Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management xxx (2007) xxx–xxx
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

A new robust relevance model in the language model framework

Xiaoyan Li *

Department of Computer Science, Mount Holyoke College, 50 College Street, South Hadley, MA 01075, USA

Received 14 January 2007; received in revised form 11 July 2007; accepted 13 July 2007

Abstract

In this paper, a new robust relevance model is proposed that can be applied to both pseudo and true relevance feedback in the language-modeling framework for document retrieval. There are at least three main differences between our new relevance model and other relevance models. The proposed model brings back the original query into the relevance model by treating it as a short, special document, in addition to a number of top-ranked documents returned from the first round retrieval for pseudo feedback, or a number of relevant documents for true relevance feedback. Second, instead of using a uniform prior as in the original relevance model, documents are assigned with different priors according to their lengths (in terms) and ranks in the first round retrieval. Third, the probability of a term in the relevance model is further adjusted by its probability in a background language model. In both pseudo and true relevance cases, we have compared the performance of our model to that of the two baselines: the original relevance model proposed by Lavrenko and Croft and a linear combination model. Our experimental results show that the proposed new model outperforms both of the two baselines in terms of mean average precision.

Keywords: Relevance models; Language modeling; Feedback; Query expansion

1. Introduction

The language-modeling framework to text retrieval was first introduced by Ponte and Croft (1998). Many research activities related to this framework have been reported since then (Hiemstra, 2001; Lafferty & Zhai, 2001; Lavrenko & Croft, 2001; Li & Croft, 2003; Liu & Croft, 2002; Miller, Leek, & Schwartz, 1999; Ponte, 1998; Song & Croft, 1999; Tao & Zhai, 2004; Zhai & Lafferty, 2001). For example, query expansion and feedback techniques (Hiemstra, 2001; Lavrenko & Croft, 2001; Miller et al., 1999; Ponte, 1998; Tao & Zhai, 2004; Zhai & Lafferty, 2001), parameter estimation methods (Lafferty & Zhai, 2001), multi-word features (Song & Croft, 1999), passage segmentations (Liu & Croft, 2002) and time constraints (Li & Croft, 2003) have been proposed to the language-modeling frameworks. Among them, query expansion with pseudo feedback can

* Tel.: +1 413 538 2554; fax: +1 413 538 3431.
E-mail address: xli@mtholyoke.edu

0306-4573/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2007.07.005

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

increase retrieval performance significantly (Lavrenko & Croft, 2001; Ponte, 1998; Zhai & Lafferty, 2001). It assumes a few top-ranked documents retrieved with ~~the~~ original query to be relevant and uses them to generate a richer query model.

However, there are two major problems that are unsolved in query expansion techniques. First, the performance of a significant number of queries will decrease when query expansion techniques are applied. Query expansion techniques are not guaranteed to work on every query, even though they usually can achieve better performance than merely using the query when measuring the mean average precision on a set of queries. Second, existing query expansion techniques are very sensitive to the number of documents used for pseudo feedback. Most approaches usually achieved the best performance when about 30 documents are used for pseudo feedback. As the number of feedback documents increases beyond 30, retrieval performance drops quickly.

Therefore, a more robust approach to query expansion in the language-modeling framework is needed. Based on the original relevance model approach by Lavrenko and Croft (2001), we propose a new relevance-based language model that improve robustness, and can be applied to both pseudo feedback and true relevance feedback. In the case of pseudo feedback for retrieval, a few of the top-ranked documents that have been initially retrieved are assumed relevant thus were used to estimate the relevance model for a query. In the case of true relevance feedback, a number of known relevant documents were used to estimate the relevance model for a query. The purpose of the experiments for true relevance feedback is to study how the proposed model behaves when more true relevant documents are given for relevance model approximation.

There are three main mechanisms in our new relevance model to improve the robustness of a relevance-based language model: *treating the query as a special document*, *introducing document-rank-related priors,* and *discounting common words*. First, the proposed model brings back the original query into the relevance model by treating it as a short, special document, in addition to a number of top-ranked documents returned from the first round retrieval for pseudo feedback, or a number of relevant documents for true relevance feedback. Second, instead of using a uniform prior as in the original relevance model, documents are assigned with different priors according to their lengths (in terms) and ranks in the first round retrieval. Third, the probability of a term in the relevance model is further adjusted by its probability in a background language model.

We have carried out experiments for both pseudo feedback and true relevance feedback to compare the performance of our model to that of the two baselines: the original relevance model (Lavrenko & Croft, 2001) and a linear combination model (Abdul-Jaleel et al., 2004). Queries on three data sets have been used:

(1) TREC title queries 101–200 on AP collections;
(2) Queries 301–400 on a heterogeneous collection which includes all data from TREC disk 4 and 5; and
(3) Queries 701–750 on the TREC terabyte collection.

In all of the three sets of experiments, the proposed new model outperforms both of the two baselines. Furthermore, the new approach is less sensitive to the number of pseudo feedback documents than the two baseline models, and it requires fewer relevant documents to achieve good performance with true relevance feedback.

The incorporation of the proposed three mechanisms was first described in a technical report (Li, 2005). We note that a very related, recent work by Tao and Zhai (2006) also considered these three aspects in improving the robustness of pseudo-relevance feedback, but using a different implementation based on the EM algorithm, and only for the case of pseudo-relevant feedback. In this paper, we also consider the case of true relevance feedback, and further provide a component analysis to show different roles of the three mechanisms. We also provide a comparison of the robustness between our approach and their approach, which indicating that ours is slightly better in robustness.

The rest of the paper is structured as follows. In Section 2, we briefly introduce relevance-based language models and then a simple variation of relevance models. Our method of constructing a new robust relevance model and a theoretic justification are described in Section 3. Section 4 provides experimental results in comparing the new relevance model to two baselines based on experimental results with TREC queries. An analysis of the components of the new relevance model is given in Section 5. Finally, Section 6 summarizes the paper with conclusions and future work.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

2. Related work: baseline approaches

Our new relevance model is based on the relevance-based language model proposed by Lavrenko and Croft (2001). Therefore, before we introduce the new robust relevance model, we will briefly describe the relevance-based language model, referred as "original relevance model" in the rest of this paper. Then, a slight variation of the relevance models proposed by Abdul-Jaleel et al. (2004) that linearly combines the query and the relevance model is described.

2.1. Original relevance model

The relevance-based language model was proposed by Lavrenko and Croft (2001). It is a model-based query expansion approach in the language-modeling framework (Ponte & Croft, 1998). A relevance model is a distribution of words in the relevant class for a query. Both the query and its relevant documents are treated as random samples from an underlying relevance model R.

