-7 and TREC-8 test collections but our model was able to perform statistically significantly better than the other models for all the used test collections.

6.3. *Retrospective Experiments*

Retrospective experiments use all the judged documents for training in order to get an upper bound of the performance for comparison. Similar to Wu et al. (2006), these retrospective experiments are used to validate our retrieval models because: (a) the experiments can reveal the potential of the models; (b) they can isolate the problems of the models from those of the parameter estimation; and (c) they can provide information about the major factors affecting the retrieval effectiveness of the models. In the

retrospective experiments here, we use the entire initial retrieval list instead of using top 20 documents from the initial retrieval list for relevance feedback.

Table 4: Retrospective results from our BILM, LM, the modified MRF and SVM

TREC	LM	MRF	SVM	Ours
6	.584	.591*	.813+$	.799+$
7	.537	.562*	.806+$	.775*+$
8	.591	.598*	.793+$	.788+$
2005	.608	.621*	.812+$	.796+$

* - The result compared with SVM is statistically significantly different with a 95% C.I.

\+ - The result compared with MRF is statistically significantly different with a 95% C.I.

$ - The result compared with LM is statistically significantly different with a 95% C.I.

Table 4 shows the results of the retrospective experiments using our model, the LM model using EM algorithm, the modified MRF model and SVM. From the results, we can see that SVM on average outperforms our model and the MRF model in retrospective experiments for all four collections tested. SVM performs statistically significantly better than MRF in all collections tested with 95% C.I. When compared with our model, only TREC-7 is statistically significantly better for SVM whereas the other three test collections are not statistically significant. Good SVM performance is probably due to the fact that SVM optimizes its performance for each query in each of the collections whereas our model, the language model and the MRF model are calibrated using TREC-2005 and are tested on the four collections using the same parameter values. Our model outperforms the highly effective language model and MRF model statistically significantly in the four TREC collections with a 95% C.I.

6.4. *Precision-Oriented Results*

Table 5 shows the precision-oriented results for LM, MRF, SVM and our BILM in both the relevance feedback experiments and retrospective experiments. The precision-oriented results are based on the precision for the top 10 documents (i.e., P@10). Only the top 10 documents are measured because there are generally more than 10 relevant documents per query to ensure that P@10 can reach 100% for each query. The language model results are presented in Table 5 and we observe that they are generally lower than others, because the MRF results can serve as an upper bound performance for the language model as the ranking formula for MRF is the same as the language model with the additional interpolated proximity matching components.

Table 5: Precision-oriented results based on precision at top 10 (P@10) of our proposed model (BILM), SVM, (modified) MRF and Language Model (LM)

	Relevance Feedback Experiments				Retrospective Experiments			
TREC	LM	MRF	SVM	Ours	LM	MRF	SVM	Ours
6	.376	.414	.405	.426	.700	.816	.930	.922

7	.406	.472	.468	.477	.782	.863	.986	.896
8	.454	.480	.475	.486	.770	.859	.992	.884
2005	.568	.576	.566	.581	.844	.875	.992	.912

For relevance feedback experiment, the P@10 performance of our BILM is the best compared with LM, MRF and SVM. However, there is not much difference between our BILM and others, so no statistical significant differences are shown. However, since our BILM numerically is better than the corresponding performance of LM, SVM and MRF for all the tested collections, we are more confident that our BILM is the best compared with LM, MRF and SVM for the relevance feedback experiment.

For retrospective experiments, the P@10 performance of SVM is the best numerically, followed by our BILM, then by MRF and lastly by LM. Since the top 10 precisions are not very stable, we cannot find statistical significant differences between the results of LM, MRF, SVM or our BILM for the retrospective experiments. However, since for all the tested collections, SVM is better than ours numerically, we believe that SVM is better as it has optimizes its performance. Similarly, our BILM is better numerically than MRF and LM for all the tested collections, therefore we are more confident that our BILM is better than MRF and LM.

If we combine all the relevance feedback and retrospective experiments as 8 samples differing by some factors (like query factor or document collection factor), then assuming (1) in our random model that the probability that the performance is better than the other is 0.5, and (2) testing based on the paired sign tests, SVM and our model has no statistical significance, and our model (BILM) compared with MRF and LM is statistically significant with a *p*-value of 0.004 or with a confidence level of 99.6%. In this way, we may conclude that our model (BILM) is numerically better than MRF and LM with statistical significance with a confidence level of at least 95%. While we can claim a confidence level of 99%, we feel that the random model is very coarse, and that it is more consistent to claim statistical significance of 95% for the confidence level which is the same as that of the MAP case.

