A Simple Enhancement for Ad-hoc Information Retrieval via Topic Modelling

Fanghong Jian[1], Jimmy Xiangji Huang, Jiashu Zhao[2], Tingting He[3] and Po Hu[3]
Information Retrieval and Knowledge Management Research Lab
[1]National Engineering Research Center for E-Learning, [3]School of Computer Science, Central China
Normal University, Wuhan, China; [2]School of Information Technology, York University, Toronto, Canada
jhuang@yorku.ca, jfhrecoba@mails.ccnu.edu.cn, jessie@cse.yorku.ca

ABSTRACT

Traditional information retrieval (IR) models, in which a document is normally represented as a bag of words and their frequencies, capture the term-level and document-level information. Topic models, on the other hand, discover semantic topic-based information among words. In this paper, we consider term-based information and semantic information as two features of query terms and propose a simple enhancement for ad-hoc IR via topic modeling. In particular, three topic-based hybrid models, LDA-BM25, LDA-MATF and LDA-LM, are proposed. A series of experiments on eight standard datasets show that our proposed models can always outperform significantly the corresponding strong baselines over all datasets in terms of MAP and most of datasets in terms of P@5 and P@20. A direct comparison on eight standard datasets also indicates our proposed models are at least comparable to the state-of-the-art approaches.

Keywords

Probabilistic Model; Dirichlet Language Model; LDA

1. INTRODUCTION

Many traditional IR models are based on the assumption that query terms are independent of each other, where a document is represented as a bag of words. Nevertheless this assumption may not hold in practice. Each document may contain several different topics and terms appeared in the document might belong to different topics, which represent different semantic information. Many researchers have been working on term topic information in IR [1, 10, 15, 16]. However, the nature of the associations among query terms still awaits further study. Some cluster-based approaches consider each document has only one topic [10], which is not reasonable to model large collection of documents. Topic-based document representation is effective in the language modeling (LM) framework [1, 15, 16]. But there is no generality in BM25 [2, 6, 20] and MATF (Multi Aspect TF) [13] based frameworks.

In this paper, we present three hybrid models for enhancing traditional IR model via topic modelling. In our proposed approach, term-based information and semantic information are considered as two features of query terms. Latent Dirichlet Allocation (LDA) [3] is utilized to combine these two features and enhance three well-known traditional IR models BM25 [2], MATF [13] and Dirichlet LM [18]. In particular, three hybrid models, denoted as LDA-BM25, LDA-MATF and LDA-LM, are proposed respectively. The main contributions of this paper are as follows. First we propose three simple but effective IR models by combining traditional IR models with topic model. Second we conduct extensive experiments to confirm the effectiveness of our proposed models.

The remainder of this paper is organized as follows. We describe the related work and propose three topic-based hybrid models for ad-hoc IR in Section 2 and 3 respectively. In Section 4, we set up our experimental environment on eight TREC collections. In Section 5, the experimental results are presented and discussed. Finally, we conclude our work briefly and present future research directions in Section 6.

2. RELATED WORK

Since the 1990s, researchers started to investigate how to integrate term association into IR models [8, 12, 16, 19, 20]. The query-term associations have been modeled by different approaches according to the distance of the query terms in documents. For example, Buttcher et al. (2006) [4] used a proximity accumulator to associate each query term. Lv and Zhai (2009) [11] proposed a positional language model (PLM) that incorporated the term proximity in a model-based approach using term-propagation functions. Metzler et al. (2005) [12] proposed a Markov Random Fields (MRF) model which modeled the joint distribution over queries and documents. Song et al. (2011) [14] proposed Proximity Probabilistic Model (PPM) which used a position-dependent term count to represent both the number of occurrences of a term and the term counts propagated from other terms. Recently, topic models have been widely used to explore latent term association in knowledge discovery and other related area. Liu and Croft (2004) [10] proposed cluster-based retrieval models under the language modeling framework, which were used to smooth the probabilities in the document model. In their approach, a document is supposed to contain only one topic, which is not reasonable to model large collection of documents. Azzopardi et al. (2004) [1] showed that it was effective to use the LDA model [3] to smooth the probabilities in the document model on several small collections. Wei and Croft (2006) [15] also discussed the applications of LDA in large collections, and presented a detailed evalua-

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07...$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914750

tion of the effectiveness. Yi and Allan (2009) [17] explored the utility of Mixture of Unigrams (MU) model, LDA and Pachinko Allocation Model (PAM) [9] for IR. They showed that topic models were effective for document smoothing. More rigorous topic models like LDA provided gains over cluster-based models and more elaborate topic models that capture topic dependencies provided no additional gains. Although it is effective to integrate topic models into the language modeling framework, how to integrate topical information into other traditional IR models is not clear.