The main challenge for a relevance-based language model is how to estimate its relevance model with no relevant documents available but only queries. The basic idea in Lavrenko and Croft (2001) can be viewed as a query expansion based on the top-ranked documents retrieved. Eqs. (1) and (2) are the formulas used in their paper for approximating a relevance model for a query:

$$P_o(w|R) \approx \frac{P(w, q_1 \dots q_k)}{P(q_1 \dots q_k)} \tag{1}$$

$$P(w, q_1 \dots q_k) = \sum_{D \in M} P(D) P(w|D) \prod_{i=1}^{k} (P(q_i|D) \tag{2}$$

where $P_o(w|R)$ stands for this original relevance model of the query and its relevant documents, in which $P(w,q_1 \dots q_k)$ stands for the total probability of observing the word w together with query words $q_1 \dots q_k$. A number of top-ranked documents (say N) returned with a query likelihood language model are used to estimate the relevance model. In Eq. (2) M is the set of the N top-ranked documents used for estimating the relevance model for a query. $P(D)$ is the prior probability to select the corresponding document language model D for generating the total probability in Eq. (2). In the original relevance model approach, a uniform distribution was used for the prior. Once the relevance model is estimated, the KL-divergence between the relevance model (of a query and its relevant documents) and the language model of a document can be used to rank the document. Documents with smaller divergence are considered more relevant thus have higher ranks.

2.2. Linear combination relevance model

The original relevance model does not work well on every query, though on average it significantly outperforms the basic query likelihood language model (Lavrenko & Croft, 2003). The performance of some queries may be hurt badly by using the relevance model, when compared to using the query solely in the query likelihood language model. For such a query, putting the original query back into the relevance model may help. A simple approach to bring the original query back into its relevance model is to linearly combine the query with the relevance model, as in Eq. (3), which was used by Abdul-Jaleel et al. (2004) in their work for the 2004 TREC HARD Track:

$$P_{lc}(w|R) = \lambda P(w|Q) + (1 - \lambda) P_o(w|R) \tag{3}$$

In Eq. (3), $P_{lc}(w|R)$ stands for the relevance model obtained by linearly combining the query with the original relevance model. $P(w|Q)$ stands for the original query model that may be calculated by the maximum likelihood estimation in the experiments. $P_o(w|R)$ stands for the original relevance model described in Eq. (1). The weighting parameter λ is used for linearly combining the query and the relevance model. The best value of λ learned with the training data is 0.05, which will be used in our experiments reported in Section 4.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

3. The new relevance model

Based on the original relevance model approach, we propose a new relevance model to further improve retrieval performance and robustness. Three significant changes have been made to the original relevance model in order to estimate a more accurate relevance model for a query: *treating the original query as a special document, introducing rank-related prior,*and *discounting common words*. We will first give a theoretical justification of the three changes made in the new model in Section 3.1 and then detail each of the three improvements in Section 3.2.

3.1. Theoretical justifications of the new model

In the pure language model approach, there is no motivation for relevance feedback with a single generating document, as pointed out by Sparck-Jones, Robertson, Hiemstra, & Zaragoza (2003). They brought up two concerns: (1) how to deal with multiple relevant documents; and (2) how to handle relevance feedback in such cases. They have suggested that there should be an explicit model which generates a set of relevance documents. This treatment can be found in some more recent models (Lavrenko & Croft, 2001; Lafferty & Zhai, 2001) in handling feedback in the language model framework. However, in the original relevance models, queries and relevant documents are treated random samples from an underlying relevance model R as shown in Fig. 1. In the new relevance model, Queries are still random samples from the underlying relevance model R but relevant documents should be sampled from both the underlying relevance model R and a background language model B as shown in Fig. 2.

The original relevance model assumes that the sampling process could be different for queries and documents, even though they are sampled from the same relevance model R. Therefore, only relevant documents (top-ranked documents) are used for approximating the relevance model R. In the new relevance model, by adding the background language model B, we assume that the way that query words are sampled from the relevance model R is the same as the way that *topic words* in relevant documents are sampled from R, whereas *non-topic* words in relevant documents are sampled from the background language model B. Therefore, a query could be treated as a short document and be considered in the new relevance model together with the top-ranked, relevant documents.

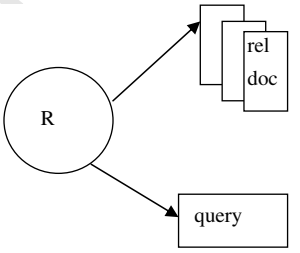

Fig. 1. Sampling process in the original relevance model.

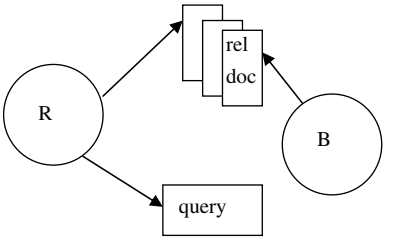

Fig. 2. Sampling process in the new relevance model.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

How could we utilize the background language model B in order to approximate a more accurate relevance model? In our approach, a *common-word-discounting* component is incorporated into the new relevance model, which will reduce the influence of non-topic words in the process of constructing the relevance model R.

Furthermore, even though top-ranked documents are not necessarily true relevant documents, documents with higher ranks are more likely to be relevant to a query thus can play a more important role than documents with lower ranks in approximating the relevance model. Therefore a *rank-relate-priors*component is designed in the new relevance model.

3.2. Three components in the new relevance model

3.2.1. Query as special document

First, the proposed model brings back the original query into the relevance model by treating it as a short, special document, instead of using a simple linear combination as in Abdul-Jaleel et al. (2004) (Section 2.2). The total probability of observing the word w together with query words $q_1 \dots q_k$ becomes:

$$P_{\text{new}}(w, q_1 \dots q_k) = \sum_{D \in S} P(D) P(w|D) \prod_{i=1}^{k} P(q_i|D) \tag{4}$$

Note that, unlike the set M including only top N documents' models in Eq. (2) for the original relevance model, the set S in Eq. (4) includes both the query model and the document models for the top N documents. The new model attempts to include the query model for a relevance model approximation so that it may lead to higher performance, especially for the queries whose performance decreased with the original relevance model.

A title query usually consists of a couple of key words and they are supposed to be the most relevant. However, the length of a query is much smaller than the average length of its relevant documents. Therefore, it is reasonable to assign a relatively small prior to the query and larger priors to relevant documents or top-ranked documents for estimating the relevance model for the query. This was implemented by using document length related priors for $P(D)$ into Eq. (4). This is further explained in the following sub-section: rank-related priors.

