TopPRF: A Probabilistic Framework for Integrating Topic Space into Pseudo Relevance Feedback

JUN MIAO, JIMMY XIANGJI HUANG, and JIASHU ZHAO,
Information Retrieval & Knowledge Management Research Lab, York University

Traditional pseudo relevance feedback (PRF) models choose top k feedback documents for query expansion and treat those documents equally. When k is determined, feedback terms are selected without considering the reliability of these documents for relevance. Because the performance of PRF is sensitive to the selection of feedback terms, noisy terms imported from these irrelevant documents or partially relevant documents will harm the final results extensively. Intuitively, terms in these documents should be considered less important for feedback term selection. Nonetheless, how to measure the reliability of feedback documents is a difficult problem.

Recently, topic modeling has become more and more popular in the information retrieval (IR) area. In order to identify how reliable a feedback document is to be relevant, we attempt to adapt the topical information into PRF. However, topics are hard to be quantified and therefore the identification of topic is usually fuzzy. It is very challenging for integrating the obtained topical information effectively into IR and other text-processing-related areas. Current research work mainly focuses on mining relevant information from particular topics. This is extremely difficult when the boundaries of different topics are hard to define. In this article, we investigate a key factor of this problem, the topic number for topic modeling and how it makes topics "fuzzy." To effectively and efficiently apply topical information, we propose a new probabilistic framework, "TopPRF," and three models, TS-COS, TS-EU, and TS-Entropy, via integrating "Topic Space" (TS) information into pseudo relevance feedback. These methods discover how reliable a document is to be relevant through both term and topical information. When selecting feedback terms, candidate terms in more reliable feedback documents should obtain extra weights. Experimental results on various public collections justify that our proposed methods can significantly reduce the influence of "fuzzy topics" and obtain stable, good results over the strong baseline models. Our proposed probabilistic framework, TopPRF, and three topic-space-based models are capable of searching documents beyond traditional term matching only and provide a promising avenue for constructing better topic-space-based IR systems. Moreover, in-depth discussions and conclusions are made to help other researchers apply topical information effectively.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Relevance feedback

General Terms: Performance, Experimentation, Modeling

Additional Key Words and Phrases: Pseudo relevance feedback, topic modeling, text mining

ACM Reference Format:
Jun Miao, Jimmy Xiangji Huang, and Jiashu Zhao. 2016. TopPRF: A probabilistic framework for integrating topic space into pseudo relevance feedback. ACM Trans. Inf. Syst. 34, 4, Article 22 (August 2016), 36 pages. DOI: http://dx.doi.org/10.1145/2956234

The work described in this article is partially supported by the Discovery grant and CREATE award from the Natural Sciences & Engineering Research Council (NSERC) of Canada, the Early Researcher Award/Premiers Research Excellence Award, and the IBM Shared University (SUR) Award. This work is also affiliated with the Information Retrieval and Knowledge Management Research Laboratory.
Authors' addresses: J. Miao, Information Retrieval and Knowledge Management Research Lab, Computer Science and Engineering Department, York University, Canada; email: jun@cse.yorku.ca; J. X. Huang (corresponding author) and J. Zhao, Information Retrieval and Knowledge Management Research Lab, School of Information Technology, York University, Canada; emails: jhuang@yorku.ca, jessie@cse.yorku.ca.

DOI: http://dx.doi.org/10.1145/2956234

1. INTRODUCTION

1.1. Background and Problem

Pseudo relevance feedback is an effective technique for improving performance in information retrieval. It assumes that top-ranked documents in the first-pass retrieval are relevant and then uses them as feedback documents to refine the representation of original queries by adding potentially related terms. These terms are called feedback terms. PRF has been shown to be effective in previous work [Andrzejewski and Buttler 2011; Lv et al. 2011; Ye et al. 2010; Ye and Huang 2014] for query expansion.

Traditionally in classic PRF models like Rocchio [1971] or RM3 [Lavrenko and Croft 2001], all the top k feedback documents are assumed to be equally relevant. The weights of candidate feedback terms in them are calculated based on their own features only. Once the documents are chosen, their reliability[1] is not considered anymore. Generally, terms in different feedback documents with the same weight (e.g., tf-idf score) are considered to be equally reliable for query expansion. According to our preliminary experiments, we use BM25 [Robertson et al. 1994] with optimal parameters to investigate how reliable the top k feedback documents are. As we know, BM25 is one of the most popular models and has been widely used as the basic model of probabilistic PRF [Huang et al. 2006; Clarke et al. 2008; Robertson and Zaragoza 2009; Miao et al. 2012]. Surprisingly, the ratio of really relevant documents in top k is not high. On the WT10G dataset with TREC2001 queries, approximately one-third in the top three, three-fifths in the top 10, two-thirds in the top 30, and four-fifths in the top 50 documents are irrelevant. Those irrelevant documents import noisy information, which can harm the overall performance of PRF significantly. Meanwhile, even if a document is relevant, it can also contain irrelevant contents. Terms in these irrelevant contents will influence query expansion negatively as well.

In a document, a relevant term is surrounded by other terms that can be either relevant or not. Without extra information, it is hard to identify the relevance of a term from terms around it, especially when the document itself can be irrelevant. Recently, researchers have begun to apply topic models [Serizawa and Kobayashi 2013; Wang et al. 2012; Yi and Allan 2009, 2008; Ye et al. 2011] for PRF to solve this problem. They attempted to find feedback terms in the most relevant topic(s) and expanded the original query with them. In other words, they used relevant topics to replace feedback documents for PRF. The advantage of this kind of method is breaking the constraint of document scope. Because topic modeling considers the co-occurrence of terms within the whole collection for training, term relations can be conducted. For example, when a term t_1 always appears with query terms, it is very likely to be relevant to the original query and have a high probability in the query-related topic K. If there is another term t_2 that co-occurs with t_1 in other documents, t_2 will have a high probability in K. It is possible that t_2 does not appear with query terms many times because query terms can have synonyms. In this case, we can find t_2 through topical information and expect that the top terms in the relevant topics are also relevant.

There is a big obstacle for this application, which we call a "fuzzy topic" problem. In this article, a topic is defined as the main theme or subject contained in a (set of) document(s), which can be represented by a list of terms with the corresponding probabilities of generating the terms from the topic. A topic can be considered as a particular distribution of terms in vocabularies. It is not a very clear concept even for human beings. In other words, topics are abstract. Hence, it is difficult to identify how many and what topics a document is really about. For example, one may consider that a document is about the "finance" topic. A different person may think the document

[1]In this article, "reliability" of a feedback document refers to how reliable it is to be relevant.

actually contains two topics, "stock" and "bond," and people who regard the document as "investment" related can also be correct.

We can think about this problem more generally. Suppose we have a corpus and define the information as global information. The global information is fixed if the corpus is not changed. When we decide to discover how many topics there are in this corpus, we actually attempt to split the global information into subinformation pieces and make topics through them. Topics are the aggregations or the segments of these subinformation pieces. Since information itself is hard to be quantified and people can interpret it differently, there are reasonably many ways to organize these subinformation pieces and generate different topics. So the "fuzzy topic" problem appears naturally and the best number of topics in a corpus cannot be determined.

Previous work tried to find relevant terms in particular topics. So the problem of selecting relevant terms changes to identifying relevant topics. However, as we can see from the previous example, the information in each topic can be divided or aggregated. This depends on how many topics we assume to be there. If we attempt to obtain particular topics through some kind of rule, the desired information in these topics can be quite different when the topic number is changed. Unfortunately, when we use popular topic models to discover topics from a corpus, there is not an appropriate way to determine the topic number. Previous researchers proposed some methods to find an optimal topic number [Blei et al. 2003a; Griffiths and Steyvers 2004; Blei and Tenenbaum 2004]. But none of them is IR or PRF oriented. As a result, the performance of the applications based on these topics will be very unstable, and the unstableness will propagate to these methods and affect the overall performance. This is why topics are coarse to some extent [Wei and Croft 2006] and hard to be applied in IR. More evidence will be shown in Section 3 to demonstrate how topics change significantly with different topic numbers.

Besides the challenge of identifying topics, we think another latent problem is the loss of topical information. Using a few topics can neglect useful information in other ones, even when they are relevant to the query topic. For example, terms with higher probabilities in the selected topic(s) are considered to be more important for PRF. However, they can also appear frequently in other topics and they are actually not so informative. In that case, terms that have even probabilities in many topics should be less important. Without the information of full topics, we will miss these kinds of features.

Generally, if we decide to utilize particular topics, we have to decide or seek an optimal topic number first. Then the effect of the "fuzzy topic" problem is inevitable due to a significant change of topics and the loss of topical information.

Instead of identifying relevant feedback terms directly, our idea is to select feedback terms based on the relevancy of feedback documents using topical information. For example, Huang et al. [2013] applied the machine-learning technique of cotraining to label feedback documents as relevant/irrelevant, and Ye et al. [2012] took the original score of feedback documents into account. Previous work started to consider whether a feedback document is really relevant or not but was not from the view of topics. In order to address the fuzzy topic issue when using topical information, we propose a new concept of "topic space" in the next section.

1.2. Topic Space

To identify whether a feedback document is relevant to the query, an intuitive way is to measure the similarity between feedback documents and really relevant documents. Since documents are represented by terms, traditional similarity measurements are also based on terms, for example, vector space model or cosine similarity. In this article, we use topics to represent a document. Topics contain more general information than

terms because the former are a distribution of all terms in the vocabulary. By using topics to represent a document, we can discover associations at a different level. To some extent, topics can reveal more semantic information than terms. So it is worthy to consider how to utilize this to evaluate feedback documents.

To effectively apply topical information for PRF and reduce the influence of fuzzy topics, we take a complete view on the documents and propose a new concept, "topic space." A document is considered as a mixture of different information. If we treat a document as a point in the space, we can use different coordinate systems to describe it and locate it. When we change the dimension and the meaning of each axis of the system, we will have a different view of the document. In the vector space model [Salton et al. 1975b], the dimension of the system is the size of the vocabulary, and each dimension is the weight of a particular term given a document. The projection of the document point on each dimension demonstrates how important a term is in this document. What will we obtain if we create a coordinate system based on topics to describe a document?

When integrating topic modeling on a corpus, suppose we set the number of topics to be M and then we will have a set of topics $z_1, z_2 \cdots z_M$. If a document d is about z_1 and z_2, the probabilities $p(z_1|d)$, $p(z_2|d)$ should be obviously higher than $p(z_i|d)$ while $i \neq 1$ *or* 2. We build a coordinate system [Stark et al. 1998] based on the topics we obtain. The coordinates $(p(z_1|d), p(z_2|d), \ldots p(z_M|d))$ are used to denote document d and the summation of $p(z_i|d)$ is 1 for $i \in \{1, 2 \ldots M\}$. We define the coordinate system as a *topic space* and documents are vectors in this space. So the system has M dimensions and each dimension denotes the conditional probability of a topic given a document. When we change M, we change the way to describe the document point, or the mixture of information in other words. No matter how we change the system, the document itself is unchanged in the topic space. In this case, we can always precisely describe the document with all the topic coordinate information. Unlike the term-based coordinate system, the matching of different documents/queries can be done beyond bag-of-words techniques in the topic space.

With this new concept, it is simple to map a document into the space and apply sophisticated space-related methods. For instance, we can define a *topic vector* as starting from the original point and ending at the document point, and then methods applied in the vector space model [Salton et al. 1975b] can be used as well. An example is shown in Figure 1. Three documents are represented as three vectors when there are three topics. If only some of the topics are used, we just investigate projections on some dimensions of the document, which is not complete and cannot describe the document accurately. When we attempt to discover useful information among documents, we always use the full topic coordinates to avoid biased topical information. We will show some experimental results in Sections 7.1 and 7.2 to justify this.

To measure the reliability of each document and choose terms to expand the original query on the evaluation, we do not focus on a particular topic. Instead, we use the coordinates of a document in the topic space to implement our ideas. To this end, we have the following two assumptions.

ASSUMPTION 1. *If two documents are similar on the topical level, their positions in the topic space will be close even when the topic number is changed.*

For instance, if two documents d_1 and d_2 are about "stock investment" with 10 topics $z_1, z_2 \ldots z_{10}$, if both z_3 and z_5 are related to "stock investment," $p(z_3|d_1)$, $p(z_5|d_1)$, $p(z_3|d_2)$, and $p(z_5|d_2)$ should be obviously higher than other conditional probabilities. If we change the topic number to five, there may be only one topic z_2 related to "stock investment." The two documents should still have similar topic coordinates while their contents are very similar. However, suppose d_1 is a query and we still set the topic

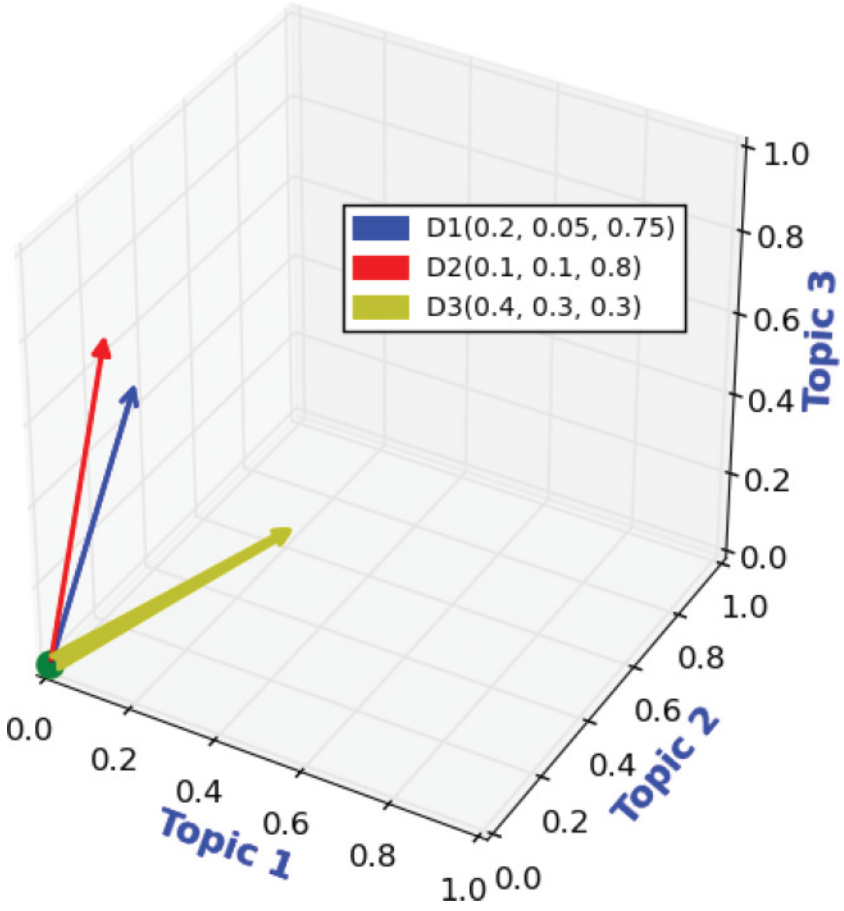


Fig. 1. Documents represented in a 3-topic space. Numbers in brackets are the coordinate values for each document.

number to 10. If we decide to choose one topic for selecting feedback terms, no matter how we choose z_3 or z_5, information in the neglected topic will be lost.