3. OUR APPROACH

For enhancing performance, topic model is integrated into traditional retrieval models. First, the latent semantic information of query terms in a document is extracted via topic modeling. Then, the term-based information is obtained through traditional retrieval models. The documents that are more related to the query according to both semantic topic-based information and term-based information are boosted in the ranking process. For clarification, Table 1 outlines the notations used throughout the paper.

Table 1: Notations

Notations	Description
c	collection
d	document
q	query
q_i	query term
dl	length of document
$avdl$	average document length
N	number of indexed documents in collection
n	number of indexed documents containing a term
tf	within-document term frequency
qtf	within-query term frequency
z	topic
K_t	number of topics
p, p_{ml}	probability function
w', w'', w	weighting function
b, k_1, k_3	parameter in BM25
μ	Dirichlet prior in Dirichlet LM
α, β	hyperparameter in LDA

3.1 Topic-based Hybrid Model

Traditional retrieval models only capture term-based information. On the other hand, topic models acquire semantic information between words. In this paper, we propose enhanced retrieval models that consider not only term frequency, document frequency and document length, but also term topics information. We treat term-based information and semantic topic-based information as two features for query terms. The enhanced retrieval models combine these two features.

Given a query q, for each term q_i in query q, $w(q_i, d)$ is the enhanced weight for document d. In order to capture the two kinds of information, we use a parameter λ to balance their importance. So the weight of a query term for a document is as follows.

$$w(q_i, d) = (1 - \lambda) \cdot w''(q_i, d) + \lambda \cdot w'(q_i, d) \tag{1}$$

where $w''(q_i, d)$ represents the explicit term-based related information in traditional retrieval model for document d, $w'(q_i, d)$ is the implicit semantic information in topic model. Finally, a document's weight for a query is given by the sum of its weight for each term in the query. When λ equals to 0, the hybrid models become traditional IR models such as BM25 and LM. When λ equals to 1, the hybrid models become topic models. Because traditional IR models and topic models are normalized independently, the value of λ changes with different combinations. It is well known that BM25, MATF and Dirichlet LM are the state-of-the-art traditional IR models and LDA is a simple but effective topic model. Therefore, we use BM25, MATF and Dirichlet LM as the traditional models and we use LDA as the topic model.

3.2 Topic Model

In general, topic model is used to capture latent semantic information of terms in document. There are a lot of topic models, such as probabilistic Latent Semantic Indexing (pLSI) [7], LDA [3] and PAM [9]. LDA is a simple and effective topic model, and is broadly used. In this paper, we use LDA as our topic model.

LDA model can generate the probability of topics in a document and the probability of words in a topic, which can obtain the generated probability of words in a document. We take the probability of a query term in a document as its implicit semantic information in the document. The probability is larger, the term is more related with the document. In order to be the same magnitude with weights in traditional models, the weight of a query term for a document in LDA uses log value of the generated probability as follows.

$$w'(q_i, d) = \log p(q_i|d) = \log \left(\sum_{z=1}^{K_t} p(q_i|z)p(z|d) \right) \tag{2}$$

The LDA model can not be solved by exact inference and use Gibbs Sampling for parameter estimation like in [5].

3.3 Traditional Information Retrieval Models

Traditional information retrieval models are mainly classified into classic probabilistic model, vector space model and statistical language model. There are several well-known strong baselines in each class, considering BM25, MATF and Dirichlet LM respectively.

In BM25, the weight of a query term is related to its within-document term frequency and query term frequency. The corresponding weighting function is as follows.

$$w''(q_i, d) = \frac{(k_1 + 1) * tf}{K + tf} * \log \frac{(N - n + 0.5)}{(n - 0.5)} * \frac{(k_3 + 1) * qtf}{k_3 + qtf} \tag{3}$$

where w'' is the weight of a query term, the k_is are tuning constants and K equals to $k_1 * ((1 - b) + b * dl/avdl)$.