3.2.2. Rank-related priors

The second component is to assign different priors to the top N documents and the query (which is as a special document) according to the ranks of document, using Eq. (5):

$$P(D) = \frac{1}{Z_1} * \frac{\alpha + |D|}{\beta + \text{Rank}(D)} \tag{5}$$

$$Z_1 = \sum_{D \in S} \frac{\alpha + |D|}{\beta + \text{Rank}(D)} \tag{6}$$

In the above equations, $|D|$ denotes the length of document D or the length of the query – the special document. Rank(D) denotes the rank of document D in the ranked list of documents returned by using the basic query likelihood language model. The rank of the query is set to 0 so that it has the highest rank among all the documents used for relevance model approximation. Z_1 is the normalization factor that makes the sum of the priors to 1 (in Eq. (6)). In using the document length $|D|$ in Eq. (5), the assumption is that the estimated language models for longer documents are likely to be smoother and more accurate than those for shorter documents. Therefore, they are more reliable to be used for the estimation of relevance models. However, longer documents could contain more noise, therefore, parameters α and β are added to the length and the rank, respectively, to control how much a document's length and its rank affect the prior of the document, respectively. If both α and β are assigned very large values, then the priors will obey a uniform distribution, which is the same as that in the original relevance model approach. Considering that the length and the rank of a document have different quantities, we have in fact tried the use a multiplier in Eq. (5) instead of the additive parameters α and β. However, experimental results show that Eq. (5) gives better performance. Therefore, in using Eq. (5), we use a normalization term Z_1 defined in Eq. (6) to partially deal with this problem. In

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

our experiments, the parameters α and β were tuned on the query set used as training data. It turned out that the best performance was achieved on the training queries when α took the value around 140 and β took the value around 50.

The change of the priors was inspired by two pieces of work. One is the time-based language model by Li & Croft (2003), in which the uniform priors were replaced by an exponential distribution to favor recent documents. The other one is the work by Wessel, Thijs, & Djoerd (2002) for entry page finding. In their work, a fixed prior probability was learned for each category of pages. We note that weighted pseudo-relevance feedback was used in Zhai, Tao, Fang, & Shang (2003). In their paper, they assumed that the probability of the relevance of a document ranked at rank r to be $1/r$. As reported in their paper, however, the performance of retrieval was not improved. We also tried to use this strategy in our model, but experiments also showed no improvements in robustness.

3.2.3. Common word discounting

The last change to the original relevance models is to discount the probabilities of words that are common in the whole collection. In the framework of the original relevance models, relevant documents are samples of the underlying relevance model. In the new relevance models, words in relevant documents can be grouped into two classes: *topical words* and *non-topical words*. Here we introduce a *background language model* in our approach for this purpose. We assume that topical words are sampled from the underlying relevance model and non-topical words are sampled from the background language model B. Therefore, discounting the probabilities of words that are common in the whole collection will help to estimate a more accurate relevance model. In Zhai & Lafferty (2001), a sophisticated approach using the EM strategy was applied in a language model that explicitly penalized common words. In this paper, we use a much simpler yet very effective approach to incorporate the common word discounting in our new relevance model. The new relevance model is described by the following equations:

$$P_{\text{new}}(w|R) = \frac{1}{Z_2}\frac{P_{\text{new}}(w, q_1 \dots q_k)}{\gamma + P(w|B)} \tag{7}$$

$$Z_2 = \sum_{w \in V} \frac{P_{\text{new}}(w, q_1 \dots q_k)}{\gamma + P(w|B)} \tag{8}$$

$P_{\text{new}}(w|R)$ denotes the probability of word w in the new relevance model. $P(w|B)$ denotes the probability of word w in the background language model B. γ is the parameter for discounting the probability of a word in the new relevance model by its probability in the background language model. Z_2 is the normalization factor that makes the sum of the probabilities of words in the new relevance model to 1 (Eq. (8)). The best value of γ learned with the training queries is 0.02.

In deriving equation (7), we have also tried to use $\gamma P(w|B)$ instead of $\gamma + P(w|B)$. The former seemed probabilistically more meaningful, but experiments showed that the performance of relevant retrieval was better using the latter. The reason could be that the changes of $\gamma + P(w|B)$ are much smoother and slower than $\gamma P(w|B)$ from word to word.

Common words discounting can also be related to the 2-Poisson model (Harter, 1975). In the 2-Poisson model, occurrences of a term in a document have a random or stochastic element, which nevertheless reflects a real but hidden distinction between those documents that are "about" the topic represented by the term, and those that are not. Those documents that are "about" this topic are described as "elite" for the term. Whereas in our common word discounting component, terms in relevant, "about" topic documents are further grouped into *topical words* and *non-topical words* (i.e., *common words*). The 2-Poisson model assumes that the distribution of the with-document frequencies of a term is Poisson for the elite documents, and also Poisson (but with a smaller mean) for the non-elite documents. However, common words do not have their elite documents and non-elite documents. Therefore, applying common word discounting helps the estimation of a more accurate relevance model on topical words.

Note that the first change (*query as a special document* – Eq. (4)) has been incorporated in Eq. (7), and the second change (*rank-related priors* – Eq. (6)) has been incorporated in Eq. (4) when the new total probability is calculated. Therefore Eq. (7) integrates all the three new components (i.e., changes).

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

In the following two sections, we will first present our experimental results of the overall performance of the new relevance model versus the original relevance model and the linear combination model. Then we will perform a component analysis in order to obtain a better understanding of the roles of each of the three changes (components).

4. Experiments and results

We have carried out two sets of experiments (with pseudo feedback and true feedback) with four TREC query sets on three data collections. We applied the new relevance model to document retrieval with both true relevance feedback and pseudo feedback. In the case of true relevance feedback, all relevant documents were assigned a same value for Rank(D) in Eq. (5), since all the documents are supposed to be equally relevant. Therefore, ranking does not affect the priors when true relevant documents are used for relevance model approximation in Eq. (5). We will discuss in Section 4.3 what could we do if this assumption is relaxed. However, weighting over the lengths of the documents is considered. In the case of pseudo feedback, both the ranks and the lengths of the relevant documents are used in the prior calculation.

We compared the new robust relevance model with two baselines. One is the original relevance model (Section 2.1) and the other is the linear combination model (Section 2.2), which linearly combines the query model with the original relevance model. In all experiments, we used the query likelihood language model (Ponte & Croft, 1998) to retrieve top-ranked documents for feedback. All experiments were performed with the Lemur toolkit (Lemur, 2006). The Krovetz stemmer (Krovetz, 1993) was used for stemming and the standard stopword list of LEMUR was used to remove *about* 420 common terms.

4.1. Data

We used four query sets from three document collections in our experiments. One query set as the training data and the other three query sets as the testing data to evaluate the proposed model and two baseline models:

(1) Queries 151–200 on AP88 and AP89 collection. This was also used in (Lavrenko & Croft, 2001). This data set was used as training data. The parameters in both the two baseline relevance models and the new relevance models were tuned on this query set.
(2) Queries 101–150 on the Associated Press data set (AP88 and AP89). This was also used in (Lavrenko & Croft, 2001; Tao & Zhai, 2004; Zhai & Lafferty, 2001). Therefore, we can consider the two-stage mixture model for pseudo feedback (Tao & Zhai, 2004) as another baseline and compare our new relevance model with it on this query set.
(3) Queries 301–400 on a heterogeneous collection TREC45 that includes all data from TREC disk 4 and disk 5.
(4) Queries 701–750 on a sub-collection of the TREC Terabyte data set. To construct the subset, the top-ranked 10,000 documents for each of the 50 queries that were retrieved using the basic query likelihood language model were selected. The subset has 466,724 unique web documents and is about 2% of the entire terabyte collection (Lavrenko & Croft, 2001).