An issue we consider is how to measure the similarity/closeness of two documents. Here we propose two models. Our purpose is to research whether the topic similarity between documents can help improve PRF instead of which similarity method will be the best. Therefore, we intuitively choose two very popular similarity measurements. Other similarity methods will be studied in future work.

First, we can consider the cosine similarity between the topic vectors of two documents. The association between topic vectors should be more stable than particular topics while we view the documents in the complete scope of topics. Sometimes it is also a useful sign when two documents both have low probabilities in a particular topic. Second, if we use the distance between two document points to measure their similarity, there will be plenty of candidate formulas to investigate (e.g., Euclidean distance). In this article, we will propose two methods named TS-COS and TS-EC to apply topic similarity scores for estimating the reliability of a feedback document. Details of these two methods will be presented in Sections 4.3 and 4.4. The higher score it has, the more likely it is relevant. The scores of these documents will affect the weights of candidate feedback terms in them. Terms in documents of high weights

are considered to have more impact for query expansion. In Figure 1, the coordinates for D1 and D2 are (0.2, 0.05, 0.75) and (0.1, 0.1, 0.8), respectively. Because both of them have a large portion of topic 3, they are very close in the topic space as shown in Figure 1. That is the feature we decide to make use of. In this article, we apply the cosine similarity and Euclidean distance to measure the closeness of two documents in the topic space. In addition, in order to apply the two methods and obtain the weights of each document, we select a small group of feedback document as samples and measure the average similarity score of each feedback document to these samples as its weight. Details about how and why we choose the samples are introduced in Section 4.3.

ASSUMPTION 2. *In PRF, the feedback documents are considered to be relevant. The fewer topics a document contains, the more reliable the document is.*

When we assume the top k feedback documents in PRF are relevant, if a document is only about one topic (i.e., one topic has a much larger probability than others given the document), we can consider that all the contents of the document are relevant, or we denote it a "pure" document. Otherwise, if the topic distribution of a document is very even, it is reasonable to think some parts of the document are not relevant. Thus, it is risky to import terms from them. In that case, a less pure document is not so reliable when evaluating the weights of candidate feedback terms. Inspired by the traditional information theory, we measure the purity of topical information in a document through "entropy" by replacing the probabilities of terms with those of topics. We also propose a TS-Entropy method on this assumption to address the negative effect of partially relevant documents. Details of the TS-Entropy method will be presented in Section 4.5.

To the best of our knowledge, our proposed approaches are novel for integrating all topical information instead of selected topics in PRF under the probabilistic framework. A document is represented as a mixture of all topics, and the latent topical information is retained to represent the meaning beyond individual terms. According to our Assumption 1 and experimental results presented in Sections 6 and 7, our proposed methods are not sensitive in terms of retrieval performance to the settings of topic numbers.

The contributions of this work are as follows. First, as far as we know, this is the first study that researches how the fuzzy topic problem affects the application of topical information on PRF. Second, we introduce a new concept, topic space, to apply full document-topical information for PRF under a probabilistic framework. This is a novel way of integrating topical information for PRF. Instead of discovering the most relevant topic(s), we use topic vectors to represent documents and mine topic-level information to evaluate the relevance of feedback documents without human efforts. The concept can extend the current bag-of-words techniques to the topical level. Third, based on the idea and two assumptions, we propose a new probabilistic framework, TopPRF, and three novel models. Extensive experiments on public datasets indicate the effectiveness of them. Finally, in-depth useful discussions and conclusions are made for further research and extension of our work.

In the rest of this article, Section 2 introduces the research work related to this study. Section 3 demonstrates how topics change with different topic numbers and the challenges of integrating topic space in PRF. In Section 4, a new framework, TopPRF, and all three methods are presented in detail. Experimental settings and baselines are introduced in Section 5. Next, extensive experimental results are shown in Section 6, and detailed analyses and discussions are made in Section 7. Finally, we make some useful conclusions in Section 8 and present some ideas about future work.

2. RELATED WORK

2.1. Pseudo Relevance Feedback

Pseudo relevance feedback via query expansion is referred to as the techniques or algorithms that reformulate the original query by adding new terms and adjusting their weights, in order to obtain a better query. It usually assumes top-ranked documents in the first-pass retrieval to be relevant. These top documents are used as feedback documents to add potentially related terms to original queries. Although the feedback documents are "pseudo," PRF has been shown to be effective with various retrieval models [Rocchio 1971; Lavrenko and Croft 2001; Carpineto et al. 2001; Cao et al. 2008; Lv and Zhai 2010; Ye et al. 2011; Miao et al. 2012; Ye and Huang 2014; Collins-Thompson 2009; Raman et al. 2010; White and Marchionini 2007; Xu and Croft 2000]. There are a large number of studies on the topic of PRF. Here we mainly review the work about PRF that is the most related to our research.

Rocchio's model [Rocchio 1971] is a classic framework for implementing (pseudo) relevance feedback via improving the query representation. It models a way of incorporating (pseudo) relevance feedback information into the vector space model (VSM) in IR. In Rocchio's model, a set of documents are utilized as the feedback information, and unique terms in this set are ranked in a descending order of their TFIDF weights. It has shown that Rocchio's performance is at least comparable with the state-of-the-art relevance models [Zhai 2008] in the language model framework when it is combined with BM25. In this article, we will conduct our research for effectively utilizing the topical information for PRF in the probabilistic IR framework.

To compare our experimental results with PRF methods in the language model framework, we choose RM3 [Lavrenko and Croft 2001; Lv and Zhai 2010] as a baseline, which is strong and widely used in previous work [Yi and Allan 2009; Lv and Zhai 2010; Ye et al. 2011]. It is a representative and the state-of-the-art approach for re-estimating query language models for PRF [Lv and Zhai 2010; Yi and Allan 2009]. Relevance language models do not explicitly model the relevant or pseudo relevant document. Instead, they model a more generalized notion of relevance R. Lv and Zhai [2010] systematically compared five state-of-the-art approaches for estimating query language models in ad hoc retrieval, in which RM3 not only yields impressive retrieval performance in both precision and recall metric but also performs steadily. In particular, we apply Dirichlet prior for smoothing document language models [Zhai and Lafferty 2004]. More details about Rocchio and RM3 will be introduced in Section 5.2.

2.2. Topic Modeling

Recently, probabilistic topic models have become more and more popular in the text mining area [Hofmann 1999; Blei et al. 2003a; Li and McCallum 2006]. These models are able to discover the latent semantic schemes in a group of documents. The basic idea of topic modeling is that the vocabulary of a document is generated from topics, and topics are represented as different probability distributions of terms in the vocabulary. A term can have various probabilities in different topics.

Probabilistic Latent Semantic Indexing (PLSI) [Hofmann 1999] was proposed in 1999, which is a significant step in the development of topic models. It models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics" [He 2011]. Thus, each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a set of topics. Latent Dirichlet Allocation (LDA) [Blei et al. 2003a] is another popular topic model that also assumes that there are topics in the

corpus, but a document can have more than one topic. LDA has a more complicated probabilistic procedure of generating a document. We applied this topic model for our methods in this article. More details will be introduced in Section 4.1. Another state-of-the-art topic model named the Pachinko Allocation Model (PAM) [Li and McCallum 2006] was proposed in 2006. Unlike PLSI and LDA, topics are not considered to be independent. The four-level PAM model utilizes a super-topic layer in a directed acyclic graph to model the correlations among topics.

Researchers also noticed that topic number is a key factor. How to optimize this parameter has attracted a lot of attention [Blei et al. 2003a; Griffiths and Steyvers 2004; Blei and Tenenbaum 2004]. Blei et al. [2003a] used perplexity to assess the generalizability of models across corpora in computational linguistics. Griffiths and Steyvers [2004] applied a Bayesian model selection method to choose parameters that can maximize the posterior probability of the model. Blei and Tenenbaum [2004] used nonparametric Bayesian statistics to automatically select topic numbers, and Teh et al. [2012] applied Dirichlet processes to reduce parameters for LDA. While most optimizations are goal oriented (e.g., minimizing perplexity), it is difficult to propose a general way for obtaining the best topic number. In addition, whether topics obtained through these methods will improve the performance of integrating topic information in IR, especially under the probabilistic framework, is still largely unexplored [Jian et al. 2016]. Topic modeling has been widely used in the areas of summarization [Bian et al. 2013], detection [Chen et al. 2013], and word embedding [Liu et al. 2015], as well as many other areas.

2.3. Topic Modeling for Pseudo Relevance Feedback

Topic models have been applied in PRF recently, but not all of the applications are effective. Yi and Allan [2008] attempted to apply different topic models for the retrieval purpose, and the proposed PRF methods CBQE, LBQE, and PBQE were all worse than the state-of-the-art RM method. LDA-RM in Yi and Allan [2009] can outperform RM in some cases but cannot obtain sustainable improvements over the latter. Andrzejewski and Buttler [2011] utilized LDA to generate latent topics and identified the latent topics that are potentially relevant, which are further manually selected by users. Then the terms that are most strongly associated with the elected topic are used to expand the original query. Ye et al. [2011] proposed three methods to obtain the most relevant topics and select feedback terms from them. Significant improvements have been made, and they found that the proposed methods performed much better in the simulated relevance feedback. Their research work indicates that the performance of topic-model-based PRF methods depends on the quality of the corpus from which topics are obtained. However, when the topic number changes, the performance of the proposed Top_k method drops significantly. Caballero and Akella [2012] incorporated topical information of relevant documents and irrelevant documents in active relevance feedback and obtained promising performance on the medical OHSUMED dataset. Their research shows that topical information can be useful in domain-specific search. Wang et al. [2012] assumed terms should be in relevant topics before and after being translated and used LDA-based PRF for the cross-language retrieval task. They used the same strategy to select topics as in Ye et al. [2011] and obtained marginal improvements. Serizawa and Kobayashi [2013] found that the precision of the topic-based relevance feedback method can be better than the word-based relevance feedback model in particular cases. But the overall performance is not good. Kotov et al. [2013] proposed the first work that leveraged geographical metadata to perform geographically focused document, query, and relevance model expansion on the Microblog collection with the application of LDA. They obtained 800 topics and combined the topic-based model with the Relevance Model. Zhiltsov and Agichtein [2013] applied cosine

similarity between latent semantic representations of PRF entity documents as features. Significant improvements were made on the entity search task. Their idea was trying to find the latent semantic relations among entities using nonnegative matrix factorization with the top three PRF entity. Their motivation is similar to ours, but they do not investigate how to make this by topic modeling.

Although the work in Wei and Croft [2006] was done mainly on the first-pass retrieval, it is still worth mentioning. They found that "LDA itself may be too coarse to be used as the only representation for IR." So they applied LDA and interpolated the Dirichlet language model by integrating the probability of generating a term through topics in a document. This finding is consistent with our investigation on the "fuzzy topic" problem, and they solved this problem by combine the term-based method with the topic-based language model. Compared to their work, our research goes deeper on integrating the coarse topical information effectively for PRF, especially under the probabilistic framework. Using topic space to represent documents is different from the generative view in the language model framework, and it is convenient to be extended with different space-based methods. In this research, we do not use offline LDA as they did so that we can easily extend our work on large collections and avoid the challenge of choosing the topic number.

To the best of our knowledge, previous work on integrating topic models in IR focused mainly on how to find the most relevant topic(s) and did not concern the fuzzy topic problem [Andrzejewski and Buttler 2011; Ye et al. 2011; Wang et al. 2012; Serizawa and Kobayashi 2013]. In fact, the problem is very important and has an extensive impact on the relevant topics they pursue. Some work is on human-involved relevance feedback [Andrzejewski and Buttler 2011; Caballero and Akella 2012]. This makes the utilization of topic models very expensive and time-consuming. Also, only little study is about how to integrate topical information into the probabilistic PRF model. Almost all the studies import extra parameters for interpolations. While topic modeling is complicated, more parameters will increase the complexity and therefore cost more computing resources.

Compared with previous studies, the uniqueness of our proposed methods is threefold. First, we focus on how to effectively apply topical information and avoid extensive fluctuation of performance with different topic numbers. Second, we only use a small group of documents so that the short processing time of topic modeling will not affect the efficiency of PRF. Finally, no new parameters are imported into our methods and therefore will not cost more time on searching optimal values for them.

3. PRELIMINARY STUDY OF "FUZZY TOPIC" IN PSEUDO RELEVANCE FEEDBACK

Before incorporating topic space in PRF, we study the characteristics of latent topics generated by LDA in this section. Section 3.1 analyzes how the terms in the topics change according to selections of topic numbers and how we plan to deal with such issues in PRF. In Section 3.2, we discuss the challenges of integrating topic space in PRF and how to find a more robust way for addressing the challenges.

