An Empirical Study of SLDA for Information Retrieval

Dashun Ma[1], Lan Rao[2], and Ting Wang[1]

[1] College of Computer, National University of Defense Technology
410073 Changsha, Hunan, P.R. China
{madashun,tingwang}@nudt.edu.cn
[2] College of Humanities and Social Sciences, National University of Defense Technology
410073 Changsha, Hunan, P.R. China
raolan21@21cn.com

Abstract. A common limitation of many language modeling approaches is that retrieval scores are mainly based on exact matching of terms in the queries and documents, ignoring the semantic relations among terms. Latent Dirichlet Allocation (LDA) is an approach trying to capture the semantic dependencies among words. However, using as document representation, LDA has no successful applications in information retrieval (IR). In this paper, we propose a single-document-based LDA (SLDA) document model for IR. The proposed work has been evaluated on four TREC collections, which shows that SLDA document modeling method is comparable to the state-of-the-art language modeling approaches, and it's a novel way to use LDA model to improve retrieval performance.

Keywords: Information Retrieval (IR), Language Model, Document Model, Pseudo-Feedback, Latent Dirichlet Allocation (LDA).

1 Introduction

The language modeling approach has been successfully applied to many IR tasks [16]. However, the state-of-the-art language model is a unigram language model because of the computational complexity. Although various heuristics (e.g. proximity [13]) and resources (e.g. WordNet [3]) have been used to improve it, the unigram language model can hardly capture semantic information in an article. For example, considering trying to match the following query in a set of articles -- *pianist*, the unigram language modeling approach intends to find documents that include words "pianist", "piano", or "musician". A sentence such as "Her hands mercilessly pounded the keys, notes cascading into the surrounding stairway." would be likely assigned a poor score, but obviously, this sentence is closely related to *pianist*.

Using topic models for document representation is an interesting and exciting research in IR. The Latent Semantic Indexing (LSI) model [5] and the probabilistic Latent Semantic Indexing (pLSI) model [9], especially the recent Latent Dirichlet Allocation (LDA) model [2], all focus on reducing high-dimensional data vectors to lower-dimensional representations. Compared with the unigram language model, LDA model has several advantages: (1) It creates a topical level between words and

M.V.M. Salem et al. (Eds.): AIRS 2011, LNCS 7097, pp. 84–92, 2011.
© Springer-Verlag Berlin Heidelberg 2011

documents [2], which gives a better generalization performance.[1] (2) Unlike the unigram language model often using interpolated score, LDA integrates syntax [8], specific information [4] and word burstiness [12] into the document generative process naturally. (3) It offers a method for using semantic information in IR; it is likely to highly rank documents that are related to the topic (even if they don't necessarily contain the exact query terms or their synonyms [4]). However, it is not optimistic about directly using the LDA modeling approach as the document representation in the IR literatures. Wang et al. [14] tested the TNG (topical n-gram model, a variant of LDA) and LDA for IR on the SJMN (San Jose Mercury News) collection, and pointed out that when the two models are directly applied to do ad-hoc retrieval, the performance is very poor (their average precisions are 0.0709 and 0.0438, which are much lower than the state-of-the-art language modeling approaches). We believe that there are two reasons which restrict the application of LDA model in IR. First, good document model does not always bring good retrieval performance [1], other factors (e.g. retrieval method, smoothing strategy, etc.) are also important. Second, training LDA model for a corpus is too inefficient and the corpus-level topics are not fit for each document in the set.

This paper proposes a generative probabilistic model for a document, which tries to deal with the constraints of applying LDA model in IR mentioned above. The model, which we call the single-document-based LDA (SLDA) model, is an extension to the LDA model. In this paper, we further investigate the parameter setting and retrieval method of SLDA, and compare it with the state-of-the-art language modeling approach (the KL-divergence retrieval model [10]) on four typical TREC collections. The experiment results show that (1) the appropriate topic number of SLDA model is less than five; (2) the query likelihood retrieval method is suitable for SLDA model; (3) compared with using the LDA model directly as the document representation, using the SLDA model can obtain better retrieval performance, which competes with the current state-of-the-art approaches.

This paper is organized as follows. The related work is reviewed in Section 2. The SLDA model is defined in Section 3. In Section 4, the experiments are presented. Finally we conclude the work in Section 5.