The statistics of the AP88&89 collection, the TREC45 collection, and the subset of terabyte collection are shown in Table 1. Here is the summary of the statistics that might help us better understand the performance of baseline and proposed methods.

(1) The average length of the documents in the TREC45 collection is 318, which is about 25% longer than the average length (254) of the news articles in the AP collection.
(2) The average frequency of terms in the TREC45 collection is about 18% more than that in the AP collection.
(3) The average length of the web documents in the terabyte collection is 2054, which is about 10 times longer than the average length (254) of the news articles in the AP collection.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

Table 1
Statistical comparison of the ~~two~~ document collections

Collection statistics	AP88&AP89	TREC45	Terabyte (GOV2)
# of documents	164,597	561,445	466,724
# of terms	41,827,813	178,893,105	958,740,730
# of unique terms	204,469	741,630	3,637,433
Length of documents	254	318	2,054
Frequency of terms	205	241	264

(4) The average frequency of terms in the subset of the terabyte collection is 30% more than that in the AP collection.

It is obvious that the three collections are very different, though the explicit impact of these factors to query expansion needs further study. In Sections 4.2 and 4.3, similar performance improvements were obtained with the testing query sets, even though the experiments were carried out on the three very different collections.

4.2. Pseudo feedback

In the case of pseudo feedback for retrieval, a few of the top-ranked documents were assumed relevant thus were used to estimate the relevance model for a query. Fig. 3 gives the performance of the proposed model, compared against that of the original relevance model and linear combination model with pseudo feedback on the training set (queries 151–200 on the AP collection). Figs. 4–6 compare the performance of the three models on three testing sets: TREC queries 101–150 on AP collection, 301–400 on the TREC45 collection, and queries 701–750 on a sub-collection of the TREC Terabyte data set.

4.2.1. Experimental results

There are two main conclusions that can be drawn based on the experimental results on the four query sets, given in Figs. 3–6, respectively.

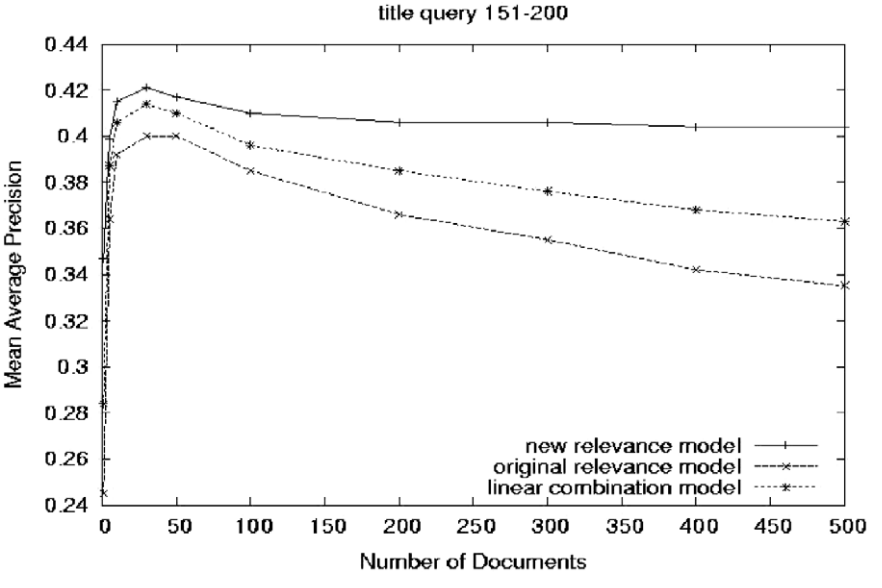

Fig. 3. Comparison between the new relevance model, the original relevance model and the linear combination model with pseudo feedback on the training set (query set 151–200).

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

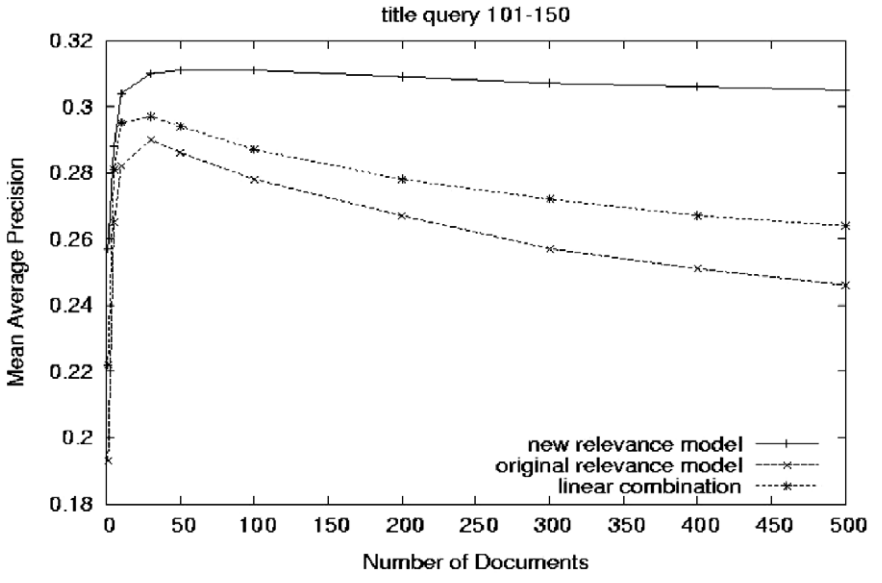

Fig. 4. Comparison between the new relevance model, the original relevance model and the linear combination model with pseudo feedback on testing query set 101–150.

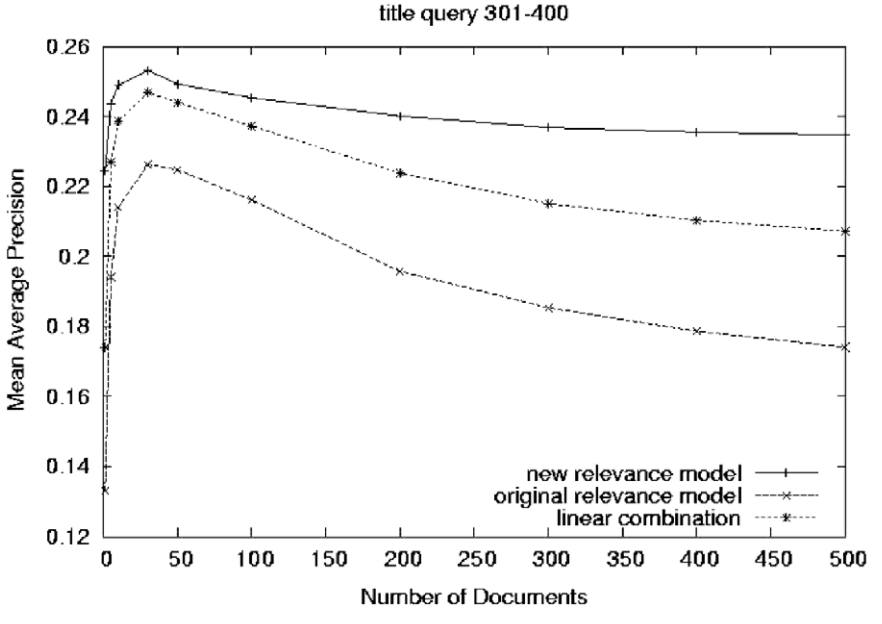

Fig. 5. Comparison between the new relevance model, the original relevance model and the linear combination model with pseudo feedback with testing query set 301–400.