3.1. Observations of Fuzzy Topic

We will show how fuzzy topics are obtained when changing the topic number for a very popular topic model, LDA [Blei et al. 2003a]. To show how a topic changes with the topic number, we choose the topic that is most likely to generate the query. This is the traditional way used in previous studies [Andrzejewski and Buttler 2011; Ye et al. 2011; Serizawa and Kobayashi 2013]. Then we check how it changes with different topic numbers. All the experiments presented in this section are done on the TREC

GOV2 collection with official queries.[2] We still use BM25 [Robertson et al. 1994], a classic probabilistic model, as the first-pass retrieval model. The top 30 documents are chosen for the LDA model. The topic number is set to five, 10, and 30, and we show the top 50 terms ranked by their probabilities given the topic. All terms are processed by Porter's Stemmer [Porter 1980]. We randomly choose two queries, 802 and 804, as examples and fix parameters except the topic number for all the experiments. So the top 30 documents are the same for LDA.

Although all the topics we obtain are assumed to be the "most relevant," they are different. If we attempt to choose feedback terms from them, we will be somewhat confused due to the ranks of these terms. As we can see from Table I, when the topic number changes, the term lists are different. Usually, 10 to 50 terms will be chosen as feedback terms for query expansion. However, when we just increase the topic number from five to 10, even the list of the top 10 terms in the table changes significantly. For example, when the topic number is 10, "ash" is ranked 36th in the topic of query 802. But when the topic number is set to five and 20, its rank raises to 10th and seventh, respectively. How can we evaluate the relevance of "ash" in this case?

If we change our view from the "terms" in these topics to "topics" themselves, there is something different. As we consider, a topic can represent a particular kind of information. When we change the topic number, the information can be divided into two or three parts or be combined with others. The similarity of two documents on a particular topic will be reflected more or less in the new coordinate system and identify their closeness in the topic space. As in the example given in Section 1.2, when two documents are both mainly about one topic, they will also have its subtopics when the topic is split into two parts. Meanwhile, we use the full topic coordinates of documents to measure the reliability and adapt them into the traditional term-based framework as an enhancement. In Section 7.1, we will show that the ranks of documents are more stable when the topic number changes. Compared to the conditional probability of a topic given the document, the internal changes of terms in this topic are more frequent and significant. Therefore, for these "fuzzy topics," it is hard to tell which term should be more relevant than the other while their relative ranks are not stable. If a researcher attempts to choose a term based on the rank or probability information in a given topic, it is very difficult for him or her to make the choice.

3.2. Challenges of Integrating Topic Space

Traditional term-based PRF approaches choose feedback terms from top feedback documents and do not consider whether these documents are entirely relevant, partially relevant, or not relevant to the query. To solve this problem, researchers have used topic modeling to extract the topics from text collections and choose feedback terms from the particular topic(s). Different strategies have been used to identify the topic(s) related to the original query, and terms appearing with higher probabilities in the topic(s) are regarded to be more semantically related to the original query. The advantage of this type of method is breaking down the constraint of document scope. Because topic modeling considers the co-occurrence of terms within the whole collection for training, the relation among terms across documents can be considered. For example, if a term t_1 always appears frequently with given query terms, t_1 is highly likely to be relevant to the original query and has a high probability in the query-related topics. If there is another term t_2 that co-occurs with t_1 in other documents, t_2 will have a high probability in these query-related topics too. It is possible that t_2 does not co-occur with

[2]In this article, in order to avoid confusion, "topic" only refers to topics obtained through topic modeling. We do not use "topic" to represent queries for all TREC datasets as some previous research does. Instead, we use the term "query."

Table I. Topics Given Topic Numbers of 5, 10, and 30 on Query 802, "Volcano Eruptions Global Temperature," and Query 804, "Ban on Human Cloning"

Term Ranks	Volcano Eruptions Global Temperature			Ban on Human Cloning		
	5	10	30	5	10	30
1	volcano	volcano	volcano	clone	clone	clone
2	nbsp	can	can	human	human	human
3	erup	volcan	flow	embryo	research	ban
4	volcan	flow	activ	ban	embryo	research
5	magma	activ	mount	research	us	embryo
6	can	erup	hazard	us	ban	us
7	flow	earthquak	**ash**	cell	cell	cell
8	activ	mount	magma	quot	will	state
9	**ash**	usg	water	will	produc	will
10	earthquak	magma	pyroclast	reproduct	purpos	creat
11	mount	hazard	scientist	state	creat	produc
12	gase	scientist	gase	mai	moral	purpos
13	second	rock	erupt	creat	onli	new
14	rock	lava	lava	stem	stem	legisl
15	usg	gase	peopl	therapeut	prohibit	onli
16	gas	erupt	state	be	mai	be
17	hazard	pyroclast	hot	purpos	be	prohibit
18	mb	peopl	lahar	onli	who	reproduct
19	lava	water	like	who	new	allow
20	monitor	caus	mile	act	transfer	who
21	water	hot	chang	moral	act	mai
22	1	dioxid	near	new	allow	need
23	erupt	debri	rock	ethic	need	stem
24	0	area	hawaii	prohibit	ethic	feder
25	dioxid	lahar	type	transfer	medic	act
26	scientist	mile	move	allow	legisl	life
27	temperatur	occur	includ	creation	feder	prohibi
28	degre	chang	st	legisl	attempt	medic
29	caus	st	ground	reason	wai	develop
30	pyroclast	monitor	earth	believ	time	time
31	measur	helen	cascad	scientif	issu	attempt
32	vent	like	caus	medic	call	issu
33	state	state	observatori	hous	practic	transfer
34	pressur	alaska	debri	genet	effect	support
35	geolog	ground	mudflow	prohibi	embryon	1
36	alaska	**ash**	cloud	attempt	state	congress
37	hot	move	system	life	life	creation
38	includ	peak	sulfur	support	now	call
39	peopl	cascad	form	like	creation	effect
40	debri	increas	locat	technolog	scientist	first
41	st	includ	danger	work	support	ethic
42	lahar	hawaii	aircraft	year	like	law
43	move	movi	onli	question	peopl	practic
44	ground	type	helen	individu	egg	technolog
45	helen	near	thousand	effect	prevent	reproduc
46	hawaii	geolog	hundr	reproduc	2	now
47	mai	continu	washington	embryon	reason	requir
48	area	carbon	call	wai	requir	genet
49	occur	produc	surfac	benefit	reproduc	2
50	type	system	crater	requir	gener	live

the query terms frequently because query terms can have synonyms. In this case, t_2 is also a good feedback term but cannot be identified by traditional term-based PRF approaches. With topical information, t_2 can be identified and utilized in PRF. However, the topic modeling approaches have randomness and generate different groups of topics if performed multiple times.

Therefore, we investigate this "fuzzy topic" issue and plan to find a more robust way to deal with it. We do not focus on identifying particular topics. Instead, we use topical information to evaluate the quality of feedback documents rather than selecting feedback terms. That is an alternative way to improve PRF. So we propose the "topic space," which uses topics as coordinates and represents feedback documents as topic vectors. The idea is inspired by the vector space model [Salton et al. 1975a] for the term-based document representation. One advantage is that we can apply similarity methods for documents on the topical level. Different from traditional term-based methods, topical information can help to discover the latent semantic similarity among documents. Our proposed idea can be a good complement for the current term-based PRF methods like Rocchio. Another advantage is that we do not deal with the dilemma of choosing topics. Our ultimate goal is to evaluate the relevancy between documents and the query. Therefore, we can address the fuzzy topic issue by using all topics to represent documents. In the next section, we will propose a new probabilistic framework called TopPRF and present how to integrate the topic space into PRF naturally instead of by dimension reduction.

4. INTEGRATING TOPIC SPACE INTO PSEUDO RELEVANCE FEEDBACK

LDA is widely used in the text mining area [Wang and Blei 2011; Tang et al. 2008; Mei et al. 2007; Griffiths and Steyvers 2004; Yi and Allan 2009], and the retrieval performance using LDA is even better than another popular model, PAM, in PRF [Yi and Allan 2009]. We therefore apply LDA to obtain latent topics for our proposed methods. In the rest of this section, we will introduce how to build the topic space via the LDA model, explain how we use topics for PRF, and propose three methods to select good feedback terms in detail.

Another issue we consider is whether we should apply LDA on the whole collection. Building topics on the whole corpora will take a large amount of time. When the size of the collection is very big, it is even harder to determine the number of topics. It should be a large number because we have millions of documents. Because of the "fuzzy topic" problem, we are not able to determine the best topic number. In addition, it can be different for various queries. Although offline topic modeling may save time for further usage, this problem cannot be avoided. Besides, even if we do an offline LDA on a medium dataset like WT10G, which contains 1,692,096 documents, it is reasonable to set a large topic number like 5,000. Otherwise, the topics we obtain will be very general in information. When we compare two documents, most features in their topic vectors will be close to 0; that is, their topic vectors are very sparse. In this case, their similarities will always be near 1 when we calculate similarity scores. Consequently, documents cannot be identified through topical information. Experimental results reported in Yi and Allan [2008] confirmed that topics discovered on the whole corpora are too coarse-grained for query expansion. In Section 6.4, we will show some experimental results over 100 and 200 topics that are obtained through offline LDA on the whole corpus. While the resource cost is much higher than our online strategy on the top documents, the performance is not improved. So in this paper, we only use the top k feedback documents as the source of LDA.

In this section, we will first introduce the LDA model, which will be used for generating topic space, and then present how we integrate the topic space into the new framework, TopPRF. Based on this framework, we will describe how we measure the

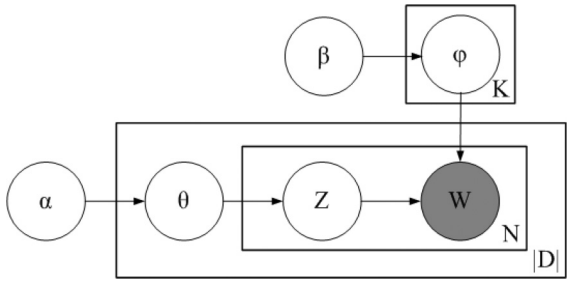


Fig. 2. Plate notation for the LDA model.

reliability of feedback documents through three newly proposed models: TS-COS, TS-EU, and TS-Entropy. These three models are built according to our assumptions in Section 1.2. TS-COS and TS-EU are two similarity-based models designed for Assumption 1, and TS-Entropy is designed to measure the purity of topical information in a document for Assumption 2.

4.1. Generating Topic Space via LDA

The LDA model [Blei et al. 2003a] deals with the problem of modeling text corpora and other collections of discrete data. LDA is applied to find short descriptions for documents of a collection and enable discovering and preserving essential statistical relationships that are useful for classification, document diversify, summarization, similarity, and relevance judgments.

When modeling text corpora, it can automatically cluster documents into mixtures of topics. Each topic is characterized by a distribution over words. In particular, the LDA model can automatically assign each document a probability distribution over topics and assign the topic distributions over the words. For example, different probabilities are assigned to the words for different topics. In our case, for the set of feedback documents to a given query, it may also contain plenty of topics.

The LDA model assumes the following generative process for each document d in a document collection D:

(1) Pick a multinomial distribution Φ_z for each topic z from a Dirichlet distribution with hyperparameter β. β is the parameter of the uniform Dirichlet prior on the per-topic word distribution.
(2) For each document d, pick a multinomial distribution θ_d from a Dirichlet distribution with hyperparameter α. α is the parameter of the uniform Dirichlet prior on the per-document topic distributions.
(3) For each word w in document d:
   (a) Choose a topic $z_n \sim Multinomial(\theta_d)$, where $n \in \{1, 2, \ldots, K\}$.
   (b) Choose the word w from the multinomial distribution of φ_{z_n}.

Thus, the probability of generating the collection D is given as the following:

$$P(d_1, \ldots, d_{|D|}|\alpha, \beta) = \iint \prod_{z=1}^{M} P(\varphi_z|\beta) \prod_{d=1}^{|D|} P(\theta_d|\alpha)(\prod_{i=1}^{N_d} \sum_{z_i=1}^{M} P(w_i|z, \varphi))d\theta d\varphi,$$

where $|D|$ is the number of documents in dataset D, N_d is the number of words in document d, and K is the number of topics in the LDA model. Figure 2 depicts the plate notation for the LDA model, which can capture the dependencies among all the variables.

The LDA model is complicated and intractable to compute. Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximate, variational approximation, Markov chain Monte Carlo [Walsh 2004], and Gibbs sampling [Geman and Geman 1984]. In this article, we use LingPipe's[3] implementation of LDA, in which the LDA model is estimated using a simplified form of the Gibbs sampler [Porteous et al. 2008]. The following probabilities can be generated from LDA: $P(z_k|d)$, $P(w|z_i)$, and $P(z_i)$, where $i \in \{1, 2 \dots M\}$ and M is the number of topics. In this article, we use $P(z_k|d)$ as the coordinates in the topic space. Documents are represented as vectors of length M and evaluated by their topic representations for PRF in Sections 4.2, 4.3, 4.4, and 4.5.

4.2. A Probabilistic Framework: TopPRF

As far as we know, little work has been done on how to integrate topical information into probabilistic PRF models naturally and effectively. So we implement our idea on the classic Rocchio model, and BM25 is utilized as the basic model. According to Zhai [2008], "BM25 [Robertson et al. 1994] term weighting coupled with the Rocchio's feedback model remains a strong baseline which is at least as competitive as any language modeling approach for many tasks." This observation is also supported in our preliminary experiments of this article. The effectiveness and flexibility of the Rocchio model make it very suitable for extensions. So we decide to integrate the topic space information into it and propose a better framework. We name the new framework TopPRF, and "Top" denotes the integration of topics.