2 Related Work

Wei and Croft [15] believed that the LDA itself may be too coarse to be used as the representation for IR, so they proposed three ways to integrate the LDA model into the language modeling framework. Their method made the LDA-based document model consistently outperform the cluster-based approach [11] and is close to the Relevance Model.

Chemudugunta et al. [4] proposed a mixture model named SWB, modeling the special words into generative model. Based on the modified AP and FR collections, SWB improves the retrieval performance, and beats the TF-IDF retrieval method.

Wang et al. [14] presented a topical n-gram (TNG) model that automatically determines unigram words and phrases according to the context and assigns mixture

[1] The document model evaluated by LDA has lower *perplexity* score than by the unigram language model.

of topics to both individual words and n-gram phrases. Although directly employing TNG gets poor retrieval performance, significant improvements still can be achieved through a combination with the basic query likelihood model.

All the previous approaches used LDA (or modified LDA) as an assistant to language model. Directly using LDA as document representation hurts retrieval performance badly. In our work, we employ SLDA document model alone, and get the comparable results to the state-of-the-art. Our goal is not to argue that SLDA model can take the place of language modeling approach in IR, but to prove that the LDA modeling approach has been underexploited, and show a novel way to use LDA model to improve retrieval performance.

3 SLDA Modeling Framework

In information retrieval, most of existing works on LDA model are set for the corpus and assume that all the documents are consistent with the same probability distribution. Figure 1(a) shows the graphical model representation of the standard LDA model. There are *C* documents and *K* is the number of topics. θ represents the document-topic multinomial and ϕ represents the topic-word multinomial. α and β are parameters of Dirichlet priors for θ and ϕ. For each document *D*, the N_D words are generated by drawing a topic *t* from the document-topic distribution θ and then drawing a word *w* from the topic-word distribution ϕ.

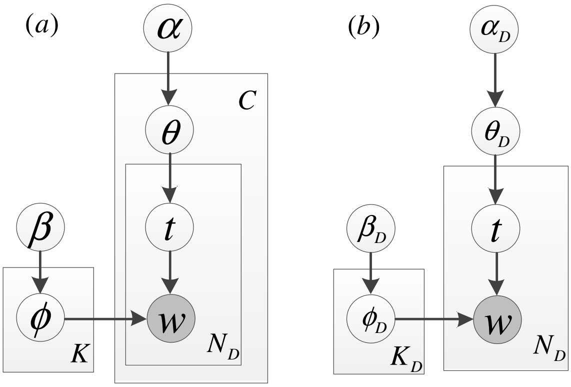

Fig. 1. Graphical models for LDA (a) and SLDA (b)

3.1 SLDA Model

For each document in the collection, corpus-level topics are too coarse. Intuitively, the number of topics in a corpus is much larger than that in single document. It is clear that using a large number of topics to represent the semantics of a document is not unavailable, but the semantic representation of the document would be too generalized, which will make no obvious semantic difference between two documents. What's more, the LDA model of whole corpus implies relationships among words of different documents, but these relationships would be useless or even harmful to the retrieval score. Therefore, we decide to establish LDA model on single document. Figure 1(b) shows the graphical model for SLDA model. It should be noted that SLDA has a similar structure to the LDA model while the difference is that all the parameters are for single document, not corpus.

3.2 Relevant Metrics

Query Likelihood Method. The basic approach for language modeling for IR is the query likelihood method, which takes the maximum likelihood of the document model generating the query terms under the "bag-of-words" assumption as the relevance between query and document. Given a query Q, the retrieval score of a document D is:

$$Score_{QL}(D,Q) = p(Q \mid \varphi_D) \ . \tag{1}$$

where φ_D is a document model of D (i.e. $p(w \mid D)$ for each word w of D). So, we can use SLDA to create φ_D for each document D, and then compute the relevant score via query likelihood method.