(1) The new relevance model consistently outperformed the original relevance model and the linear combination model no matter how many documents were used for feedback. This can be clearly seen from the four graphs.
(2) The new relevance model is less sensitive to the number of feedback documents than the two baselines.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

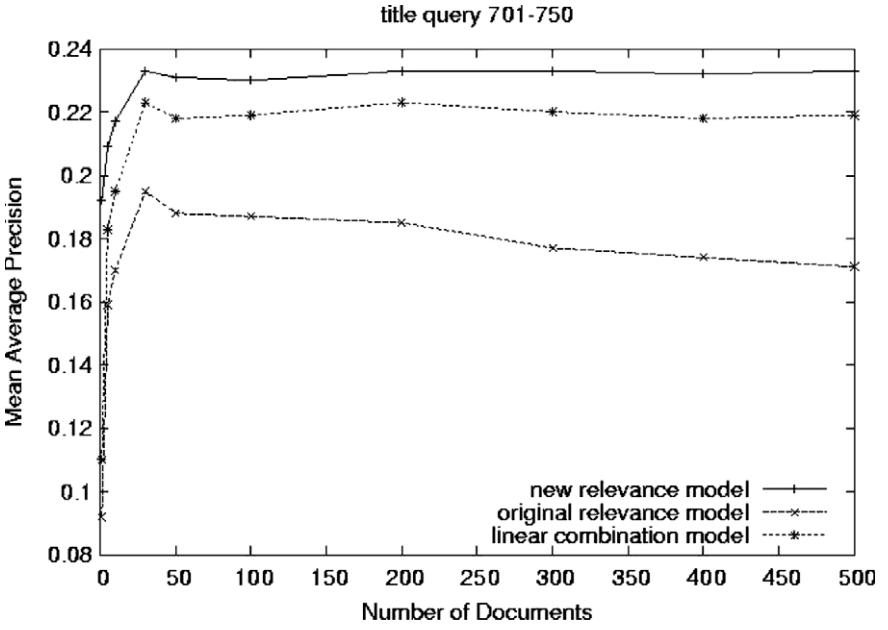

Fig. 6. Comparison between the new relevance model, the original relevance model and the linear combination model with pseudo feedback with query set 701–750 on a subset of the TREC Terabyte Track collection.

In the Figs. 3–6, both the new approach and the baselines achieved the best performance around the area where about 30 or 50 documents were used for feedback. However, for the AP collection and the TREC45 collection (Figs. 3–5), as the number of feedback documents increases, the performance of the original relevance model and the linear combination model dropped more quickly than the performance of the proposed new relevance model. On the subset of the TREC Terabyte collection, the performance of our new relevant model keep high when above 30 documents were used for feedback. As the number of feedback documents increases, the performance of the original relevance model dropped obviously. The drop of the performance of the linear combination model is not as obviously, but its performance is always lower than that of the new relevance model.

Our model is also more robust than the results reported in (Tao & Zhai, 2004) with the TREC 101–150 queries on AP88–89 collection. The sensitivity to the number of feedback documents of a two-stage mixture model was studied in (Tao & Zhai, 2004). Based on the results reported, the two-stage mixture model achieved the best performance around 30 feedback documents with the queries from 101 to 150 on AP88–89 collection. As the number of feedback documents increased to 500, the average precision dropped about 12%. Our model achieved the best performance around 50 feedback documents but only dropped less than 5% when top 500 documents were considered for relevance model approximation with the same query set on the same collection.

We also compared the robustness of our new relevance model with a more recent work by Tao & Zhai (2006), which also used the three ideas. In their method, regularized mixture models were used to estimate a query model with pseudo feedback documents for a query. Fig. 7 and Table 2 show the comparison, in which our new relevance model (NRM) approach was tested on four sets of data, and their regularized mixture model (RMM) approach was test on two sets of data. In Fig. 7, the mean average precision is shown for each case. While the absolute values of precision is not comparable with different sets of data, we compare the robustness of the two methods in terms of the difference of precision (DP), which is measured as the relative change of precision (in percentile) from the highest precision value to the lowest when the number of relevance feedback documents increase. While the mixture model approach use a more well-recognized method (the EM algorithm) to train the parameters for relevance models, Table 2 shows that the robustness of our approach is comparable to that of their approach, and is in fact slightly better.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

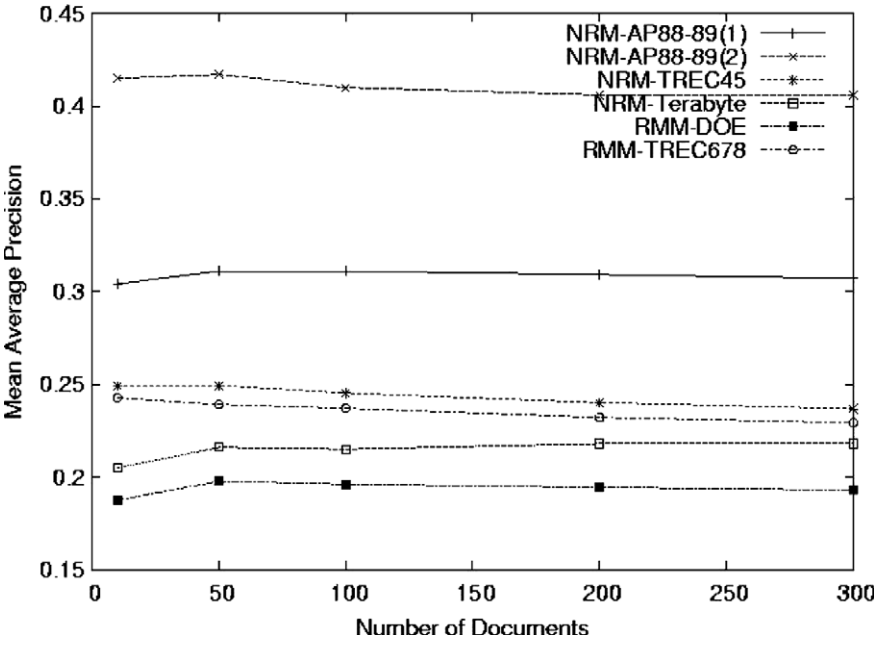

Fig. 7. Comparison of the robustness of our new relevance model (NRM) approach with the regularized mixture model (RMM) approach.