The traditional Rocchio model is based on the term vector space, and it is proven to be effective. So we keep this term information and utilize it together with information from the topic space. Matches on the term level are accurate but rigid. On the topical level, matches can bring something different while they are actually processed on groups of terms instead a particular one. So we consider the cooperation of the two spaces as promising. As we mentioned earlier, we plan to adjust the weights of candidate feedback terms according to the topical information we have. The TopPRF framework is shown as follows:

$$\vec{Q}_1 = \alpha * \vec{Q}_0 + \beta * \sum_{\vec{r} \in D_R} \frac{\vec{r} * TS(d_k)}{|R|}, \tag{1}$$

where $\vec{Q}_1$ and $\vec{Q}_0$ represent the original and first iteration query vectors, D_R is the set of pseudo relevance documents, $\vec{r}$ is the expansion term weight vector, R is the set of feedback documents, $|R|$ is the number of feedback documents, d_k is the kth feedback document in R, and $TS(d_k)$ is the score of feedback documents based on our proposed topic space methods.

In fact, we only add one factor into the Rocchio model without importing new parameters. So we will not suffer the resource-consuming process of new parameter optimization and use much time on sampling or grid search. Although we only propose three models as follows, the framework can be easily extended with different $TS(d_k)$. This will encourage researchers to discover more good models. In $\vec{r}$, we use *tf-idf* as in the traditional Rocchio model. Particularly, we set α to 1 and train an optimal β in the experiments. More details can be found in Section 5.3.

The complexity of each iteration of the Gibbs sampling for LDA is linear with the number of topics and the number of documents [Wei and Croft 2006]. In our case, the time complexity is denoted as $O(M^*|R|)$. For each query, we use at most 50 feedback

[3]http://alias-i.com/lingpipe/.

documents and 20 topics in LDA, which is constant time O(1) to the size of the collection. Therefore, our proposed approaches do not bring higher computational complexity in theory. We will have more discussion about the time cost of our proposed approaches in the experiments.

4.3. TS-COS: Measuring Topic Similarity via Cosine Formula

When integrating LDA on the top k documents $d_1, d_2, \ldots d_k$ in the retrieved list, we will have M topics $z_1, z_2, \ldots, z_M$ and a $k * M$ Document-Topic matrix that contains the probability $p(z_i|d_n)$, where n is from 1 to k and i is from 1 to M. For each document d_n, we consider it to be a probability vector of topics $p(z_1|d_n)$, $p(z_2|d_n), \ldots, p(z_M|d_n)$ in the topic space, and the sum of all elements in this vector is 1. Because a document is the mixture of these M topics, the topic distribution can reveal the topic bias of this document. Moreover, if a relevant document of a particular query has a great bias toward some topics (e.g., z_i and z_j), it is natural to consider that other documents with the same bias are more likely to be relevant.

We can use relevant documents as examples and measure the topic similarities between them and other documents. In our study, we first assume the top s documents are relevant. Different from the top k documents in traditional PRF, we attempt to ensure the s documents to be really relevant, so we must choose a very small value for it. The s documents are part of the feedback documents, and we use them to evaluate the reliability of the rest of the documents. We cannot guarantee that all the s documents are really relevant since the process is still pseudo. But according to our preliminary experiments on WT10G, smaller s can lead to a higher ratio of relevant documents in the group. We consider this s document group as the *trustable group*. Sometimes the relevant documents for a particular query cover several topics. In order to maintain the balance of document relevance and topic diversity, we first choose three as a reasonable number for s in our research and then try different values to see how s impacts the performance.

For the proposed method named TS-COS, we measure the similarity between the topic vectors of two documents via the cosine formula. Thus, the topic similarity of documents i and j is as follows:

$$cos(d_i, d_j) = \frac{\sum_{t=1}^{M} (p(z_t|d_i) * p(z_t|d_j))}{\sqrt{\sum_{t=1}^{M} (p(z_t|d_i))^2} \times \sqrt{\sum_{t=1}^{M} (p(z_t|d_j)^2}}, \tag{2}$$

where z_t is the tth topic, and M is the total topic number.

When we set $s = 3$, $TS(d_k)$ is calculated as follows:

$$TS(d_k) = \frac{\sum_{i=1}^{3} cos(d_k, d_i)}{3}, \tag{3}$$

where the scores of the top three documents are set to 1 since they are supposed to be relevant. $TS(d_k)$ is normalized to (0.5,1) to avoid a significant difference between two documents when $k > 3$.

Intuitively, we think all the feedback documents are somewhat relevant while they are the top-ranked ones obtained through models like BM25. More or less, they have something related to one or more query terms. After all, the process of weighting feedback documents is still pseudo. We are not 100% sure about the relevance of the sample documents. So we set the floor of the final similarity scores to 0.5 to narrow down the differences between the best and the worst documents. The normalization is simple. We just divide the normal cosine similarity by 2 and plus 0.5.

4.4. TS-EU: Measuring Topic Similarity via Euclidean Distance

In order to see how different similarity methods perform, we change the cosine similarity method to another popular one, Euclidean distance. For this method, we actually consider documents as the points in the topic space. The distance between two document points indicate their closeness, or similarity. The distance between document i and j in the topic space is as follows:

$$Euclidean(d_i, d_j) = \frac{\sqrt{\sum_{t=1}^{M} (p(z_t|d_i) - p(z_t|d_j))^2}}{M}, \tag{4}$$

where z_t is the tth topic, and M is the total topic number.

Equation (1) is still used to choose feedback terms. Unlike Equation (2), large Euclidean distance between two documents means they are not similar. So when $s = 3$, $TS(d_k)$ is calculated as follows:

$$TS(d_k) = 1 - \frac{\sum_{i=1}^{3} Euclidean(d_k, d_i)}{3 \times M}. \tag{5}$$

We name this method TS-EU as EU represents Euclidean distance and $TS(d_k)$ is also normalized to [0.5, 1] as in Equation (3).

4.5. TS-Entropy: Measuring Document Purity via Topic Entropy

When we use the feedback documents for query expansion, the whole content of each document is supposed to be relevant, which is not true in most cases. Terms in the irrelevant part of a document can bring useless information and harm the performance of PRF. In general, we consider that documents that contain few topics are more reliable for choosing feedback terms. In other words, terms in a "pure" feedback document are likely to be relevant than those from a document containing multiple topics. To measure the purity of a document, we import a new concept called *topic entropy*.

Entropy is an important concept in the information theory. It is a measure of unpredictability of information content. Suppose a discrete random variable X has possible values $x_1, \ldots, x_n$; then the entropy of X can be calculated as

$$H(X) = \sum_i P(x_i) I(x_i) = -\sum_i P(x_i) \log_b P(x_i), \tag{6}$$

where b is the base of the logarithm used.

If we consider a document d as the variable X, its value depends on the topic it is about. So x_i in Equation (6) is actually topic z_t given d. The topic entropy of document d is calculated as

$$H(d) = \sum_{t=1}^{M} P(z_t|d) I(z_t|d) = -\sum_{t=1}^{M} P(z_t|d) \log_b P(z_t|d). \tag{7}$$

Large topic entropy means multiple topics have nonignorable probability given d and this document is not pure. Documents with small topic entropy should therefore be more reliable than others for PRF. Again, we modify Equation (10) to apply this TS-Entropy method.

$TS(d_k)$ in Equation (1) is calculated as follows:

$$TS(d_k) = 1 - \frac{H(d)}{\log_b M}, \tag{8}$$

where the value is normalized to [0, 1], and b is set to 2 in this article.

Table II. Information About the Test Collections

Collection	Queries	# Docs
disk1&2	51–200	741,856
disk4&5	301–450, 601–700	528,155
WT2G	401–450	247,491
WT10G	451–550	1,692,096
GOV2	701–850	25,178,548

5. EXPERIMENTAL SETTINGS

In this section, we describe the settings in our experiments. Experiments are conducted on five standard TREC datasets described as in Section 5.1. Section 5.2 presents three baselines models for comparison with our proposed approaches, including the classic BM25 and two state-of-the-art approaches, RM3 and BM25-based Rocchio. In Section 5.3, we discuss the metrics for evaluation, how to set and optimize the parameters in the baselines, and how to train the parameters in our proposed approaches.

5.1. Collections

We evaluate our proposed methods on five public TREC[4] datasets with ad hoc queries, including Disk1&2, Disk4&5, WT2G, WT10G, and GOV2, which are different in size and genre. The Disk1&2 and Disk4&5 collections contain newswire articles from various sources, such as the Associated Press (AP), Wall Street Journal (WSJ), and Financial Times (FT), which are usually considered high-quality text data with little noise. The WT2G collection is a general crawl of web documents, which has 2 gigabytes of uncompressed data. This collection was used in the TREC 8 web track. The WT10G collection is a medium-size crawl of web documents, which was used in the TREC 9 and 10 web tracks. It has 10 gigabytes of uncompressed data. GOV2 is a very large crawl of the .gov domain, which has more than 25 million documents with an uncompressed size of 423 gigabytes. The TREC tasks and query numbers associated with each collection are presented in Table II.

Queries for these datasets are provided by TREC in the past 10 years, and the datasets are widely used for IR [Zhao et al. 2011; Zhai and Lafferty 2004; Culpepper et al. 2014; Zhao et al. 2014; Cummins et al. 2015]. We only use the title part of the queries to retrieve because users usually only input several keywords when searching in the real world. For the preprocessing of the collections, we use the Porter Stemmer [Porter 1980] and general stopword remover [Allan et al. 2000] with 418 stopwords removed.

For implementation, there are several well-known open-source IR systems supporting the probabilistic retrieval models. For instance, the Lemur project [Strohman et al. 2005] develops the Lemur Toolkit and the Indri search engine, which combines the inference nets and language modeling in an architecture designed for large-scale applications. The Terrier search engine [Ounis et al. 2006] implements indexing and retrieval functionalities and combines ideas from probabilistic theory, statistical analysis, and data compression techniques. The proposed approaches can be implemented on any of these IR systems.

5.2. Baselines in Comparison

In the experiments, we compare our proposed methods with the traditional probabilistic model, BM25, and two state-of-the-art pseudo relevance feedback models, RM3 and Rocchio. These three baselines are described as follows.

[4]http://trec.nist.gov/.

BM25 is a famous traditional weighting model, which has been recognized for its good performance in IR. In BM25, the weight of a search term is assigned based on its within-document term frequency and query term frequency [Robertson et al. 1994]. The corresponding weighting function is as follows:

$$BM25 = \frac{(k_1+1)*tf}{K+tf} * \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} * \frac{(k_3+1)*qtf}{k_3+qtf} \\ \oplus \quad k_2 * nq * \frac{(avdl-dl)}{(avdl+dl)}, \tag{9}$$

where w is the weight of a query term, N is the number of indexed documents in the dataset, n is the number of documents containing a specific term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is the within-document term frequency, qtf is the within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is are tuning constants, K equals $k_1 * ((1-b)+b*dl/avdl)$, and $\oplus$ indicates that its following component is added only once per document, rather than for each term.

Rocchio's method has the following steps to incorporate (pseudo) relevance feedback information into the retrieval process [Ye and Huang 2016]:

(1) All documents are ranked for the given query using a particular information retrieval model. For fair comparison, we use the BM25 model (Equation (9)) in this article. This step is the first-pass retrieval. The $|D_f|$ highest-ranked documents are identified as the pseudo relevance set D_f.
(2) An expansion weight $w(t, D_f)$ is assigned to each term appearing in the set of the D_f highest-ranked documents. In general, $w(t, D_f)$ is the mean of the weights provided by a weighting model, for example, the TF-IDF weighting model ([Salton et al. 1975a]) in this article.
(3) The vector of the query term weight is finally modified by taking a linear combination of the initial query term weights with the expansion weight $w(t, D_f)$ as follows:

$$Q_1 = \alpha * Q_0 + \beta * \sum_{r_i \in D_f} \frac{r_i}{|D_f|}, \tag{10}$$

where Q_0 and Q_1 represent the original and first iteration query vectors, r_i is the expansion term weight vector for the ith feedback document, $|D_f|$ is the number of feedback documents for PRF, and α and β are tuning constants controlling how much we rely on the original query and the feedback information. When we obtained a list of documents through BM25, the top k documents were chosen as feedback document set D_R. After calculating the $tf*idf$ score for each term in each feedback document, we have a set of vectors $\vec{r}$. Then we can figure out the new query vector $\vec{Q}_1$ through Equation (10). Because $\vec{Q}_1$ is a vector containing the scores of all the terms, we usually choose a few top-ranked terms as feedback terms to expand the original query. In the rest of the article, we use Rocchio to represent this BM25-based Rocchio.

RM3 is an interpolated version of the relevance model [Lavrenko and Croft 2001], which is a representative and state-of-the-art approach for re-estimating query language models for PRF [Lv and Zhai 2009]. Relevance language models do not explicitly model the relevant or pseudo relevant document. Instead, they model a more generalized notion of relevance R. The formula of RM1 is

$$p(w|R) \propto \sum_{\theta_D} p(w|\theta_D)p(\theta_D)P(Q|\theta_D). \tag{11}$$

The relevance model $p(w|R)$ is often used to estimate the feedback language model θ_F and then interpolated with the original query model θ_Q in order to improve its estimation as follows:

$$\theta_{Q'} = (1 - \alpha) * \theta_Q + \alpha * \theta_F. \tag{12}$$

This interpolated version of the relevance model is RM3. Lv and Zhai [2009] systematically compare five state-of-the-art approaches for estimating query language models in ad hoc retrieval, in which RM3 not only yields impressive retrieval performance in both precision and recall metric but also performs steadily. In particular, we apply a Dirichlet prior for smoothing document language models [Zhai and Lafferty 2001].