7. Conclusion

This article presents a new type of language model, called the binary independence language model (BILM), which integrates two document-context based language models into one using the log-odds ratio. Each of these two models incorporates link dependencies and multiple query term dependencies. The probabilities of these models are estimated by smoothing the relative frequency estimate with the background probabilities. We evaluated BILM against other highly effective retrieval models (i.e., support vector machine (SVM), modified Markov random field model and language model) in a simulated relevance feedback environment across four TREC collections. We observe that mean average precision (MAP) of the BILM was statistically significantly better than the MAP of the other highly effective, competing retrieval models at 95% confidence level across all TREC collections using fixed parameter values. This

demonstrates that BILM is highly effective. We also evaluated BILM in a retrospective study. The MAP of the BILM was statistically significantly better than the MAP of the language model and the modified Markov random field model across all the tested TREC collections. Although the MAP of BILM was statistically significantly lower than SVM for the TREC-7 collection, the MAP of BILM was not statistically significantly lower than SVM for the other TREC collections. For precision at the top 10 (P@10) documents, we cannot find any statistical significant differences between LM, MRF, SVM and our BILM. This may be due to the fact that the P@10 measure is not very stable unlike MAP which makes the measurement from the top 1000 documents instead of just top 10 documents. However, we are more confident that our BILM is better than LM and MRF for the relevance feedback experiments because the P@10 performance of our BILM is numerically better than the corresponding P@10 performance of LM and MRF for all the tested collections in both relevance feedback and retrospective experiments with a confidence level of 95% based on a paired sign test.

Our future work includes applying this model to pseudo-relevance feedback and the initial retrieval rather than relevance feedback. In this case, we may assume that the top documents are relevant and the bottom ranked documents are non-relevant. Our hope is that the estimation needs not be accurate to produce effective retrieval similar to other proximity matching models. We may also work on cheaper estimation methods for the probability values. Such estimation may be approximations rather than accurate estimation that guarantee certain statistical properties. Our hope is that the estimation needs not be very accurate to produce effective retrieval.

Acknowledgments

This work is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 5259/09E).

References

References are to be listed in the order cited in the text in Arabic numerals. They can be typed after punctuation marks, e.g. "…in the statement [2]." Or used directly, e.g. "see [6] for examples." Please list using the style shown in the following examples. For journal names, use the standard abbreviations. Typeset references in 9 pt Times Roman with line spacing of 11pt.

References

[1] K.-D. Althoff, A. Birk, S. Hartkopf, and W. Müller, Managing software engineering experience for comprehensive reuse, in *Proc. 11th Int. Conf. on Software Engineering*, Kaiserslautern, Germany, 1999.

[2] K.-D. Althoff, A. Birk, and C. Tautz, The experience factory approach: Realizing learning from experience in software development organizations, in *Proc. 10th German Workshop on Machine Learning* (*FGML'97*), University of Karlsruhe, 1997, pp. 6–8.