Table 2
Comparison of the robustness of our approach with the regularized mixture model approach

	NRM – AP88–89(1)	NRM – APP88–89(2)	NRM – TREC45	NRM – Terabyte	RMM – DOE	RMM – TREC678
DP	−1.29%	−2.2%	−5.5%	+0.5%	−2.53%	−5.67%
Average DP	−2.46%				−4.15%	

4.2.2. Discussion

Relevance models can be viewed as a way of query expansion in the sense that they introduce more words into the query representation. Query expansion techniques are not guaranteed to work on every query though they usually can achieve better performance than using the query when measuring the mean average precision on a set of queries. The performance of some queries may be hurt using query expansion techniques while some queries can get significant improvements. Table 3 showed how many queries were affected significantly by using the new relevance model and two baseline models. In the table, Ni denotes the number of queries whose performance increased by 40% in terms of average precision compared to the performance of the query likelihood language model. Nd denotes the number of queries whose performance decreased by 40%. We have the following observations based on our experiments.

First, there were more queries whose performance increased significantly but fewer queries whose performance was hurt badly using the new relevance model than using the original relevance model and the linear combination model. This is obvious in Table 3 in that Ni of the new relevance model is almost always the highest among the three models, whereas Nd of the new model is always the lowest among the three.

Second, for queries 101–150 and 151–200 on the AP collection, there are more queries whose performance was improved significantly than the queries whose performance was hurt badly, with all the three models. This is also true for queries 301–400 on the TREC45 collection, with an exception for the original relevance model. There are 28 queries whose performance was significantly increased but with the performance of 30 queries decreased.

However, this is not true on the subset of the Terabyte collection. For queries 701–750 in Table 3, the performance of a large number of queries decreased significantly. Based on our experimental results, all three rel-

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

Table 3
Query-based comparisons of relevance models to query likelihood language models (Orig.: the original relevance model, LC: linear combination model, New: the new relevance model)

Query\method	AP 101–150		AP 151–200		TREC45 301–400		Terabyte 701–750	
	Ni	Nd	Ni	Nd	Ni	Nd	Ni	Nd
Orig.	20	12	20	9	28	30	3	23
LC	17	7	24	3	25	16	7	21
New	21	7	25	4	33	17	7	19

evance models implemented in this paper did not improve retrieval performance with the queries 701–750 on the TREC terabyte collection. This observation is similar to the findings by the groups who applied relevance models or query expansion techniques in the TREC Terabyte Track. Nevertheless, even with this query set, the new robust relevance model performed the best among the three.

Third, compared to queries 101–200 on the AP collection, the performance improvement for queries 301–400 on the TREC45 collection using the new relevance model is not as significant. We notice that the TREC45 collection is composed of news articles from many different resources. Some of the documents are very long and may span multiple topics. When long, cross-topic documents are used for feedback, words related to other topics in the documents will play a negative role in constructing the relevance models for a query, therefore drive the estimated relevance models to drift away from the true relevance model of the query.

This explanation is further verified with our experiments with data from the TREC Terabyte track. The Terabyte collection is more diverse than the TREC45 collection. It has many noisy web pages as well as long documentations spanning multiple topics. Similar observations were also made in Melzler, Srohman, Turtle, & Croft (2004) with the Terabyte track.

Passage retrieval was reported effective on collections that have long cross-topic documents (Liu & Croft, 2002). Therefore, a future extension of the new relevance model is to incorporate passage retrieval for a consistent retrieval performance for queries over heterogeneous collections. Our ongoing experiments have shown the promise in this direction.

4.3. True relevance feedback

In the case of true relevance feedback, a number of known relevant documents were used to estimate the relevance model for a query. Fig. 8 shows the average performance of the new relevance model, the original relevance model and the linear combination model with true relevance feedback on 18 queries selected from the TREC queries 101–150. The purpose of the experiments for true relevance feedback is to study how the three models behave when more truly relevant documents are given for relevance model approximation. The criterion in selecting the queries was: each of the 18 queries used in this experiment have at least 30 relevant documents within the 200 top-ranked documents from the first round retrieval. Note that this test does not include queries from the training set (queries 151–200).

The equal relevance assumption could be relaxed if truly relevant documents have different degrees of relevance. In such a case, a similar approach as in the pseudo-relevance feedback case may be applied.

Three main conclusions can be drawn based on the experimental results given in Fig. 8:

(1) The performance of both the baselines and the new relevance models increases as the number of feedback relevance documents increases.
(2) As the number of feedback relevant documents increases, the new relevance model consistently outperforms the two baselines in terms of mean average precision.
(3) The new relevance model can achieve even better performance than the two baselines when using fewer relevant documents. The new relevant model achieves about 0.57 of mean average precision when 15 relevant documents are used for feedback. But the original relevance model can only achieved the same performance (0.57) and the linear combination model achieves 0.56, respectively, with as many as 30 relevant documents used each for feedback.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

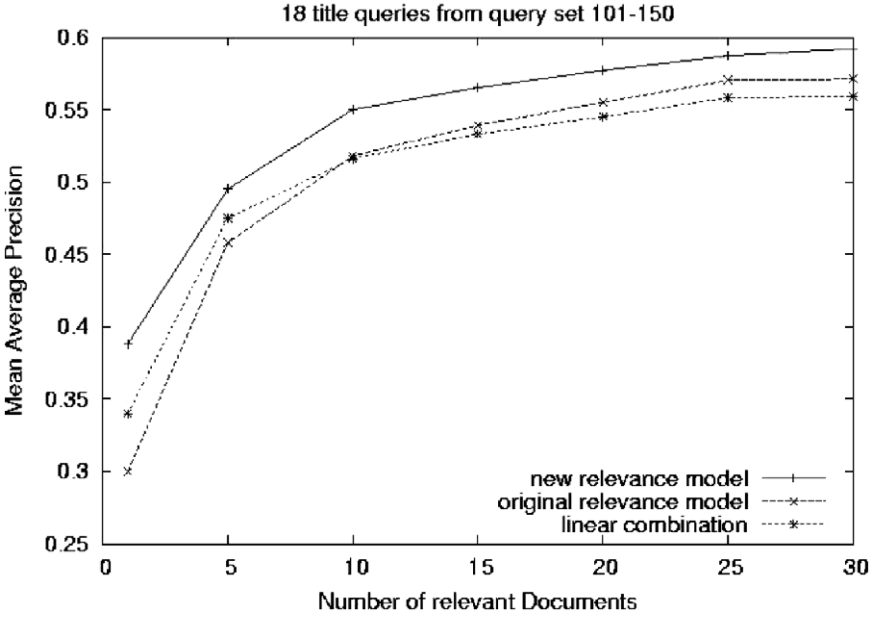

Fig. 8. Comparison between the new relevance model, the original relevance model and the linear combination model with true relevance feedback on 18 queries chosen from testing queries 101–150.