5.3. Parameter Settings and Optimization

Particularly, for the basic retrieval model, we use the BM25 model and search optimal b from 0.1 to 0.9 with the step 0.1. In addition, we search β from 0.1 to 0.9 with the step 0.1 for the Rocchio model as well. Because we will evaluate the impact of topical information on the feedback document, we fix the feedback term number to 30. Feedback document numbers are set to 10, 20, 30, and 50, respectively. The number of topics is set to be five, 10, and 20, which are reasonable when the feedback document number is not large. The commonly used Mean Average Precision (MAP) is the metric for evaluations. The MAP metric reflects the overall accuracy, and the detailed descriptions for it can be found in Voorhees and Harman. [2000]. A language model with a Dirichlet prior is used as the basic model for another baseline, RM3. For the smoothing parameter μ, we sweep over values from 500 to 2,000 with an interval of 100. The interpolation parameter α for RM3 is set from 0.1 to 0.9 with the step 0.1. All the experimental results are evaluated through twofold cross-validation. The TREC queries are partitioned into two sets by the parity of the TREC queries' number on each dataset. Parameters trained on one set are applied to the other set and then vice versa for evaluation, as in Ye et al. [2011]. In our experiments, we use Okapi BSS (Basic Search System) [Beaulieu et al. 1997; Robertson and Walker 1994] as our main search system and conduct our information retrieval experiments using the improved Okapi system [Huang et. al 2005; Huang et al. 2006; Huang and Hu 2009; Miao et al. 2012; Yin et al. 2013; Zhao et al. 2014; Ye and Huang 2014, 2016].

6. EXPERIMENTAL RESULTS AND ANALYSES

This section presents the experimental results and compares the proposed approaches with the baselines. In Section 6.1, we demonstrate that the classic BM25 is a reasonable basic model for PRF in our proposed approaches. For fair comparison, BM25 is also adopted in Rocchio with the same settings. The performance of baseline models is shown in Section 6.2, and the performance of the proposed topic-space-based approaches is discussed in Section 6.3. We study the impact of using LDA over the whole collection for PRF in Section 6.4. Further analyses about the experimental results are provided in Section 6.5.

6.1. Comparison of Basic Retrieval Models

As we mentioned in the previous section, the results of both models are obtained by twofold cross-validation with optimal parameters. It is therefore fair to compare them on these five collections. As shown in Table III, BM25 slightly outperforms LM with a Dirichlet prior on the Disk1&2 and WT2G collection. The results of these two models are almost the same over the Disk4&5, WT10G, and GOV2 collections. This comparison indicates that the classic BM25 model is generally comparative to LM, and it is reasonable to use them as the basic models of the PRF baselines and our proposed methods.

Table III. BM25 vs. LM on the Five TREC Collections

	disk1&2	disk4&5	WT2G	WT10G	GOV2
BM25	0.2380	0.2494	0.3124	0.2055	0.3034
LM	0.2320	0.2510	0.2995	0.2063	0.3040

6.2. Performance of Baseline Models

All the experimental results are shown in Table IV and Figure 3. $|D_f|$ is the number of feedback documents for PRF. Rocchio in Table IV actually denotes Rocchio's model with BM25 as the first-pass retrieval model, and RM3 denotes LM+RM3. As we can see from Table IV, the Rocchio model generally outperforms BM25 in most cases. On these five collections, the Rocchio model achieves its best performance on four of the five collections and the second-best performance on the GOV2 collection when $|D_f|$ is 20. When more feedback documents are chosen, the performance drops dramatically and is even worse than BM25 on WT2G and WT10G. This indicates that documents are more and more unreliable when their ranks are lower.

RM3, which is a state-of-the-art model PRF for language modeling, generally outperforms the Rocchio model on the WT2G and WT10G collections, but not very significantly. This indicates that the Rocchio model is still a very strong baseline for IR research work. Compared to the Rocchio model, the results of RM3 are more stable when the number of feedback documents changes.

6.3. Performance of Topic-Space-Based Models

Although our proposed methods are all based on the Rocchio model, we can see that their performance is quite different. In most cases, TS-COS achieves the best results. It is only surpassed by RM3 on the WT10G collection, while its base model, Rocchio, does not work well. Even under that condition, the average performance of TS-COS is just slightly weaker than RM3's. Also, its average performance outperforms that of the Rocchio model on all collections significantly. These experimental results justify the effectiveness of the TS-COS method and the application of topic similarity.

In Figure 3, we have a more clear view on the general performance of each method.[5] The performance of TS-EU changes with TS-COS. It is a little worse than TS-COS except in three cases. Although they are based on the same idea, their performance is different. The similarity model we choose can affect the overall results significantly. Consequently, it is possible to have better performance if we investigate more similarity models. On average, TS-EU also outperforms Rocchio's on all the collections.

Different from the other two methods, the TS-Entropy method is not so outstanding. But its results do verify our assumptions. It is better than the Rocchio model on four out of five collections, while its results are usually worse than TS-COS and TS-EU. With the increase of collection size, the performance gap between TS-Entropy and the other two methods becomes larger and larger. In a large collection, the ratio of irrelevant documents is larger as well. So the purity of documents will not be helpful to identify irrelevant documents. Nevertheless, the weights of pure relevant documents are still enhanced by the method, and mostly this factor makes improvements. Its performance highly depends on its base, the Rocchio model. On WT10G, the average performance of the Rocchio model is marginally better than BM25, and the MAP of BM25 is only 0.2055. In that case, the average performance of the TS-Entropy method is 1.70% worse than that of the Rocchio model. On the contrary, the Rocchio model obtains the best performance on GOV2 with MAP 0.3338, and the TS-Entropy method obtains the largest improvement on average (3.25%) over it. So the experimental results justify

[5]In order to show the performance clearly, we choose a value around the average performance of all the methods as the base.

Table IV. MAP Obtained by the Baselines, TS-COS, TS-EU, and TS-Entropy. A "*" and a "+" Symbol Indicate a Statistically Significant Improvement Over the RM3 and the Rocchio Baselines According to the Wilcoxon Matched-Pairs Signed-Ranks Test at the 0.05 Level. The Percentage in the Parentheses Is the Improvement Over Them. The Best Performance in Each Line Is in Bold

$\|D_f\|$	BM25	RM3	Rocchio	TS-COS	TS-EU	TS-Entropy
				disk1&2		
10	0.2380	0.2665	0.2962	**0.3019**$^{*+}$	0.3017^{*+}	0.3014^{*+}
				(13.28%, 1.92%)	(13.21%, 1.86%)	(13.10%, 1.73%)
20	0.2380	0.2652	**0.3095**	0.3073*	0.3056*	0.3048*
				(15.87%, −0.71%)	(15.23%, −1.26%)	(14.93%, −1.54%)
30	0.2380	0.2632	0.2950	**0.3075**$^{*+}$	0.3060^{*+}	0.3046^{*+}
				(16.83%, 4.24%)	(16.26%, 3.73%)	(15.73%, 3.15%)
50	0.2380	0.2610	0.2953	**0.3054**$^{*+}$	0.3035^{*+}	0.3022^{*+}
				(17.01%, 3.42%)	(16.28%, 2.78%)	(15.79%, 2.28%)
Average	0.2380	0.2640	0.2990	**0.3055**	0.3042	0.3033
				(15.74%, 2.18%)	(15.24%, 1.74%)	(14.88%, 1.40%)
				disk4&5		
10	0.2494	0.2720	0.2876	**0.3035**$^{*+}$	0.3020^{*+}	0.2979^{*+}
				(11.85%, 5.53%)	(11.03%, 5.01%)	(9.52%, 3.46%)
20	0.2494	0.2709	0.2894	**0.3028**$^{*+}$	0.2998^{*+}	0.2973^{*+}
				(11.78%, 4.63%)	(10.67%, 3.59%)	(9.75%, 2.66%)
30	0.2494	0.2695	0.2801	**0.2927**$^{*+}$	0.2898^{*+}	0.2868^{*+}
				(8.61%, 4.50%)	(7.53%, 3.46%)	(6.42%, 2.34%)
50	0.2494	0.2576	0.2688	**0.2824**$^{*+}$	0.2772^{*+}	0.2720^{*+}
				(9.63%, 5.06%)	(7.61%, 3.13%)	(5.59%, 1.18%)
Average	0.2494	0.2675	0.2815	**0.2954**	0.2922	0.2885
				(10.41%, 4.93%)	(9.23%, 3.81%)	(7.85%, 2.44%)
				WT2G		
10	0.3124	0.3244	0.3219	**0.3261**$^{+}$	0.3214	0.3146
				(0.52%, 1.30%)	(−0.92%, −0.16%)	(−3.02%, −2.32%)
20	0.3124	0.3255	0.3233	0.3338^{*+}	**0.3379**$^{*+}$	0.3283^{+}
				(2.55%, 3.25%)	(3.81%, 4.52%)	(0.86%, 1.52%)
30	0.3124	**0.3222**	0.2979	0.3198^{+}	0.3176^{+}	0.3076^{+}
				(−0.74%, 7.35%)	(−1.43%, 6.61%)	(−4.53%, 3.15%)
50	0.3124	0.3234	0.3092	**0.3131**	0.3104	0.3049
				(−3.18%, 1.26%)	(−4.02%, 0.39%)	(−5.72%, −1.41%)
Average	0.3124	**0.3239**	0.3131	0.3232	0.3218	0.3139
				(−0.21%, 3.23%)	(−0.63%, 2.79%)	(−3.10%, 0.25%)
				WT10G		
10	0.2055	0.2164	0.2045	0.2172^{+}	**0.2183**$^{+}$	0.2093^{+}
				(0.37%, 6.21%)	(0.88%, 6.75%)	(−3.88%, 2.29%)
20	0.2055	0.2151	**0.2193**	0.2171	0.2102	0.2077
				(0.93%, −1.00%)	(−2.28%, −4.15%)	(−3.44%, −5.58%)
30	0.2055	**0.2123**	0.1993	0.2078^{+}	0.2021	0.2007
				(−2.12%, 4.26%)	(−4.80%, 1.40%)	(−5.46%, 0.70%)
50	0.2055	**0.2098**	0.2010	0.2008	0.1991	0.1926
				(−4.29%, −0.10%)	(−5.10%, −0.95%)	(−8.20%, −4.36%)
Average	0.2055	**0.2134**	0.2060	0.2107	0.2074	0.2026
				(−1.25%, 2.28%)	(−2.80%, 0.68%)	(−5.07%, −1.70%)

(Continued)

Table IV. Continued

	GOV2					
10	0.3034	0.3172	0.3343	**0.3550**$^{*+}$ (11.92%, 6.19%)	0.3532^{*+} (11.35%, 5.65%)	0.3498^{*+} (10.28%, 4.43%)
20	0.3034	0.3167	0.3345	**0.3578**$^{*+}$ (12.98%, 6.97%)	0.3529^{*+} (11.43%, 5.50%)	0.3455^{*+} (9.09%, 3.18%)
30	0.3034	0.3160	0.3354	**0.3527**$^{*+}$ (11.61%, 5.16%)	0.3449^{*+} (9.15%, 2.83%)	0.3463^{*+} (9.59%, 3.15%)
50	0.3034	0.3138	0.3309	**0.3491**$^{*+}$ (11.25%, 5.50%)	0.3410^{*+} (8.67%, 3.05%)	0.3384^{*+} (7.84%, 2.22%)
Average	0.3034	0.3160	0.3338	**0.3537** (11.94%, 5.95%)	0.3480 (10.15%, 4.26%)	0.3450 (9.20%, 3.25%)

that those relevant documents that are "pure" in topics can help improve the overall performance of PRF. Generally, TS-Entropy is useful when Rocchio performs well, and this feature can be used in other text mining applications to evaluate the diversity of a document.

The performance of all three methods has similar trends in different cases while they are implemented under the same framework, TopPRF. In Figure 3, we can see that the bars of these three methods go up or bend down almost synchronously when the conditions change. For instance, if TS-COS obtains the best result when $|D_f|$ is 30, the other will also get their best performance in this situation. TS-COS is mostly the best one. TS-EU is a little worse than TS-COS, and TS-Entropy is always the worst among the three. On disk1&2, disk4&5, and GOV2, it is obvious that all the three methods are much better than RM3. Their base model, Rocchio, contributes some, while its performance is better than RM3 too. On WT10G, however, the bad performance of Rocchio also pulls down the three methods, while TS-COS can be better than RM3 sometimes. Although the proposed methods are generally better than their base, Rocchio, they are affected by the performance of Rocchio significantly because they actually use the same feedback documents. While we can obtain solid improvements by costing little in adjusting the weights of these documents, the experimental results are still encouraging.

6.4. Impact of Using LDA with the Whole Collection for PRF: A Case Study

One reason we choose only part of the feedback documents for topic modeling instead of the whole collection is the issue of time complexity. As we have mentioned in Section 4.2, when we only choose a fixed small set of documents for topic modeling and a fixed small number of topics, the time complexity of each iteration for LDA is O(1). However, the time complexity of using the whole collection will be linear with the number of topics and the number of documents $O(M * N)$, where M is the number of topics and N is the size of the collection. Here we have conducted a case study of using the whole collection on a relatively small dataset, Disk 1&2. We spent 15 hours building the topics offline on our server with Intel(R) Xeon(R) CPU E5410 @ 2.33GHz, 32G RAM. Our proposed approaches do not need this time for building topic space, since it costs constant time for building it online. Actually, the time spent in our experiments is comparable to Rocchio. According to our experiments on the GOV2 collection for 150 queries, Rocchio takes 2,056 seconds[6] and TS-COS takes 2,391 seconds with 20 topics. TS-COS takes about 10% more time than the Rocchio model. In general, the time complexity for our

[6]The experiments are conducted on our Intel(R) Core(TM) i7-2600 CPU, 8G RAM workstations. The time cost may change with different environments.