In 2013, Jiaul H. Paik [13] proposed a novel TF-IDF term weighting scheme MATF that employed two different within document term frequency normalizations to capture two different aspects of term saliency. One component of the term frequency is effective for short queries, while the other performs better on long queries. The final weight is measured by taking a weighted combination of these components, which is determined on the basis of the length of the corresponding query. Experiments carried out on a set of news and web datasets show that MATF outperforms several well-known state-of-the-art TF-IDF baselines with significantly large margin.

Dirichlet LM presented by Zhai and Lafferty in 2001 [18] used the likelihood probability of query terms in a document to rank relevance between query and document. In order to better computing, the weight of a query term uses the log value of the probability as follows.

$$w''(q_i, d) = \log p(q_i|d) = \log \left(\frac{dl}{dl + \mu} p_{ml}(q_i|d) + \frac{\mu}{dl + \mu} p_{ml}(q_i|c) \right) \tag{4}$$

4. EXPERIMENTAL SETTING

We conduct experiments on eight standard collections, which include AP88-89 with queries 51-100, AP88-90 with queries 51-150, FBIS with queries 351-450, FT(91-94) with queries 301-400, LA with queries 301-400, SJMN(1991) with queries 51-150, WSJ(87-92) with queries 151-200 and WT2G with queries 401-450. These datasets are different in size and genre [15, 19]. Queries without judgments are removed.

	Eval Metric	AP88-89	AP88-90	FBIS	FT	LA	SJMN	WSJ	WT2G
BM25	MAP	0.2710	0.2198	0.2606	0.2600	0.2490	0.1965	0.3156	0.3156
	P@5	0.4360	0.4566	**0.3735**	0.3726	0.3571	0.3404	0.5240	0.5280
	P@20	0.3860	0.3894	0.2685	0.2389	0.2194	0.2564	0.4410	0.3930
LDA-BM25	MAP	0.3021*	**0.2617***	**0.2661**	0.2769*	0.2592*	0.2297*	**0.3471***	0.3230
		(+11.476%)	(+19.064%)	(+2.111%)	(+6.500%)	(+4.96%)	(+16.902%)	(+9.981%)	(+2.345%)
	P@5	0.5020	**0.5232***	0.3679	0.3621	0.3673	0.3809	**0.5520**	**0.5360**
		(+15.138%)	(+14.595%)	(-1.499%)	(-2.818%)	(+2.856%)	(+11.889%)	(+5.344%)	(+1.515%)
	P@20	**0.4388***	**0.4505***	0.2691	0.2416	0.2291*	0.2915	**0.4640***	0.4030
		(+13.679%)	(+15.693%)	(+0.223%)	(+1.130%)	(+5.105%)	(+13.697%)	(+5.215%)	(+2.545%)
MATF	MAP	0.2771	0.2238	0.2553	0.2660	0.2502	0.2095	0.3029	0.3340
	P@5	0.4531	0.4707	0.3605	**0.3789**	0.3571	0.3723	0.5240	0.5240
	P@20	0.3980	0.4086	0.2673	0.2426	0.2240	0.2809	0.3950	0.4110
LDA-MATF	MAP	**0.3041***	**0.2617***	0.2634*	**0.2781***	0.2586*	**0.2309***	0.3343*	**0.3393**
		(+9.744%)	(+16.935%)	(+3.173%)	(+4.549%)	(+3.357%)	(+10.215%)	(+10.366%)	(+1.587%)
	P@5	0.4898	0.5131	0.3580	0.3621	**0.3694**	**0.3915**	0.5200	**0.5360**
		(+8.100%)	(+9.008%)	(-0.693%)	(-4.434%)	(+3.444%)	(+5.157%)	(-0.763%)	(+2.290%)
	P@20	0.4378*	0.4465*	**0.2784**	**0.2453**	**0.2337***	**0.2989**	0.4300*	**0.4150**
		(+10.000%)	(+9.276%)	(+4.153%)	(+1.113%)	(+4.330%)	(+6.408%)	(+8.861%)	(+0.973%)
LM	MAP	0.2672	0.2157	0.2525	0.2571	0.2427	0.2009	0.3047	0.3118
	P@5	0.4571	0.4465	0.3506	0.3684	0.3429	0.3532	0.5120	0.5000
	P@20	0.4041	0.4146	0.2500	0.2311	0.2235	0.2697	0.3910	0.3920
LDA-LM	MAP	0.2980*	0.2560*	0.2628*	0.2774*	**0.2603***	0.2254*	0.3344*	0.3165*
		(+11.527%)	(+18.683%)	(+4.079%)	(+7.896%)	(+7.252%)	(+12.195%)	(+9.747%)	(+1.507%)
	P@5	**0.5102***	0.5010*	0.3630	0.3600	**0.3694***	0.3830*	0.5200	0.5080
		(+11.617%)	(+12.206%)	(+3.537%)	(-2.280%)	(+7.728%)	(+8.437%)	(+1.563%)	(+1.600%)
	P@20	0.4276*	0.4414*	0.2599*	0.2426*	0.2286	0.2904*	0.4320*	0.3950
		(+5.815%)	(+6.464%)	(+3.960%)	(+4.976%)	(+2.282%)	(+7.675%)	(+10.486%)	(+0.765%)