[3] V. R. Basili, G. Caldiera, and H. D. Rombach, The experience factory, in *Encyclopedia of Software Engineering*, ed. J. J. Marciniak, (John Wiley & Sons, 1994), pp. 469–476.
[4] V. R. Basili and H. D. Rombach, The TAME project: Towards improvement-oriented software environments, *IEEE Trans. on Software Engineering* **SE-14**(6) (1988) 758–773.
[5] U. Becker-Kornstaedt and R. Webby, A Comprehensive Schema Integrating Software Process Modelling and Software Measurement, Fraunhofer IESE-Report No. 047.99 (Ed.: Fraunhofer IESE, 1999), http://www.iese.fhg.de/Publications/Iese_reports/.
[6] R. Bergmann and U. Eisenecker, Case-based reasoning for supporting reuse of object-oriented software: A case study, in *Proc. Expert Systems* **95** (1996) 152–169.
[7] D. N. Card and R. L. Glass, *Measuring Software Design Quality* (Prentice Hall, Englewood Cliffs, NJ, 1990).
[8] D. Deridder, A concept-oriented approach to support software maintenance and reuse activities, in Proc. Workshop on Knowledge-Based Object-Oriented Software Engineering at 16th European Conference on Object-Oriented Programming (ECOOP 2002) Málaga, Spain, 2002.
[9] C. Estay and J. Pastor, Improving action research in information systems with project management, in *Proc. Americas Conference on Information Systems*, Long Beach, CA, USA, 2000, pp. 1558–1561.
[10] R. A. Falbo, C. S. Menezes, and A. R. Rocha, Using ontologies to improve knowledge integration in software engineering environments, in *Proc. 4th Conference on Information Systems Analysis and Synthesis*, Orlando, Florida, USA, 1998.
[11] Azzopardi, L. 2007. Explicitly considering relevance within the language modeling framework. In *Proceedings of the International Conference in Theory of Information Retrieval*, pp. 125-134.
[12] Büttcher, S., Clarke, C.L.A. & Lushman, B. 2006. Term proximity scoring for ad-hoc retrieval on very large text collections. In Proceedings of ACM SIGIR 06, pp. 621-622.
[13] Chen, S.F. & Goodman, J. 1996. An Empirical Study of Smoothing Techniques for Language Modelling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pp. 310-318.
[14] Diaz, F. 2008. Improving relevance feedback in language modeling with score regularization. In *Proceedings of ACM SIGIR 08*, pp.807-808.
[15] Dang, E.K.F., Luk, R.W.P., Allan, J., Ho, K.S., Chung, F.L., Lee, D.L. & Chan, S.C.F. 2010. A new context-dependent term weight computed by boost and discount using relevance information. *Journal of the American Society for Information Science and Technology* 61, 12, 2514-2530.
[16] Dang, E.K.F., Luk, R.W.P. & Allan, J. 2014. Beyond bag-of-words: bigram-enhanced context-dependent term weights. *Journal of the American Society for Information Science and Technology* 65, 6, 1134-1148.
[17] Dang, E.K.F., Luk, R.W.P. & Allan, J. 2016. A context-dependent relevance model. *Journal of the Association for Information Science and Technology* 67, 3, 582-593.
[18] Dang, E.K.F., Wu, H.C., Luk, R.W.P. & Wong, K.F. 2009. Building a framework for the probability ranking principle by a family of expected weighted rank. *ACM Transactions on Information Systems*, 27, 4.
[19] Fuhr, N. 1992. Probabilistic models in information retrieval. *The Computer Journal*, 35, 3, 243-255.
[20] Gao, J., Nie, J.Y., Wu, G., & Cao, G. 2004. Dependence language model for information retrieval. In *Proceedings of ACM SIGIR 04*, pp. 170 - 177.
[21] Harman, D. 2004. *Private communication*. (at NTCIR-4).
[22] Hiemstra, D. 1998. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of European Conference on Digital Libraries*, pp. 569-584.

[23] Huston, S. & Croft, W.B. 2014. A comparison of retrieval model using term dependencies. In *Proceedings of ACM CIKM '14*, pp. 111-120.
[24] Jelinek, F. & Mercer, R. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pp. 381-402.
[25] Joachims, T. 1999. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods, Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press.
[26] Kolmogorov, A.N. 1950. *Foundations of the Theory of Probability*. New York: Chelsea.
[27] Kong, Y.K., Luk, R.W.P., Lam, W., Ho, K.S. & Chung, F.L. 2004. Passage-based retrieval based on parameterized fuzzy operators. In *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval*.
[28] Lafferty, J. & Zhai, C.X.. 2001. Document language models, query models and risk minimization for information retrieval. In *Proceedings of ACM SIGIR 01*, pp. 111-119.
[29] Lafferty, J. & Zhai, C.X. 2003. Probabilistic relevance models based on document and query generation. In Croft, W.B. and Lafferty, J. *Language Modeling for Information Retrieval*, pp. 1- 10.
[30] Lavrenko, V. & Croft, W.B. 2001. Relevance-based language model. In *Proceedings of ACM SIGIR 01*, pp. 120-127.
[31] Lease, M. 2008. Incorporating Relevance and Pseudo-relevance Feedback in the Markov Random Field Model. *TREC-2008*.
[32] Li, J. & Yan, H. 2006. Peking University at the TREC 2006 terabyte track. In *Proceedings of the Fifteenth Text Retrieval Conference*.
[33] Lidstone, G.J. 1920. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8, 182-192.
[34] Lv, Y. & Zhai, C.X. 2009. Positional language models for information retrieval. In *Proceedings of ACM SIGIR 09*, pp. 299-306.
[35] Lv, Y. & Zhai, C.X. 2010. Positional relevance model for pseudo-relevance feedback. In *Proceedings of ACM SIGIR 10*, pp. 579-586.
[36] Lv, Y. & Zhai, C.X. 2015. Negative query generation: bridging the gap between query likelihood retrieval models and relevance. Information Retrieval Journal 18, 4, 359-378.
[37] Luk, R.W.P. 2008. On event space and rank equivalence between probabilistic retrieval models. *Information Retrieval*, 11, 6, 539-561.
[38] Maisonnasse, L., Gaussier, E. & Chevallet, J.P. 2007. Revisiting the dependence language model for information retrieval. In *Proceedings of ACM SIGIR 07*, pp. 695 - 696.
[39] Manning, C.D., Raghavan, P. and Schutze, H. 2008. *Introduction to Information Retrieval*. Cambrdige University Press, Cambridge, UK.
[40] Metzler, D. & Croft, W.B. 2005. A Markov random field model for term dependencies. In *Proceedings of ACM SIGIR 05*, pp. 472-479.
[41] Metzler, D. & Croft, W.B. 2007. Latent concept expansion using Markov random fields. In *Proceedings of ACM SIGIR 07*, pp. 311-318.
[42] Ney, H., Essen, U. & Kenser, R. 1994. On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language*, 8, 1–38.
[43] Peng, T. & Liu, L. 2015. Clustering-based topical web crawling for topic-specific information retrieval guided by incremental classifier. International Journal of Software Engineering and Knowledge Engineeering, 25, 1, 147-168.
[44] Pickens, J. & MacFarlane, A. 2006. Term context models for information retrieval. In *Proceedings of ACM CIKM 06*, pp. 559-566.
[45] Ponte, J. & Croft, W.B. 1998. A language modeling approach in information retrieval. In *Proceedings of ACM SIGIR 98*, pp. 275-281.
[46] Rasolofo, Y. and Savoy, J. 2003. Term proximity scoring for keyword-based retrieval systems. In *Proceedings of ECIR '03*, pp. 207-218.