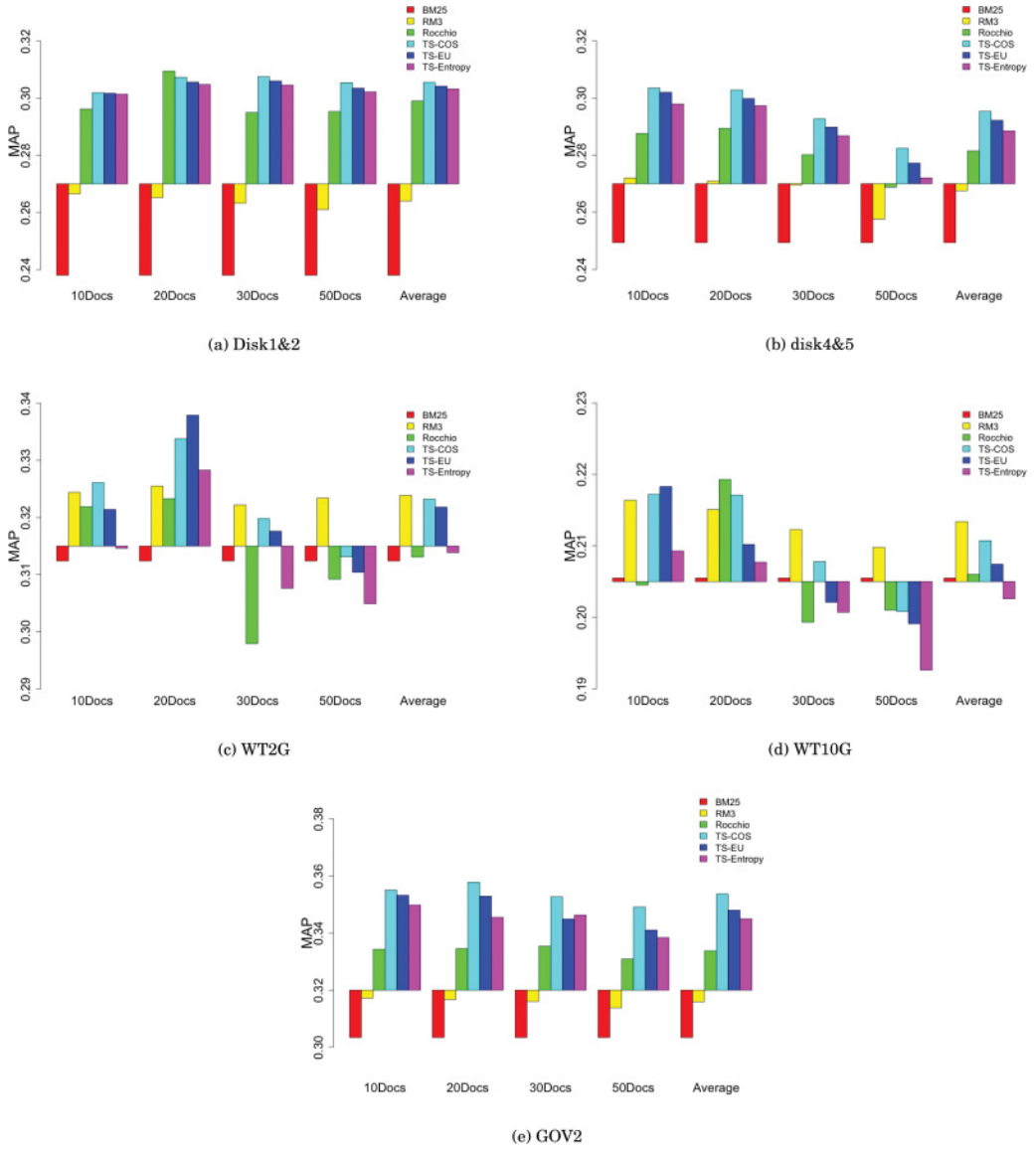


Fig. 3. Performance of the BM25, RM3, Rocchio, TS-COS, TS-EU, and TS-Entropy methods on the five TREC collections.

proposed methods is quite reasonable when only the top documents are used in building LDA.

Regarding the retrieval performance, we conduct experiments on using LDA with the whole collection for PRF and compare with our strategy of using top documents. Experiments are conducted on Disk1&2 with TREC queries 51 to 100. The MAP values of these two scenarios with different feedback document numbers and topic numbers are shown in Table V. We can observe that the results are very similar. The best result for each feedback document number is in bold, and using LDA on the top documents gave better performance than using LDA on the whole collection. The reason can be that using the whole collection will consider a broader area of topics, which may not be

Table V. MAP Comparison Between Building Topic Space from Topic Documents and Whole Collection: Experiments Are Conducted on Disk1&2 with TREC Queries 51–100, and All Topic Spaces Are Built via LDA

	LDA on the Top Documents			LDA on the Whole Collection	
$\lvert D_f \rvert$ \ topic	5	10	20	100	200
10	0.2747	0.2744	**0.2756**	0.2734	0.2691
20	0.2757	0.2767	**0.2769**	0.2762	0.2700
30	0.274	**0.2753**	0.275	0.2751	0.2680
50	0.2734	**0.2738**	0.2732	0.2718	0.2690

related to the given query. On the other hand, building LDA on the top documents will make the topic space less sparse and therefore the differences between relevant and irrelevant documents become more obvious. In other words, more focused topics will be generated, which are more likely to be relevant to the query. Similar trends can be observed on the other datasets.

6.5. Analyses

The improvements made by our methods come from the combination of pseudo relevance feedback and topic modeling. The term-based pseudo relevance feedback model, BM25-based Rocchio (Section 5.2), performs significantly better than the basic weighting model, BM25. For example, Rocchio has a 10.2% improvement (from 0.3034 to 0.3343 in Table IV) over BM25 on dataset GOV2. If we look at topic modeling approaches in the past, [Yi and Allan 2008] explored several different types of topic models for retrieval purposes, and their experimental results indicated that none of the topic model approaches can outperform RM on any dataset. Moreover, from Table IV, we can see that BM25-based Rocchio generally performs better than RM3. Therefore, none of the pure topic modeling approaches can significantly outperform Rocchio. Our proposed methods, incorporating topic space into feedback, can bring further contribution for boosting performance in most of the cases, compared to either pseudo-relevance-feedback-only approaches or topic-modeling-only approaches. For example, TS-COS significantly improves BM25-based Rocchio by 6.19% (from 0.3343 to 0.3550 in Table IV) on GOV2. Further, we have also studied the impact of using LDA with the whole collection for PRF. It is observed that using only the top documents is more efficient and effective than using the whole collection.

Besides the challenge of identifying topics, we consider that another latent problem is the loss of topical information. Using a few topics can neglect useful information in other ones, even when they are relevant to the query topic. For example, terms with higher probabilities in the selected topic(s) are considered to be more important for PRF. However, they can also appear frequently in other topics and they are actually not so informative. In that case, terms that have even probabilities in many topics should be less important. Without the information of full topics, we will miss these kinds of features. Choosing particular topics is a kind of dimension reduction. When topics are not stable, it is better to keep all the topic information. Therefore, we do not consider other dimension reduction methods like PCA or NMF in the scope of this article.

7. FURTHER EXPERIMENTS AND DISCUSSIONS

In this section, we will conduct an in-depth discussion and analysis based on our experimental results. At first, we will show two case studies on the same queries used in Section 3.1 and demonstrate how the ranks of feedback documents change with the topic number. As a result, our proposed model can achieve stable performance based on

the relative ranks of these documents. Next, we will compare TS-COS with the state-of-the-art Topk_LDA method. The experimental results indicate that the performance of our proposed method is much more stable and at least as good as the latter. Finally, we will show how the size of the trust group affects the performance.

7.1. Discussions of Two Case Studies

As we show in Section 3.1, the number of topics has an impact on the term distributions in the "most related" topic extensively. In this section, we will demonstrate how the topic number influences our proposed methods. Particularly, we make TS-COS the representative of our three proposed methods. Both TS-EU and TS-Entropy perform similarly to TS-COS.

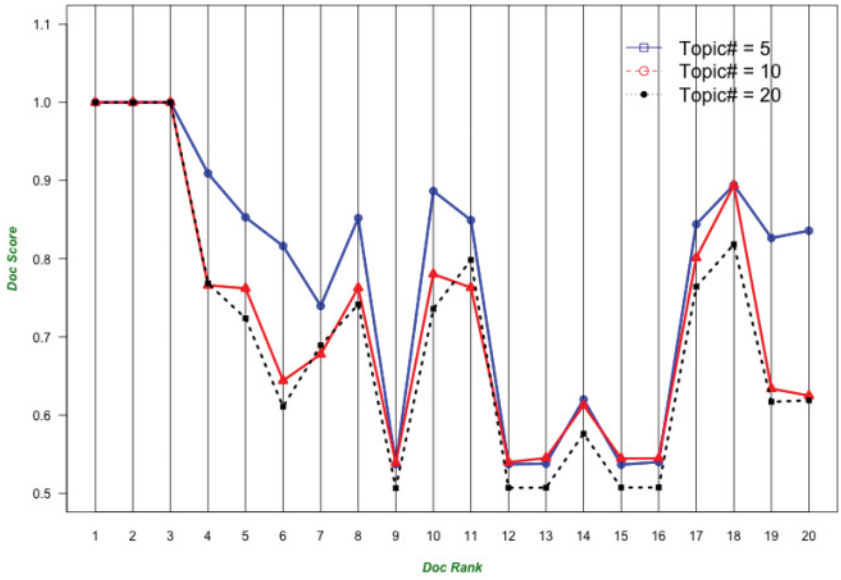
In order to make a fair comparison, we use the same parameter settings as in Section 3.1. Queries 802 and 804 are still used as examples. To demonstrate the results more clearly, we plot the scores of the top 20 documents graphically in Figure 4. The score for each document with different topic numbers can be found in Appendix A. As we can see from Figure 4, the score curves do not fall monotonously. Some documents obtain higher scores than those that are above them. That indicates that documents with lower ranks in the retrieved list can be similar to the trustable group on the topical level and assigned more weights.

When the topic number increases, the curves become more smooth. While there are more topics or we can say that topics are more fine grained, the topic distributions of two documents have more trivial differences. Consequently, the maximum similarity score of the feedback documents (except the trustable group) will be smaller when the topic number increases, and the range of the scores will be narrower while we normalize them to be above 0.5. In summary, the topic number can be used to adjust the differences among feedback documents.

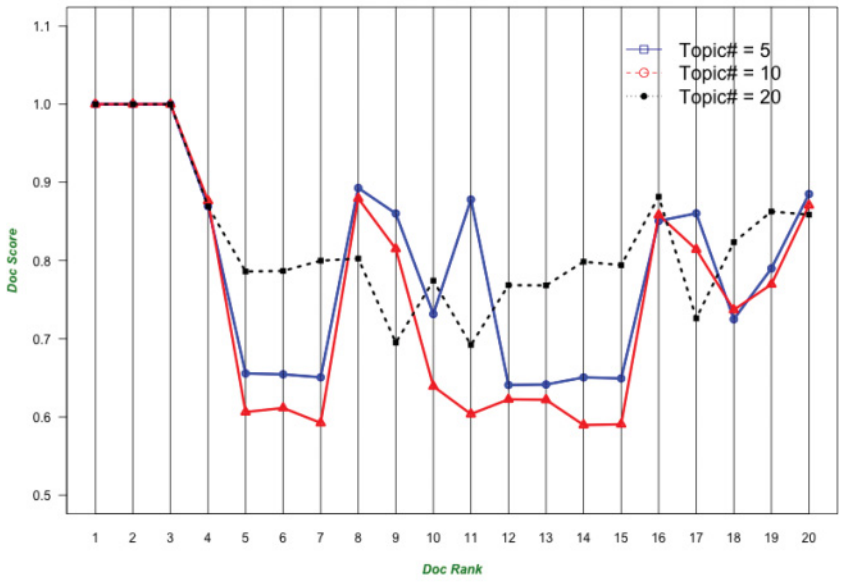
For query 802, the score rank of each document is almost unchanged for different topic numbers. But it is not that perfect for query 804. For instance, the score of document 10 is higher than document 11 when the topic number is five but is lower when the topic number is 10 or 30. With the changes in topic number, the score trends of the top k documents are generally stable, especially when compared with the term ranks in Table I.

The difference between the curves for queries 802 and 804 can be caused by the quality of the trustable group. The P@3 of BM25 for query 802 is 1.0, which means all the documents in the trustable group are relevant. Our proposed TS-COS method benefits from the high quality of the top three documents. Query 802, "Volcano eruptions global temperature," is also very clear. So the related information in the top three documents should be very close. If a document is similar to the first document, it will be similar to the other two. In other words, documents that are relevant can consistently gain high scores when the topic number changes because they will always get high similarity scores from each of the top three documents. The differences among documents are very clear. Compared with the performance of BM25 (AP 0.3241) and Rocchio (AP 0.3939), TS-COS obtains 0.4079, 0.4080, and 0.4084, respectively, for topic numbers of five, 10, and 30.

Meanwhile, the P@3 result of BM25 for query 804 is only 0.3333, and only the third document in the trustable group is relevant. Relevant documents have very similar information, which makes it easy to identify them from the irrelevant ones. On the contrary, irrelevant sample documents are not very helpful to identify other irrelevant documents, since their contents can be totally different. In this case, the relative ranks of feedback documents can be unstable when the topic number changes due to the unpredictable matching. When the topic number increases, most documents will not have overlaps on the topics of the irrelevant samples and obtain close scores. That can



(a) Query 802



(b) Query 804

Fig. 4. Cosine similarity scores of top 20 documents for queries 802 and 804.

Table VI. The Performance Change of Topk_LDA and TS-COS on Disk1&2 When Topic Number Is 5, 10 and 20. The Percentage in the Parentheses Is the Designated MAP over the MAP for Topic Number 5. "*" Indicates a Statistically Significant Improvement over Topk_LDA According to the Wilcoxon Matched-Pairs Signed-Ranks Test at the 0.05 Level

	Topk_LDA		
topic / $\lvert D_f \rvert$	5	10	20
10	0.2897	0.2871 (−0.90%)	0.2648 (−8.60%)
20	0.2967	0.2893 (−2.49%)	0.2813 (−5.20%)
30	0.2963	0.2897 (−2.22%)	0.2818 (−4.89%)
50	0.2941	0.2912 (−0.99%)	0.2876 (−2.21%)
	TS-COS		
topic / $\lvert D_f \rvert$	5	10	20
10	0.3056*	0.3055* (−0.03%)	0.3060* (−0.20%)
20	0.3108*	0.3112* (0.13%)	0.3108* (0.00%)
30	0.3107*	0.3106* (−0.03%)	0.3110* (−0.10%)
50	0.3087*	0.3087* (0.00%)	0.3085* (−0.64%)

explain why the curve becomes smoother when the topic number is 20. An interesting phenomenon is that the performance of TS-COS is not worse than the Rocchio model in this case. The average precision (AP) performance of Rocchio is 0.5632, and TS-COS gets 0.5762, 0.5761, and 0.5759, respectively, for topic numbers of five, 10, and 20. A possible reason is that when the trustable group is not good, the score differences of these documents are not huge. So terms are compared mainly on their term features in the feedback documents. That ensures the performance will not be much worse than the Rocchio model. In addition, only the relevant sample can be helpful in identifying feedback documents.