Table 2: Comparison with baselines. The best result obtained on each dataset is in bold. "*" denotes statistically significant improvements over corresponding baselines (Wilcoxon signed-rank test with $p < 0.05$). The percentages below are the percentage improvement of proposed models over corresponding baselines.

For all test collections used, each term is stemmed by using Porter's English stemmer. Standard English stopwords are removed. The official TREC evaluation measure is used in our experiments, namely Mean Average Precision (MAP). To investigate top retrieved documents, P@5 and P@20 are also used for evaluation. All statistical tests are based on Wilcoxon Matched-pairs Signed-rank test.

For fair comparisons, we use the following parameter settings for both the baselines and our proposed models, which are popular in the IR domain for building strong baselines. First, in BM25, setting k_1, k_3 and b to 1.2, 8 and 0.35 respectively gave the best MAP for most datasets in [20]. Second, in Dirichlet LM, $\mu = 1000$ was shown in [15] to achieve best MAP for most datasets. Finally, in LDA model, we use symmetric Dirichlet priors with $\alpha = 50/K_t$ and $\beta = 0.01$, which are common settings in the literature and shown in [15] that retrieval results were not very sensitive to the values of these parameters. The number of topics K_t is set to be 400 as recommended in [15].

5. EXPERIMENTAL RESULTS

5.1 Comparison with Baselines

We first investigate the performance of our proposed topic-based models compared with the corresponding strong baselines BM25, MATF and Dirichlet LM. The experimental results are presented in Table 2. As shown by the results, our proposed models outperform the corresponding baselines on almost all datasets in terms of MAP, P@5 and P@20. Statistically significant improvement can be observed on most of datasets in terms of MAP and P@20. According to the results in Table 2, each hybrid model has its advantage on some aspects. However, there is no single hybird model that can achieve the best performance on all the datasets.

5.2 Parameter Sensitivity

An important issue that may affect the robustness of our proposed models is the sensitivity of their parameter λ to retrieval performance. Since the weights of query terms in traditional retrieval models and topic model are normalized independently, the value of λ reflects the influence of using topic-based model. Figure 1 plots the evaluation metrics MAP obtained by the proposed hybrid models over λ values ranging from 0 to 1 on all the datasets. It is clear that hybrid models perform better than either traditional models or topic model on all data sets. As we can see from Figure 1, our proposed models LDA-BM25, LDA-MATF and LDA-LM generally perform well over different datasets when λ has a smaller value.

We also study the performance of our proposed topic-based models with different number of topics compared with the corresponding baselines in terms of MAP. In Figure 2, the traditional models are shown as straight lines since the performance does not change over the number of topics. All the results are presented in Figure 2, which shows that our proposed models with different number of topics outperform corresponding baselines in terms of MAP over all datasets. Figure 2 shows that the proposed hybrid models tend to perform better when the number of topics increases. When the number of topics reaches a certain value, the retrieval performance tends to become more stable. The performance tendency of our proposed models with different number of topics is surprisingly consistent on all the datasets. Similar trends for λ and with different number of topics can also be observed in terms of P@5 and P@20.

5.3 Comparison with CRTER2 and LBDM

In addition, we compare our proposed models with two state-of-the-art approaches. Zhao etc. [19, 20] showed that bigram cross term model (CRTER2) is at least comparable to major probabilistic proximity models PPM [14] and BM25TP [4] in BM25-based framework. Xing and Allan [17], which is most close to our proposed model LDA-LM, showed their LDA-based model (LBDM) [15] achieved the best performance in topic-based LM framework. So we make a direct comparison with CRTER2 and LBDM. The results in terms of MAP are presented in Table 3. "↑" denotes LDA-BM25 outperforms CRTER2, while "⇑" denotes LDA-LM outperforms LBDM. Among eight datasets, LDA-BM25 wins five times and LDA-LM wins four times. By comparison, we can conclude that our proposed models LDA-BM25 and LDA-LM are at least comparable to the state-of-the-art models CRTER2 and LBDM.