(4) The linear combination model outperforms the original model only when a small number of relevant documents are used for estimating relevance models. However, as more feedback relevant documents are used, the performance of the original relevance model is closer to and even better than the performance of linear combination model.

5. Component analysis

In the new relevance model, there are three new components added to the original relevance model: *treating the query as a special document, introducing document-ranking-related priors*, and *discounting common words*. To separate the contribution of the three components, we have carried out a set of experiments to breakdown the performance of the new relevance model on both the training data set (queries 151–200) and a testing data set (queries 101–150).

Our first step was to study the contribution of treating queries as special documents by removing query from the set S in Eq. (4). Therefore, in this case, only top-ranked documents were used for relevance model approximation. The curves labeled by "no query" in Figs. 9 and 10 stand for the performance of experiments without the *query as special document* component. Compared to the performance of the new relevance model with all three components, the performance on average dropped about 2.5% for the training queries 151–200 and 1.8% for the testing queries 101–150, respectively.

Our second step was to explore the role of the *rank-related priors* component in the new relevance model. In Figs. 9 and 10, the curves labeled by "no ranking" mean that document ranking will not be used in adapting document priors, which was implemented by assigning a very large value to β in Eq. (5). Compared to the performance of the new relevance model with all the three components, the performance without *rank-related priors* component got about the same performance with 50 feedback documents used for the training query set 101–150, and with 30 feedback documents used for the test query set 151–200. But the performance without rank-related priors dropped more with more documents for feedback on both query sets, up to as large as about 12% when the number of documents is 500.

Our last step of component analysis was to study the role of the *common word discounting* component. We removed the *common word discounting* component from the framework of the new relevance model but kept

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

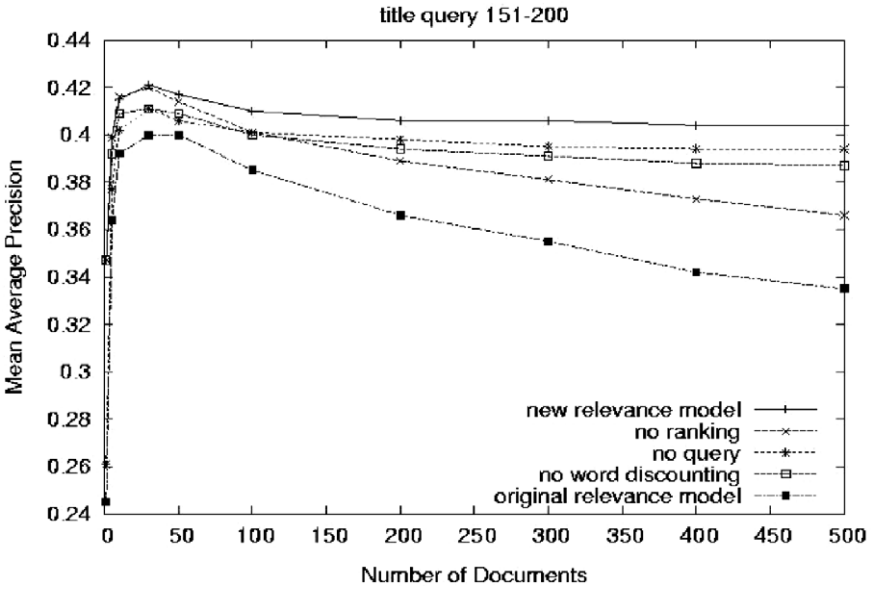

Fig. 9. New relevance model component analysis on training query set 151–200 (No query: query is not considered for relevance model approximation; no ranking: document rank is not considered; no word discounting: doesn't discount probabilities of the words that are common in the collection).

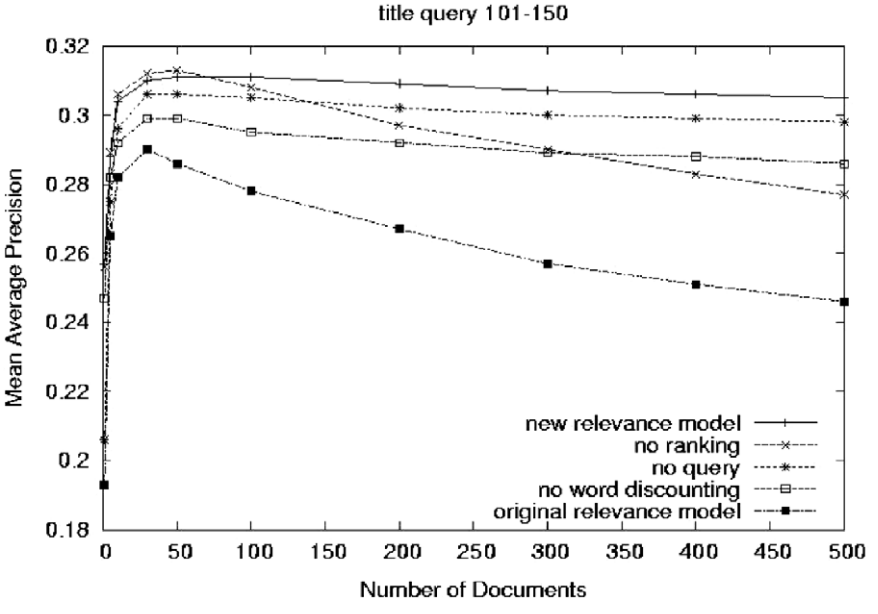

Fig. 10. New relevance model component analysis on testing query set 101–150 (No query: query is not considered for relevance model approximation; no ranking: document rank is not considered; no word discounting: doesn't discount probabilities of the words that are common in the collection).

the other two components. The performance is represented by the curves by "no word discounting" in Figs. 9 and 10. Compared to the performance of the new relevance model with all three components, the performance on average dropped about 5% on for queries 101–150 and about 3% for queries 151–200.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

Figs. 9 and 10 compared the contribution of each component on both the training query set 151–200 and the test query set 101–150. From the above discussions, we can draw several conclusions:

(1) Replacing the uniform priors in the original relevance model with document-ranking-related priors given in Eq. (5) makes the model less sensitive to the number of pseudo feedback documents.
(2) Both considering query as a special document and discounting common word probabilities can improve the performance in terms of mean average precision.
(3) Most of the performance gain of the new relevance model on average seem to be obtained by the word discounting component, but more performance gain is caused by rank-related priors when more documents are used for feedback.

6. Conclusions and future work

In this paper, a new robust relevance model has been proposed. It was applied to both pseudo feedback and true relevance feedback in the language-modeling framework. The main contributions of this work are the follows.