Like language model, SLDA need to take some smoothing strategy to handle the sparseness problem of assignment zero probability to unseen words. In this work, we take the Jelinek-Mercer (fixed coefficient interpolation) smoothing method for SLDA.

$$p(w \mid D) = (1-\lambda) p_{slda}(w \mid D) + \lambda p_{slda}(w \mid Ref) \ . \tag{2}$$

where $p_{slda}(w \mid Ref)$ is the reference model, i.e. the maximum likelihood estimate of word w in the background collection. And $p_{slda}(w \mid D)$ represents the SLDA model for a document, its construction process is that: at first, we estimate the parameters θ and ϕ using Gibbs Sampling [6, 7], after get their posterior estimates $\hat{\theta}$ and $\hat{\phi}$, then we calculate the probability of a word in a document as follows,

$$p_{slda}(w \mid D) = p(w \mid D, \hat{\theta}, \hat{\phi}) = \sum_{t=1}^{K} p(w \mid t, \hat{\phi}) p(t \mid \hat{\theta}, D) \ . \tag{3}$$

Negative KL-divergence and JS-divergence. LDA model can infer a new model on a different set of data using existing model on old dataset, so we get another idea to compute the relevance of document and query. After get each document SLDA model, we use it to infer the query SLDA model, so we get document-topic and query-topic multinomial distributions on the same topic collections, then we can use the Kullback-Leibler (KL) divergence (a non-symmetric measure of the difference between two probability distributions) of these two models to measure how close they are to each other and use their negative distance as a score to rank documents as follows:

$$Score_{nKL}(D,Q) = -D(\theta_Q \parallel \theta_D) = -\sum_{t=1}^{K} p(t \mid Q) \log \frac{p(t \mid Q)}{p(t \mid D)} \ . \tag{4}$$

where θ_Q represents query-topic distribution and θ_D represents document-topic distribution.

In probability theory and statistics, the Jensen-Shannon (JS) divergence is a popular method of measuring the similarity between two probability distributions. It is a symmetrized and smoothed version of the KL-divergence:

$$Score_{JS}(D,Q) = (D(\theta_Q \,||\, \theta_M) + D(\theta_D \,||\, \theta_M))/2 \ . \tag{5}$$

where $\theta_M = (\theta_Q + \theta_D)/2$, we also use it as the metric.

4 Experiments

4.1 Dataset and Experiment Setup

The proposed work has been evaluated on four collections from TREC: AP (Associated Press News 1988-1989), FR (Federal Register), SJMN (San Jose Mercury News) and TREC8 (the ad hoc data used in TREC8) with three different TREC topic sets, TOPIC 51-100, TOPIC 101-150 and TOPIC 401-450. Queries are taken from the *title* field of topics. Table 1 shows some basic statistics on these data sets.

Table 1. Statistics of data sets

Collection	query	avgdl	#docs	#qrels
AP	51-100	462	164,597	6101
FR	51-100	1495	45,820	502
SJMN	51-150	408	90,257	4881
TREC8	401-450	480	528,155	4728

All the experiments make use of Lemur toolkit[2] and Gibbs Sampling LDA toolkit[3] for implement. Both the queries and documents are stemmed with the Porter stemmer. Besides stemming, a total of 418 stop words from the Lemur stoplist are removed.

4.2 Comparison of Relevant Metrics

First of all, we compare the relevant metrics on AP dataset. Besides the query likelihood method (QL), negative KL-divergence method (nKL) and JS-divergence method (JS) mentioned above, we add two additional strategies. The motivation is that when we take document-topic and query-topic as two K-dimensional vectors, we can use angle and Euclidean distance of these two vectors (named VA and VD) to measure their relevance. We use Dirichlet priors in the SLDA estimation with