[47] Robertson, S.E. 1977. The probability ranking principle in IR. *Journal of Documentation* 33, 4, 294-304.

[48] Robertson, S.E. 2005. On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8, 319-329.

[49] Robertson, S.E. & Sparck Jones, K. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 3, 129-146.

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

[51] Srikanth, M. & Srihari, R. 2002. Biterm language models for document retrieval. In *Proceedings of ACM SIGIR 02*, pp. 425 - 426.

[52] Song, F. & Croft, W.B. 1999. A general language model for information retrieval. In *Proceedings of ACM SIGIR 99*, pp. 279-280.

[53] Song, R., Yum L., Wen, J.-R. and Hon, H-W. 2016. A proximity probabilistic model for information retrieval.

[54] Strohman, T., Metzler, D., Turtle, H. & Croft, W. 2004. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*.

[55] Vechtomova, O. & Robertson, S.E. 2000. Integration of collocation statistics into the probabilistic retrieval model. In *Proceedings of the 22nd British Computer Society - Information Retrieval Specialist Group Conference*, pp. 165-177.

[56] Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F. and Sparck Jones, K. 1998. Okapi at TREC-6: automatic ad hoc, vlc, routing, filtering and QSDR. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*. NIST Special Publication.

[57] Wang, X., Fang, H. and Zhai, C.X. 2007. Improve retrieval accuracy for difficult queries using negative feedback. In *Proceedings of ACM CIKM 07*, pp. 991-994.

[58] Wang, X., Fang, H. and Zhai, C.X. 2008. A study of methods for negative relevance feedback. In *Proceedings of ACM SIGIR 08*, pp. 219-226.

[59] Wu, H.C., Luk, R.W.P., Wong, K.F. & Kwok, K.L. 2006. A retrospective study of a hybrid document-context based retrieval model. *Information Processing & Management*, 43, 5, 1308-1331.

[60] Wu, H.C., Luk, R.W.P., Wong, K.F. & Kwok, K.L. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26, 3.

[61] Zhai, C.X. & Lafferty, J. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of ACM CIKM 01*, pp. 403-410.

[62] Zhai, C.X. & Lafferty, J. 2003. A risk minimization framework for information retrieval. In *Proceedings of the ACM SIGIR 2003 Workshop on Mathematical/Formal Methods in IR*.

[63] Zhai, C.X. & Lafferty, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22, 2, 179-214.

[64] Zhao, F., Fang, F., Yan F, Jin H. and Zhang, Q. 2012. Expanding approach to information retrieval using semantic similarity analysis based on wordnet and Wikipedia. International Journal of Software Engineering and Knowledge Engineering 22, 2, 305-322.