In summary, our proposed methods can provide very good and stable results when the samples are relevant. If most of the samples in the trustable groups are not relevant, our proposed method can prevent the performance from dropping too much by taking the term features into account. At the same time, the score range of these documents will be narrower due to the diversity of irrelevant information. In this case, the impact of topical information is reduced when evaluating the weight of feedback terms.

7.2. Comparisons with Topk_LDA

To support our argument that the performance is more robust by integrating topic space into PRF under our proposed probabilistic framework, we compare one of our methods, TS-COS, with the best method, Topk_LDA, in Ye et al. [2011] and see how the performance changes according to topic numbers on five standard TREC collections. Topk_LDA is a state-of-the-art approach in integrating topical information on PRF, which chooses a set of the top topics with weights higher than a given threshold and selects terms based on their probabilities given these topics.

In order to make fair comparisons, we set feedback term numbers as {10, 20, 30, 40, 50} as in Ye et al. [2011]. This is different from the setting in Section 6, where the feedback term number is fixed to be 30. The rest of the settings are the same as described in Section 5.3. The comparison results with Topk_LDA are shown in Tables VI, VII, VIII, IX, and X. Since more feedback term numbers are screened, these results are slightly better than those in Table IV, but the trends are the same. In addition, Ye et al. [2011] use a different query set from our experiments, and here we implement our approach, TS-COS, with the query set in Ye et al. [2011] in Table VII for a fair comparison.

Table VII. The Performance Change of Topk_LDA and TS-COS on Disk4&5 When Topic Number Is 5, 10, and 20. To Compare with Topk_LDA, We Only Use the Same Queries, 301–450. So the Performance Is Quite Different from What We Show in Table IV. The Percentage in the Parentheses Is the Designated MAP over the MAP for Topic Number 5. "*" Indicates a Statistically Significant Improvement over Topk_LDA According to the Wilcoxon Matched-Pairs Signed-Ranks Test at the 0.05 Level

	Topk_LDA		
topic / $\lvert D_f \rvert$	5	10	20
10	0.2581	0.2522 (−2.29%)	0.2396 (−7.17%)
20	0.2628	0.2579 (−1.86%)	0.2523 (−4.00%)
30	0.2631	0.2555 (−2.89%)	0.2490 (−5.36%)
50	0.2569	0.2527 (−1.63%)	0.2506 (−2.45%)
	TS-COS		
topic / $\lvert D_f \rvert$	5	10	20
10	0.2635	0.2622* (−0.49%)	0.2616* (−0.72%)
20	0.2625	0.2644 (0.72%)	0.2634* (0.34%)
30	0.2553	0.2540 (−0.51%)	0.2538* (−0.66%)
50	0.2498	0.2506 (0.32%)	0.2488 (−0.4%)

Table VIII. The Performance Change of Topk_LDA and TS-COS on WT2G When Topic Number Is 5, 10, and 20. The Percentage in the Parentheses Is the Designated MAP over the MAP for Topic Number 5. "*" Indicates a Statistically Significant Improvement over Topk_LDA According to the Wilcoxon Matched-Pairs Signed-Ranks Test at the 0.05 Level

	Topk_LDA		
topic / $\lvert D_f \rvert$	5	10	20
10	0.3171	0.3031 (−4.41%)	0.3091 (−2.52%)
20	0.3161	0.3082 (−2.50%)	0.3039 (−3.86%)
30	0.3170	0.3121 (−1.55%)	0.3130 (−1.26%)
50	0.3174	0.3147 (−0.85%)	0.3129 (−1.42%)
	TS-COS		
topic / $\lvert D_f \rvert$	5	10	20
10	0.3204	0.3202 (0.06%)	0.3208 (0.18%)
20	0.3384	0.3384* (0.00%)	0.3385* (0.03%)
30	0.3197	0.3207 (0.31%)	0.3192 (−0.16%)
50	0.3112	0.3112 (0.00%)	0.3117 (0.16%)

First of all, the performance of Topk_LDA shows a big difference when the topic number changes. For example, in Table V on Disk1&2, Topk_LDA loses about 8% performance when the topic number is changed from five to 20 for the feedback document number 10. In Table VIII, on WT10G, Topk_LDA gained about 4% when changing the topic number from five to 20 for feedback document number 30. However, our proposed method TS-COS is much more robust. The results are not sensitive with respect to the topic numbers. Figure 5 shows the performance of Topk_LDA and TS-COS on the five TREC collections. All results are averaged based on the number of feedback documents and converted to percentages based on the lowest value on each collection. We can observe that Topk_LDA and TS-COS behave very differently when the topic number changes from five to 20. Topk_LDA's performance is highly sensitive to the change of topic numbers. On the other hand, TS-COS usually has very similar performance with different topic numbers. These results show that the usage of topic space can reduce

Table IX. The Performance Change of Topk_LDA and TS-COS on WT10G When Topic Number Is 5, 10, and 20. The Percentage in the Parentheses Is the Designated MAP over the MAP for Topic Number 5

	Topk_LDA		
$\lvert D_f \rvert$ \ topic	5	10	20
10	0.2310	0.2283 (−1.17%)	0.2297 (−0.56%)
20	0.2290	0.2289 (−0.04%)	0.2333 (1.87%)
30	0.2220	0.2285 (2.93%)	0.2325 (4.73%)
50	0.2267	0.2273 (0.26%)	0.2312 (1.99%)
	TS-COS		
$\lvert D_f \rvert$ \ topic	5	10	20
10	0.2157	0.2224 (3.11%)	0.2212 (2.55%)
20	0.2204	0.2148 (−2.54%)	0.2152 (−2.41%)
30	0.2004	0.2062 (2.89%)	0.2059 (2.74)%
50	0.2020	0.1988 (−1.58%)	0.1967 (−2.62%)

Table X. The Performance Change of Topk_LDA and TS-COS on GOV2 When Topic Number Is 5, 10, and 20. The Percentage in the Parentheses Is the Designated MAP over the MAP for Topic Number 5. "*" Indicates a Statistically Significant Improvement over Topk_LDA According to the Wilcoxon Matched-Pairs Signed-Ranks Test at the 0.05 Level

	Topk_LDA		
$\lvert D_f \rvert$ \ topic	5	10	20
10	0.3445	0.3327 (−3.43%)	0.3352 (−2.69%)
20	0.3446	0.3333 (−3.28%)	0.3357 (−2.58%)
30	0.3443	0.3335 (−3.14%)	0.3322 (−3.51%)
50	0.3488	0.3473 (−0.43%)	0.3412 (−2.18%)
	TS-COS		
$\lvert D_f \rvert$ \ topic	5	10	20
10	0.3589*	0.3580* (−0.25%)	0.3576* (−0.36%)
20	0.3605*	0.3587* (−0.50%)	0.3588* (−0.49%)
30	0.3559	0.3546 *(−0.37%)	0.3541* (−0.51%)
50	0.3506	0.3496 (−0.29%)	0.3486 (−0.57%)

the sensitivity to the topic number when integrating full topic-document information for PRF. This is because the feedback terms' weights are adjusted based on the scores of corresponding feedback documents. Top-ranked documents are more likely to be relevant to the query, and the terms appearing in the top documents are more likely to be good feedback terms. Our proposed approaches keep the highly weighted documents and evaluate the reliability of other feedback documents that can potentially provide more relevant terms according to the proposed topic-based approaches. The feedback term list keeps the good feedback terms in the top documents and is expanded by more relevant terms based on the term weight and the reliability of the feedback document. Therefore, the feedback term list is stable no matter how the topics are changed with different topic numbers. On the other hand, approaches only relying on selected topics, such as Topk_LDA, change the feedback term set according to the topics. And the good feedback terms can be neglected if the topics are not well selected. That is the reason Topk_LDA could be significantly affected by the change of generated topics.

Also, the TS-COS method generally outperforms Topk_LDA significantly on four out of five collections, Disk12, Disk45, WT2G, and GOV2. Although we have used Rocchio

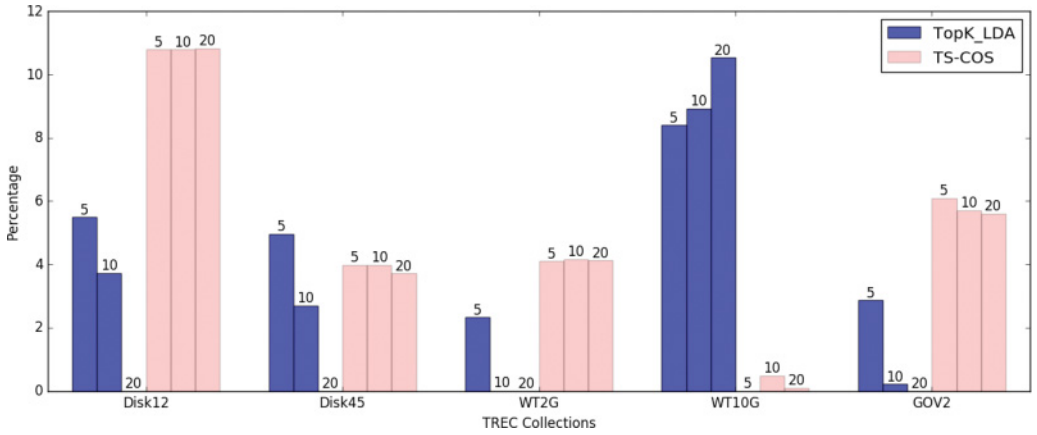


Fig. 5. Comparisons between Topk_LDA and TS-COS with different number of topics. Results are percentages based on the lowest value on each collection.

in our approaches, which has better performance than RM3[7] on these two collections, these experimental results justify that our idea can make solid improvements over strong baselines. On WT10G, Topk_LDA has better performance than TS-COS. The reason could be that the feedback documents from the first-pass retrieval on WT10G are more irrelevant to the query, since the MAP on WT10G is low. Thus, there is more noise in the generated topics. The Topk_LDA method has removed the noisy topics in the feedback process. TS-COS is able to capture all topical information and could be affected by the noisy topics. However, our proposed TS-COS has more advantages for further improving the basic models with high accuracies. When the basic model does not perform well, TS-COS can also improve the retrieval performance. But the improvement is not as high as the case that the basic model has good performance.

To summarize, the integration of topic space in PRF makes the performance of our proposed methods more robust than methods like the state-of-the-art Topk_LDA. When the fluctuation caused by the "fuzzy topics" is relieved, the application of topic space can enhance the performance of the classic Rocchio model. When using topics to represent documents and mine latent relations among documents from them, the overall performance can be more stable than methods selecting particular topics. Our proposed methods will not filter topical information even when it is identified as "irrelevant." Information hidden in all topics can be useful when measuring the similarity between two documents. If two documents have quite similar distributions over all topics, we will know they are similar in semantics and do not have to identify which topics are really relevant. It is better to identify how relevant a document is rather than a topic while the latter is not stable with different topic numbers. At the same time, significant improvements over the strong baseline BM25-based Rocchio model also shows that the integration of topical information can benefit the term-based PRF by importing information on a different grade. Thus, the integration of topic space brings both robustness and significant improvements over strong baseline models. We can conclude that topic space is a very beneficial complement for traditional term-based matching. In the future, it is promising to integrate topic space into other topic modeling applications.

[7]Actually, the performance of RM3 in Ye et al. [2011] is better than that in this article because Ye et al. tried more parameter values (e.g., term numbers). The performance of TS-COS in Tables VI, VII, VIII, IX, and X is better than that in Table IV for the same reason.

Table XI. Impact of P@n on the TS-COS Method: The Best Performance Under Each Condition Is in Bold

$\|D_f\|$	1	2	3	5	10
			disk1&2		
10	0.3014	0.3009	**0.3019**	0.3016	NA
20	0.3066	0.3069	**0.3073**	0.3062	0.3056
30	0.3071	**0.3076**	0.3075	0.3071	0.3070
50	0.3049	0.3051	**0.3054**	0.3046	0.3043
P@n	0.5733	0.5267	0.5244	0.5187	0.5053
			disk4&5		
10	0.3021	0.3029	**0.3035**	0.3007	NA
20	0.2997	0.3014	**0.3028**	0.3002	0.3017
30	**0.2935**	0.2929	0.2927	0.2926	0.2926
50	0.2822	0.2824	0.2824	0.2831	**0.2836**
P@n	0.5582	0.5321	0.5261	0.4980	0.4345
			WT2G		
10	0.3181	0.3127	**0.3261**	0.3183	NA
20	0.3371	0.3343	0.3338	**0.3377**	0.3327
30	0.3196	**0.3226**	0.3198	0.3157	0.3199
50	0.3123	**0.3136**	0.3131	0.3121	0.3133
P@n	0.5800	0.5800	0.5400	0.5040	0.4840
			WT10G		
10	0.2164	0.2157	**0.2172**	0.2143	NA
20	**0.2176**	0.2084	0.2171	0.2107	0.2076
30	**0.2088**	0.2085	0.2078	0.2044	0.2005
50	0.2008	**0.2015**	0.2008	0.2013	0.2003
P@n	0.4900	0.4400	0.4200	0.3840	0.3280
			GOV2		
10	**0.3556**	0.3527	0.3550	0.3531	NA
20	0.3580	**0.3594**	0.3578	0.3544	0.3493
30	0.3540	0.3537	0.3527	**0.3543**	0.3477
50	0.3503	0.3525	**0.3537**	0.3516	0.3461
P@n	0.6970	0.6465	0.6431	0.6182	0.5818

7.3. Trustable Group Size *s*

In this section, we will discuss how the size of the trustable group s affects the performance of the topic-similarity-based methods. Because TS-EU performs similarly to TS-COS, we only focus on TS-COS.