[2] http://www.lemurproject.org/

[3] http://gibbslda.sourceforge.net/

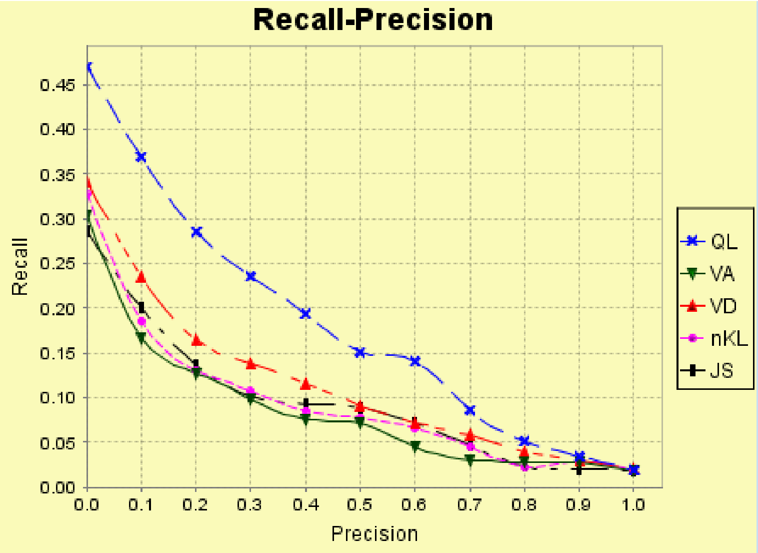

Fig. 2. The comparison of five metrics

$\alpha = 50/K$ and $\beta = 0.1$, which are the common and default settings in current research [7]. We will try different values of K and Gibbs sampling iterations in following experiments, so that we just fix the number of topics with K=100 and set the number of iterations with 2000. The interpolated Recall-Precision curve represents the results in Figure 2, which shows that the QL method significantly outperforms the others, so we use query likelihood as the retrieval strategy in the rest experiments.

4.3 Parameter Settings

There are several parameters that need to be determined in our experiments. For the SLDA model, the number of topics and the number of iterations are very important in topic modeling. At the current stage of our work, we select these parameters through exhaustive search manually. We have tried different iteration numbers with different numbers of topics to see the MAP (Mean Average Precision) values of retrieval results on the AP set. [4]

Table 2 shows the retrieval results on AP with different number of topics (K) and iterations. We find that the performance impact of different numbers of iterations is not very obvious; and the best selection of K is less than five, performance is significantly lower when there are more than 10 topics. Therefore, 50 iterations and K=3 are a good tradeoff between accuracy and efficiency.

In order to choose a suitable value of λ on (**2**), we take a similar experiment process as above on the AP collection and find 0.5 to be the best value for performance.

[4] The results on others corpus show the same trends, so we only list results on AP.

Table 2. Results (MAP) on AP with different *K* and iterations

K	**iterations**		
	50	200	1000
2	0.2460	0.2461	0.2458
3	0.2467	0.2465	0.2467
5	0.2442	0.2445	0.2421
10	0.2398	0.2372	0.2371
20	0.2327	0.2315	0.2303
30	0.2307	0.2268	0.2266
50	0.2261	0.2257	0.2249

4.4 Comparison with Language Model

We compare the performance of the SLDA model (with query likelihood retrieval method) with the KL divergence language model [10] (noted as LM) with Dirichlet prior smoothing (we set the smoothing parameter to 2000, which are the common settings in current research [17]) on the TREC collections. There are two comparison experiments named Rank and Re-rank. For the "Rank" experiment, we use the two models to retrieve top-ranked 1000 documents on the whole collections for each query and compare their retrieval performance. In the other "Re-rank" experiment, we first use the baseline model (i.e. LM) to retrieve 2000 documents for each query, and organize these initial retrieved documents as a subset of the corpus, then re-rank the subset and use top 1000 documents for all runs to compare performance.

Table 3. The comparison of LM and SLDA retrieval results

Metrics	*Rank*							
	AP		FR		SJMN		TREC8	
	LM	SLDA	LM	SLDA	LM	SLDA	LM	SLDA
MAP	0.2568	0.2464	0.1127	0.0982	0.1921	0.1755	0.2312	0.2209
P@10	0.3980	0.3780	0.0680	0.0640	0.2760	0.2530	0.4360	0.4160
#rel_ret	3454	3553	232	272	2987	2943	2764	2752
Metrics	*Re-rank*							
	AP		FR		SJMN		TREC8	
	LM	SLDA	LM	SLDA	LM	SLDA	LM	SLDA
MAP	0.1810	0.1686	0.0708	0.0575	0.1592	0.1182	0.1823	0.1641
P@10	0.3280	0.2700	0.0560	0.0440	0.2630	0.1960	0.2820	0.3100
#rel_ret	2928	2940	153	176	2590	2349	2398	2269

The re-rank step could be considered as using the pseudo-feedback technology in retrieval task. Generally, the language modeling approach using pseudo-feedback documents to re-estimate the query model. However, in our experiment, we employ the query likelihood retrieval model which cannot accommodate the feedback information naturally [16]. Therefore, unlike the traditional method, we use the subset to train the background model (reference model) and re-estimate the document model. In order to facilitate a fair comparison, we also use the pseudo-feedback documents to re-estimate document model in language modeling approach. Table 3 shows the comparison of the two models.