Table 3: Comparison with CRTER2 and LBDM

	CRTER2	LDA-BM25	LBDM	LDA-LM
AP88-89	0.2789	0.3021↑	0.3051	0.2980
AP88-90	0.2268	0.2617↑	0.2535	0.2560⇑
FBIS	0.2738	0.2661	0.2636	0.2628
FT	0.2717	0.2769↑	0.2750	0.2774⇑
LA	0.2604	0.2592	0.2630	0.2603
SJMN	0.2095	0.2297↑	0.2234	0.2254⇑
WSJ	0.3406	0.3471↑	0.3359	0.3344
WT2G	0.3359	0.3230	0.3108	0.3165⇑

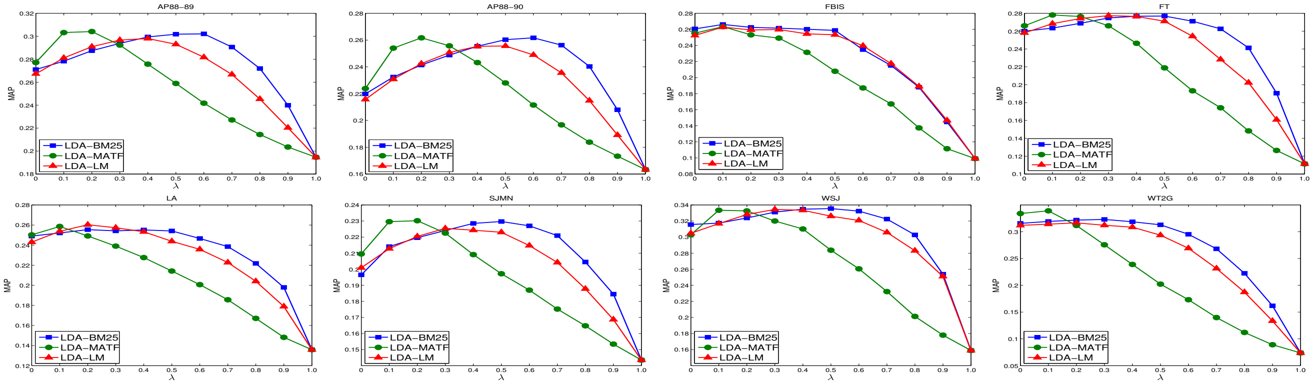

Figure 1: Parameter sensitivity of λ on all data sets

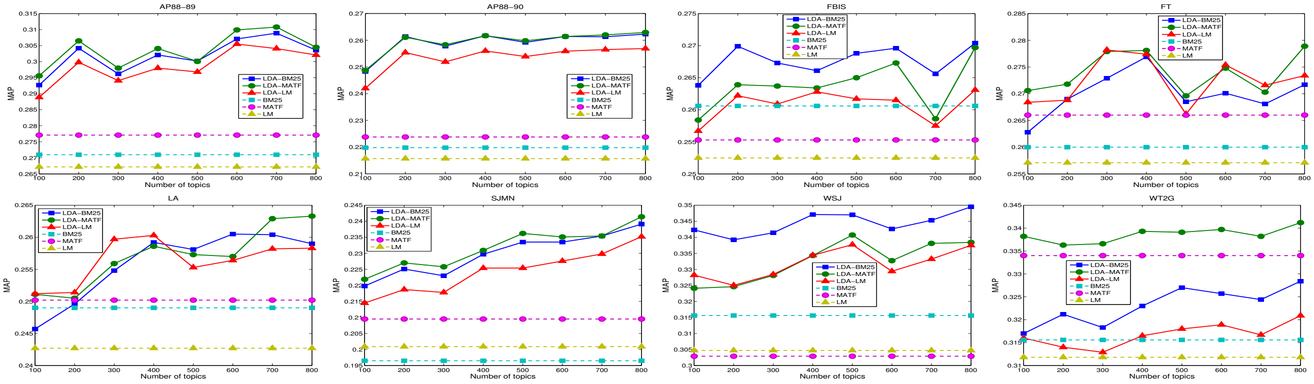

Figure 2: Parameter sensitivity of the number of topics on all data sets

6. CONCLUSIONS AND FUTURE WORK

In this paper, a simple enhancement for ad-hoc IR is proposed by combining traditional retrieval model and topic model. Specifically, we present three hybrid models LDA-BM25, LDA-MATF and LDA-LM for enhancing traditional IR models via topic modeling. These three models capture both term-based information and latent semantic topic-based information at the same time. Experimental results on eight standard datasets show that the proposed models are effective, and outperform the corresponding strong baselines on most of datasets in terms of MAP, P@5 and P@20. Meanwhile, our proposed models are at least comparable to the state-of-the-art `CRTER2` and topic-based model `LBDM`. Additionally, we carefully analyze the influence of λ to our proposed models and the performance of our proposed models with different number of topics.

There are several interesting future research directions to further explore. We would like to study the optimal topic number on each dataset. It is also interesting to conduct an in-depth study on the combination traditional IR model with topic model and find the best combination. We also plan to evaluate our models on more datasets including some real datasets and apply our models into real world applications.

7. ACKNOWLEDGMENTS

This research is supported by a Discovery grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada, an NSERC CREATE award and also supported by the National Natural Science Foundation of China. We thank anonymous reviewers for their thorough comments.

8. REFERENCES

[1] L. Azzopardi, M. Girolami, and C. J. Van Rijsbergen. Topic Based Language Models for ad hoc Information Retrieval. In *Proceedings of the International Joint Conference on Neural Networks*, pages 3281–3286, 2004.

[2] M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. Williams. Okapi at TREC-5. In *Proc. of TREC*, pages 143–166, 1996.