(1) Three features are studied that have impact on the performance of document retrieval, based on well-designed experiments. The features include key words from original queries, relevance ranks of documents from the first round retrieval, and common words in the background data collection.
(2) The features are seamlessly incorporated into the original relevance-based language model to improve its performance and robustness. The three corresponding incorporations are: bringing back the original title query as a special document, introducing document-ranking-related priors, and discounting common words.

Three main conclusions have been drawn from the experimental results queries on three data collections: queries 101–150 and 151–200 on the AP88&89 collection, queries 301–400 on the TREC45 collection for the TREC ad-hoc retrieval task, and queries 701–750 on a sub-collection of the TREC Terabyte data set.

First, the new model outperforms both the original relevance model and the linear combination model in terms of mean average precision on document retrieval with both pseudo-relevance feedback and true relevance feedback.

Second, all three models achieved their best performance when about 30–50 top-ranked documents were used for relevance model approximation, but our new model is more robust in the sense that it is less sensitive to the number of documents considered for pseudo feedback than the two baseline models compared. Therefore, the new relevance model can benefit from a large number of feedback documents while the performance drops quickly with the original relevance model and the linear combination model as the number of feedback documents increases.

Third, in case of true relevance feedback, the new relevance model achieves a better performance with less relevant documents. This property is very important and desirable because relevance judgments are expensive and usually very hard to obtain.

We note here that although the new relevance model outperforms the original relevance model, there are still some queries, whose retrieval performance in fact is decreased when using pseudo feedback. Future work will focus on query-based relevance models that allow parameters in the new relevance models to have different values for different queries. A possible way is to incorporate selective query extension techniques, such as the work by Cronen-Townsend, Zhou, & Croft (2004), into the new relevance model. Queries may be first grouped into two classes. Queries belonging to the first class are likely to have better performance with query expansion techniques and queries belonging to the second class are likely to decrease performance with query expansion techniques. Therefore, the new relevance model may learn different parameter values for the two different classes of queries.

As another future work, new approaches to query expansion techniques need to be developed for retrieval on heterogeneous collections (e.g., the Terabyte collection), which may include web documents, blogs, emails

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

as well as news articles. In this case, incorporating passage retrieval and features like metadata into relevance models may be helpful.

7. Uncited reference

Q1 Clarke, Craswell, & Soboroff (2004).

Acknowledgements

This work was supported in part by Center for Intelligent Information Retrieval at the University of Massachusetts at Amherst, and by DARPA under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsors. Earlier versions of this work was first appeared in a technical report (Li, 2005), and then was presented at the *Fourth IASTED International Conference on Communications, Internet and Information Technology* (Li, 2006).

References

Abdul-Jaleel, N. et al. (2004), UMASS at TREC2004. In *The thirteen text retrieval conference (TREC 2004) notebook*.
Clarke, C., Craswell, N., & Soboroff, I. (2004). Overview of the TREC 2004 terabyte track. In *The thirteen text retrieval conference (TREC 2004) notebook*.
Cronen-Townsend, S., Zhou, Y., Croft, W. B. (2004). A framework for selective query expansion. In *Proceedings of thirteenth international conference on information and knowledge management (CIKM'04)* (pp. 236–237).
Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing, Part 1: On the distribution of speciality words in a technical literature, Part 2: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science, 26*, 197–206, and 280–289.
Hiemstra, D. (2001). Using Language Models for Information Retrieval. PhD thesis, University of Twente.
Krovetz, R. (1993). Viewing morphology as an inference process. In *Proc. 16th ACM SIGIR conference on research and development in information retrieval (SIGIR'93)* (pp. 191–202). Pittsburgh, June 27–July 1.
Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *24th ACM SIGIR conference on research and development in information retrieval (SIGIR'01)* (pp. 111–119).
Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In *24th ACM SIGIR conference on research and development in information retrieval (SIGIR'01)* (pp. 120–127).
Lavrenko, V., & Croft, W. B. (2003). Relevance models in information retrieval. In W. Bruce Croft & John Lafferty (Eds.), *Language modeling for information retrieval* (pp. 11–56). Kluwer Academic Publishers.
Lemur, (2006). Lemur toolkit for language modeling and information retrieval. The LEMUR PROJECT by CMU and UMASS (http://www.lemurproject.org).
Li, X. (2005). Improving the robustness of relevance-based language models, CIIR Technical Report, IR-401, Department of Computer Science, University of Massachusetts Amherst, 2005. http://ciir.cs.umass.edu/pubfiles/ir-401.pdf.
Li, X. (2006). Robust relevance-based language models, In *Proceedings of the fourth IASTED international conference on communications, internet and information technology (CIIT 2006)*, November 29–December 1, St. Thomas, US Virgin Islands.
Li, X., & Croft, W. B. (2003). Time-based language models. In *Proceedings 12th international conference on information and knowledge management (CIKM'03)* (pp. 469–475).
Liu, X. & Croft, W. B. (2002). Passage retrieval based on language models. In *Proc. 11th international conference on information and knowledge management (CIKM'02)* (pp. 375–382).
Melzler, D., Srohman, T., Turtle, H., & Croft, W. B. Indri at TREC 2004: terabyte track. In *The thirteen text retrieval conference (TREC 2004) notebook*.
Miller, D. H., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. In *22nd ACM SIGIR conference on research and development in information retrieval (SIGIR'99)* (pp. 214–221).
Ponte, J. (1998). A language modeling approach to information retrieval. PhD thesis, UMass-Amherst.
Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *21st ACM SIGIR conference on research and development in information retrieval (SIGIR'98)* (pp. 275–281).
Song, F. & Croft, W. B. (1999). A general language model for information retrieval. In *22nd ACM SIGIR conference on research and development in information retrieval (SIGIR'99)* (pp. 279–280).
Sparck-Jones, K., Robertson, S. E., Hiemstra, D., & Zaragoza, H. (2003). Language modelling and relevance. In W. B. Croft & J. Lafferty (Eds.), *Language modeling for information retrieval* (pp. 57–71). 2003: Kluwer Academic Publishers.
Tao, T., & Zhai, C. (2004). A two-stage mixture model for pseudo feedback. In *Proc. 27th ACM SIGIR conference on research and development in information retrieval (SIGIR'04)* (pp. 486–487).

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005

Tao, T., & Zhai, C. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proc. 29thACM SIGIR conference on research and development in information retrieval (SIGIR'06)* (pp. 162–169).

Wessel, K., Thijs, W., & Djoerd, H. (2002). The importance of prior probabilities for entry page search. In *25th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)* (pp. 27–34).

Zhai, C., & Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proc. tenth international conference on information and knowledge management (CIKM'01)* (pp. 403–441).

Zhai, C., Tao, T., Fang, H., & Shang, Z. (2003). Improving the robustness of language models – UIUC TREC-2003 robust and genomics experiments. In *The 12th text retrieval conference (TREC 2003) notebook*.

Please cite this article in press as: Li, X. , A new robust relevance model in the language model framework, *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.07.005