In Table 3, we can observe that on the AP and TREC8, SLDA is comparable to the LM; on the FR and SJMN, SLDA falls a bit behind. Fortunately, on all the collections, the recall of SLDA method is good, even higher than LM approach. Therefore, there are much room for improvement. What's more, on SJMN, SLDA (MAP 0.1755, recall 2943) is more superior to TNG (MAP 0.0709, recall 2450) [14] and LDA (MAP 0.0438, recall 2257), which probably means that, SLDA has an advantage over traditional LDA-like models on document representation in IR.

5 Conclusions and Future Work

Using LDA as an aid can improve the retrieval performance; however, directly using LDA as representation of document hurts the retrieval performance [14, 15]. In this paper, we propose the SLDA model which employs LDA model directly on single document representation. Our experiment results show that SLDA document model is close to the current state-of-the-art language modeling approaches, which is better than traditional LDA models to improve information retrieval performance. We think that the LDA modeling approach has been underexploited, our goal is not to argue that SLDA model can take the place of language modeling approach in IR, but to show a novel way to use LDA model to improve the retrieval performance.

We further study the parameter settings and retrieval model of SLDA. Experiment results on four TREC test collections show that the appropriate topic number of SLDA model is less than five and the query likelihood retrieval method is suitable.

Our work can be extended in several directions: First, although we have found empirically that document-level topics are better than corpus-level topics for document representation, how to determine the number of topics is still a very important problem. Second, for different documents, setting different number of topics instead of a fixed number of topics for all the docs is an interesting direction. Finally, it is challenging to develop a method to define the Gibbs Sampling iterations.

Acknowledgments. The research is supported by National Natural Science Foundation of China (60873097, 60933005). We thank the anonymous reviewers for their useful comments.

References

1. Azzopardi, L., Girolami, M., van Risjbergen, K.: Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In: Proc. of 26th SIGIR, pp. 367–370 (2003)

2. Blei, M., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Cao, G.H., Nie, J.Y., Bai, J.: Integrating Word Relationships into Language Models. In: Proc. of 28th SIGIR, pp. 298–305 (2005)
4. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: Proc. of 19th NIPS, pp. 241–248 (2006)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
6. Geman, S., Geman, D.: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions Pattern Analysis and Machine Intelligence 6, 721–741 (1984)
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. of the National Academy of Sciences 101(suppl. 1), 5228–5235 (2004)
8. Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J.: Integrating topics and syntax. In: Proc. of 17th NIPS, pp. 537–544 (2005)
9. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. of 22nd SIGIR, pp. 35–44 (1999)
10. Lafferty, J.D., Zhai, C.X.: Document language models, query models, and risk minimization for information retrieval. In: Proc. 24th of SIGIR, pp. 111–119 (2001)
11. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proc. of 27th SIGIR, pp. 186–193 (2004)
12. Madsen, R.E., Kauchak, D., Elkan, C.: Modeling Word Burstiness Using the Distribution. In: Proc. of 22nd ICML, pp. 298–305 (2005)
13. Tao, T., Zhai, C.X.: An Exploration of Proximity Measures in Information Retrieval. In: Proc. of 30th SIGIR, pp. 295–302 (2007)
14. Wang, X.R., McCallum, A., Wei, X.: Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In: Proc. of IEEE 7th ICDM, pp. 697–702 (2007)
15. Wei, X., Croft, W.B.: LDA-Based Document Models for Ad-hoc Retrieval. In: Proc. of 29th SIGIR, pp. 178–185 (2006)
16. Zhai, C.X.: Statistical Language Models for Information Retrieval: A Critical Review. Foundations and Trends in Information Retrieval 2(3), 137–213 (2008)
17. Zhai, C.X., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In: Proc. of 24th SIGIR, pp. 334–342 (2001)