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

[4] S. Buttcher, C. L. A. Clarke, and B. Lushman. Term Proximity Scoring for Ad-hoc Retrieval on Very Large Text Collections. In *Proceedings of the 29th ACM SIGIR*, pages 621 – 622, 2006.

[5] T. L. Griffiths and M. Steyvers. Finding Scientific Topics. In *Proceeding of the National Academy of Sciences*, pages 5228–5235, 2004.

[6] B. He, J. X. Huang, and X. Zhou. Modeling term proximity for probabilistic information retrieval models. *Information Sciences*, 181(14):3017–3031, 2011.

[7] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd ACM SIGIR*, pages 50–57, 1999.

[8] Q. Hu, J. X. Huang, and X. Hu. Modeling and Mining Term Association for Improving Biomedical Information Retrieval Performance. *BMC Bioinformatics*, 13(2):18 pages, 2012.

[9] W. Li and A. McCallum. Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In *Proc. of ICML*, pages 577–584, 2006.

[10] X. Liu and W. B. Croft. Cluster-Based Retrieval Using Language Models. In *Proceedings of the 27th ACM SIGIR*, pages 186–193, 2004.

[11] Y. Lv and C. Zhai. Positional Language Models for Information Retrieval. In *Proceedings of the 32nd ACM SIGIR*, pages 299–306, 2009.

[12] D. Metzler and W. B. Croft. A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th ACM SIGIR*, pages 472–479, 2005.

[13] J. H. Paik. A Novel TF-IDF Weighting Scheme for Effective Ranking. In *Proc. of the 36th ACM SIGIR*, pages 343–352, 2013.

[14] R. Song, L. Yu, J. R. Wen, and H. W. Hon. A Proximity Probabilistic Model for Information Retrieval. *Tech. Rep., Microsoft Research*, 2011.

[15] X. Wei and W. B. Croft. LDA-Based Document Models for Ad-hoc Retrieval. In *Proc. of the 29th ACM SIGIR*, pages 178–185, 2006.

[16] X. Wei and W. B. Croft. Modeling Term Associations for Ad-Hoc Retrieval Performance within Language Modeling Framework. In *Proceedings of the 29th European Conference on IR research*, pages 52–63, 2007.

[17] X. Yi and J. Allan. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval (ECIR'09)*, pages 29–41, 2009.

[18] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM TOIS*, 22(2):179–214, 2004.

[19] J. Zhao, J. X. Huang, and B. He. CRTER: Using Cross Terms to Enhance Probabilistic IR. In *Proc. of the 34th ACM SIGIR*, pages 155–164, 2011.

[20] J. Zhao, J. X. Huang, and Z. Ye. Modeling Term Associations for Probabilistic Information Retrieval. *ACM Trans. Inf. Syst.*, 32(2):1–47, 2014.