In order to investigate the impact, we also demonstrate the ratio of the relevant documents in the trustable group, which is actually P@n of the basic model BM25. We set the size of the group s to one, two, three, five and 10. In total, we have four different $|D_f|$ of 10, 20, 30, and 50 for all five collections. We consider the combination of a particular $|D_f|$ and a particular collection as a certain condition, and therefore we have 20 conditions to compare the performance of TS-COS with different sizes of trustable groups. All the results are shown in Table XI .

From the table, we can see that P@n decreases when n is larger. This evidence shows that the quality of the trustable group does fall down if we use more documents. P@10 of BM25 is less than 0.5 on three out of five collections, and only a little higher than that on the Disk1&2 dataset. Accordingly, only under one condition does TS-COS obtain the best result with a 10-document trustable group (on Disk4&5 when $|D_f|$ is 50). So it is better to choose a small s for similarity calculation.

Furthermore, although P@1 is usually much higher than other P@n results, TS-COS does not benefit much from it. The one-document group performs the best in four out of 20 conditions, but none of these four results are significantly better than that obtained when s is three. Generally, the performance of TS-COS using different s is close on Disk1&2, Disk4&5, and WT2G. On WT10G, we obtain significant improvements when s is one and three over others. This indicates that when the overall quality of the trustable group is not good (i.e., P@n is comparatively low), the performance of TS-COS is very sensitive to the values of s. When s is 10, the performance drops significantly compared to the best performance. Meanwhile, three is a good choice under eight conditions. This justifies the assumption that the relevant topics are not covered by the first feedback document in many cases.

Generally, it is better to choose a small s, especially when the results obtained through the first-pass retrieval are not good. To take the diversity of the relevant topics into account, one is not the best choice in most cases. According to the experimental results in this section, it is safe to set s to be three for different datasets.

8. CONCLUSIONS AND FUTURE WORK

In this article, we propose a probabilistic framework, TopPRF, and three new models based on the topic-document information without importing any new parameter. A new concept, "topic space," is introduced to evaluate the reliability of each candidate feedback document, and then the weights of terms are adjusted according to the reliability scores of the documents they belong to. Generally, extensive experiments show that topical information can make significant improvements over the classic Rocchio model with BM25 optimal parameter settings through our proposed methods, and also outperform the state-of-the-art RM3 model in many cases.

The contributions of the proposed framework are fivefold. First of all, we investigate the "fuzzy topic" obstacle and provide evidence to justify how it affects the application of topics significantly, especially for methods relying on particular topics. Because topic modeling and the usage of topics are becoming more and more popular, this problem is very important and cannot be ignored. To this end, we propose a new probabilistic framework, TopPRF, by introducing a new concept topic space. Using topic space coordinates to describe documents and comparing them with complete topic-document information can bring very stable results. To identify which documents are more reliable, we need to weight the feedback documents by integrating topical information. Using our methods, the relative ranks of feedback documents according to their scores (e.g., cosine similarity scores) are very robust. No matter how the topic distributions change, terms in those highly ranked documents will be consistently more important than others for query expansion. This is an important finding for integrating topical information for IR, especially when there is not an optimal topic number for corpora. By using all topical information in the feedback documents, our proposed approaches have more advantages for further improving strong basic models.

Second, based on the new framework, we find that topic similarity is effective for evaluating the reliability of each feedback document on its relevance. However, different similarity functions will lead to different performance. For instance, TS-COS performs better than TS-EU in most cases. Third, because TopPRF is derived from the Rocchio model, when the performance of the latter is not good, our proposed methods are affected. The average performance of TS-COS is better than the Rocchio model. Therefore, how to transfer the information of a document in the topic space into weights needs very careful consideration. Fourth, the TS-Entropy performs not as well as TS-COS or TS-EU. But it obtains better results than the baselines for most cases. So "purity" should be a useful feature when most feedback documents are really relevant. It can also be considered as a useful feature when measuring the quality of a document

for other applications. Finally, when the P@n performance of the basic model is good enough (e.g., above 0.5), the size of the trustable group will not affect it much. When most documents in the trustable group are actually irrelevant or the size of the collection is large, the performance of our proposed methods will drop significantly when the group contains more than 10 documents. By default, three is a good choice for different collections. This also verifies the assumption that more than one document is needed to cover the related topics for a query. In summary, the "fuzzy topic" problem deserves more concern and the usage of "topic space" will be a promising solution for further applications of topics.

In the future, we plan to research other similarity formulas that can affect the performance significantly. Since there are many choices, it is promising to have better results based on the topic similarity. Also, we can combine the topic similarity feature with TS-Entropy or other features to investigate how to integrate them effectively. Additionally, with the development of topic modeling, we can study more approaches for extracting topics to investigate how topics obtained through them influence our methods and why that happens. The Top-PRF framework can also be extended with ontology-based knowledge for semantic search. Finally, the conclusions in the article can be used for other text-processing-related areas to discover high-quality documents or measure document similarity. We can also apply our proposed probabilistic framework and approaches to passage-based retrieval, genomics/clinical IR, and medical search.

ACKNOWLEDGMENTS

We gratefully appreciate the anonymous reviewers for their valuable and detailed comments that greatly helped to improve the quality of the article. In particular, we thank the associate editor for reviewing and providing important feedbacks for this article. We also thank Dr. Zheng Ye for providing the source code of comparison baselines. Finally, we acknowledge Professor Stephen E. Robertson for his valuable feedback on this article.

REFERENCES

J. Allan, M. E. Connell, W. B. Croft, F. Feng, D. Fisher, and X. Li. 2000. INQUERY and TREC-9. In *Proceedings of the 9th Text REtrieval Conference*, 13.

D. Andrzejewski and D. Buttler. 2011. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM Conference on Knowledge Discovery and Data Mining*, 600–608. ACM, New York, NY.

M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. Williams. 1997. Okapi at TREC-5. In *Proceedings of the 5th Text REtrieval Conference*. NIST Special Publication SP, 143166.

J. Bian, Y. Yang, and T. Chua. 2013. Multimedia summarization for trending topics in microblogs. In *22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*. 1807–1812.

D. M. Blei, A. Y. Ng, and Jordan, M. I. 2003a. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

G. Blei and J. Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems* 16:17–25.

K. L. Caballero and R. Akella. 2012. Incorporating statistical topic information in relevance feedback. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1093–1094. ACM, New York, NY.

G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 243–250.

C. Carpineto, R. de Mori, G. Romano, and B. Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)* 19(1):1–27.

Y. Chen, H. Amiri, Z. Li, and T. Chua. 2013. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, 43–52.

C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 659–666. ACM.

K. Collins-Thompson. 2009. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. 837–846. ACM, New York, NY.

J. S. Culpepper, S. Mizzaro, M. Sanderson, and F. Scholer. 2014. Trec: Topic engineering exercise. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1147–1150. ACM, New York, NY.

R. Cummins, J. H. Paik, and Y. Lv. 2015. A pólya urn document language model for improved information retrieval. *ACM Transactions on Information Systems (TOIS)* 33(4):21:1–21:34.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 6(6):721–741.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

J. He 2011. *Exploring Topic Structure: Coherence, Diversity and Relatedness*. ISBN 9789490371814.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 50–57. ACM, New York, NY.

J. X. Huang, J. Miao, and B. He. 2013. High performance query expansion using adaptive co-training. *Information Processing & Management* 49(2):441–453.

X. Huang and Q. Hu. 2009. A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY.

X. Huang, Y. R. Huang, M. Wen, A. An, Y. Liu, and J. Poon. 2006. Applying data mining to pseudo-relevance feedback for high performance text retrieval. In *Proceedings of the 6th IEEE International Conference on Data Mining*. 295–306. IEEE.

X. Huang, M. Zhong, and L. Si. 2005. York University at TREC 2005: Genomics track. In *Proceedings of the 14th Text REtrieval Conference.*

F. Jian, J. X. Huang, and J. Zhao. 2016. A simple enhancement for ad-hoc information retrieval via topic modelling. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

A. Kotov, Y. Wang, and E. Agichtein. 2013. Leveraging geographical metadata to improve search over social media. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13 Companion)*, 151–152.

V. Lavrenko and W. B. Croft. 2001. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 120–127.

W. Li and A. McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*. 577–584, New York, NY, USA. ACM.

Y. Liu, Z. Liu, T. Chua, and M. Sun. 2015. Topical word embeddings. In *Proceedings of the 29th Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas,USA.*. 2418–2424.

Y. Lv and C. Zhai. 2009. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the International Conference on Information and Knowledge Management*. 1895–1898. ACM.

Y. Lv and C. Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 579–586. ACM.

Y. Lv, C. Zhai, and W. Chen. 2011. A boosting approach to improving pseudo-relevance feedback. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–174. ACM.

Q. Mei, X. Shen, and C. Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*, KDD'07, 490–499, New York, NY, USA. ACM.

J. Miao, J. X. Huang, and Z. Ye. 2012. Proximity-based Rocchio's model for pseudo relevance. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 535–544, New York, NY, USA. ACM.

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma 2006. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*. 18–25.

I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. 2008. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*. 569–577. ACM.

M. Porter. 1980. An algorithm for suffix stripping. *Program*, 14:130–137.

K. Raman, R. Udupa, P. Bhattacharyya, and A. Bhole. 2010. On improving pseudo-relevance feedback using pseudo-irrelevant documents. In *Proceedings of 32nd European Conference on Information Retrieval*. 573–576, 2010.

S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval.* 3(4): 333-389. Now Publishers Inc. Hanover, MA, USA.

S. E. Robertson, S. Walker, S. Jones, Hancock-M. Beaulieu, and Gatford, M. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*.

S. E. Robertson and S. Walker 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA, 232–241.

J. Rocchio. 1971. *Relevance feedback in information retrieval*, 313–323. Prentice-Hall Englewood Cliffs.

G. Salton, A. Wong, and C. Yang. 1975a. A vector space model for information retrieval. *Journal of American Society for Information Retrieval*, 18(11):613–620.

G. Salton, A. Wong, and C. S. Yang. 1975b. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

M. Serizawa and I. Kobayashi. 2013. A study on query expansion based on topic distributions of retrieved documents. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, 369–379. Springer Berlin Heidelberg.

H. Stark, Y. Yang, and Y. Yang. 1998. *Vector space projections: A numerical approach to signal and image processing, neural nets, and optics*. John Wiley & Sons, Inc. ISBN:0471241407.

T. Strohman, D. Metzler, H. Turtle, and W. B. Croft 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*. Vol. 2. 2–6.

J. Tang, R. Jin, and J. Zhang. 2008. A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of the 8th IEEE International Conference on Data Mining*, 1055–1060. IEEE.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2012. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006. 101[476]:1566–1581.

E. M. Voorhees and D. Harman. 2000. Overview of the sixth text retrieval conference. *Information Processing and Management: an International Journal*, 36:3–35.

B. Walsh. 2004. Markov chain Monte Carlo and Gibbs sampling. *Lecture Notes for EEB 581*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.131.4064.

C. Wang and D. M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*, 448–456, New York, NY, USA. ACM.

X. Wang, Q. Zhang, X. Wang, and Y. Sun. 2012. LDA based pseudo relevance feedback for cross language information retrieval. In *Cloud Computing and Intelligent Systems (CCIS)*, volume 03, 1511–1516.

X. Wei and W. B. Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178–185. ACM.

R. W. White and G. Marchionini. 2007. Examining the effectiveness of real-time query expansion. *Information Processing and Management*, 43(3):685–704, 2007.

J. Xu and W. B. Croft. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

Z. Ye, B. He, X. Huang, and H. Lin. 2010. Revisiting Rocchio's relevance feedback algorithm for probabilistic models. In *Information Retrieval Technology*, volume 6458, 151–161. Springer Berlin Heidelberg.

Z. Ye and J. X. Huang. 2014. A simple term frequency transformation model for effective pseudo relevance feedback. *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 323–332.

Z. Ye and J. X. Huang. 2016. A learning to rank approach for quality-aware pseudo-relevance feedback. *Journal of the Association for Information Science and Technology* 67(4): 942–959.

Z. Ye, J. X. Huang, and H. Lin. 2011. Finding a good query-related topic for boosting pseudo-relevance feedback. *Journal of the American Society for Information Science and Technology* 62(4):748–760.

Z. Ye, J. X. Huang, and J. Miao. 2012. A hybrid model for ad-hoc information retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1025–1026, New York, NY, USA. ACM.

X. Yi and J. Allan. 2008. Evaluating topic models for information retrieval. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*. 1431–1432. ACM.

X. Yi and J. Allan. 2009. A comparative study of utilizing topic models for information retrieval. In *Proceedings of the 31st European Conference on Information Retrieval*. 29–41, Berlin, Heidelberg. Springer-Verlag.

X. Yin, J. Huang, Z. Li, and X. Zhou. 2013. A survival modeling approach to biomedical search result diversification using wikipedia. *IEEE Trans. Knowl. Data Eng.* 25, 6, 12011212.

C. Zhai. 2008. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2:137–213.

C. Zhai and J. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 334–342, New Orleans, LA.

C. Zhai and J. Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *Foundations and Trends in Information Retrieval* 22(2):179–214.

J. Zhao, J. X. Huang, and B. He. 2011. CRTER: Using cross terms to enhance probabilistic information retrieval. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 155–164, New York, USA. ACM.

J. Zhao, J. X. Huang, and Z. Ye. 2014. Modeling term associations for probabilistic information retrieval. *ACM Transactions on Information Systems (TOIS)*, 32(2), 7, 47.

N. Zhiltsov and E. Agichtein. 2013. Improving entity search over linked data by modeling latent semantics. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. 1253–1256. ACM.

Received November 2015; revised April 2016; accepted June 2016