A Re-ranking Method Based on Tag-Topic Model

Maoyuan Zhang
Academy of Computer Science,
Central China Normal University
zhangmy@mail.ccnu.edu.cn

Shuiyin Chen
Academy of Computer Science,
Central China Normal University
sychen.ccnu@gmail.com

Fanli He
Academy of Computer Science,
Central China Normal University
HeFanli2013@163.com

***Abstract*—We describe a novel approach to improve the accuracy of the IR which combines the tag-topic model with natural language processing. Tag-topic model extended the Latent Dirichlet We describe a novel approach to improve the accuracy of the IR which combines the tag-topic model with natural language processing. We get tag though semantic fingerprint. Tag-topic model extended the Latent Dirichlet Allocation(LDA) model by adding the data set with 901446 documents, and train the model with different number of topics, we can obtain three important distributions: the document-tag distribution, the tag-topic distribution and the topic-word distribution by using the Tag-topic model. Then through the matrix operation we can get the tag-word distribution which quantify the importance of each word of the document. Finally, based on this distribution these documents are re-ranked. Experiments on NTCIR-5 document collection for SLIR(Single Language IR) show that this method achieves an 13.6% and 19.6% improvement comparing to the initial retrieval method without any re-ranking**.

Keywords-Information Retrieval; Re-ranking; LDA;Tag-topic model

I. INTRODUCTION

Given a query, information retrieval(IR) system should return a ranked list of retrieval documents to users. Retrievaled documents are ranked in the order of their probabilities of relevance to the query.

In 2001, Lafferty and Zhai proposed a statistical language model of KL distance retrieval model. In this model each query and document correspond to a certain language, KL distance of the query language and the language of the document as a correlation measure.

All these model above have their own advantages, however the shortcomings are also obvious, one of the drawbacks is that data sparseness problem, secondly, these algorithms lack of mining for the document semantics. In order to solve these problems, It appears the Latent Semantic Analysis(LSA)[2], Probabilistic Latent Semantic Analysis(PLSA)[3], LDA[4] and Tag-LDA[5] model recent years, compare to the other models, The Tag-LDA model takes both semantic related words and words themselves into consideration.

II. TAG-TOPIC MODEL

A. Tag-Topic Model theory

The Tag-Topic model is a 4-level hierarchical Bayes generative probability model which are generated as follows, We select a topic based on the distribution of tags over the topics, and then generate the corresponding word through the distribution of topics over the words. Repeat this operation to get a document.

This model has the following assumptions: for each word w in a document d, a tag t is sampled from the tag distribution ψ_d, and then a topic z is drawn from θ conditioned on the tag t, following the word w is drawn from ϕ conditioned on the topic z. The document d is generated by repeating the process Nd times, which is the number of word tokens in document d.

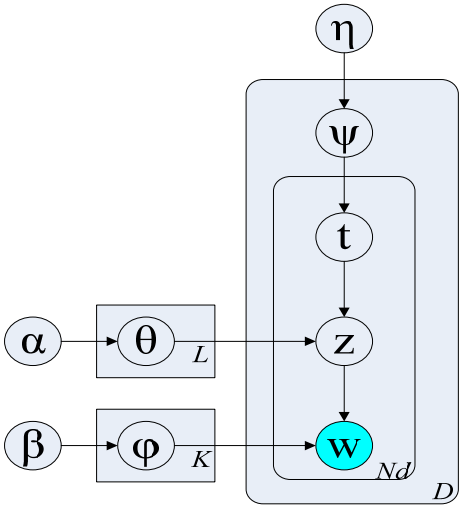

Fig. 1 Plate notation representing the Tag-Topic Model

In Figure 1, ψ is a D×L matrix drawn from a symmetric Dirichlet(η) prior, each row is the tag mixture proportion

for a document d. θ is a L×T matrix drawn from a symmetric Dirichlet(α) prior, and denotes the multinomial distribution over topics for the L tags. Φ denotes the topic-word distribution as described in LDA. For each word w, t and z denote the tag and topic responsible for generating that word.

III. RETRIEVAL SYSTEM BASED ON TAG-TOPIC MODEL

We have successfully developed a retrieval system based on tag-topic model. It has document module, preprocessing module, Index building module, Tag-Topic model parsing module, Inverted index module, User inputing query module, Similarity calculating module and Final ranking documents module.

It is shown in Figure2. We will discuss some important parts in this section.

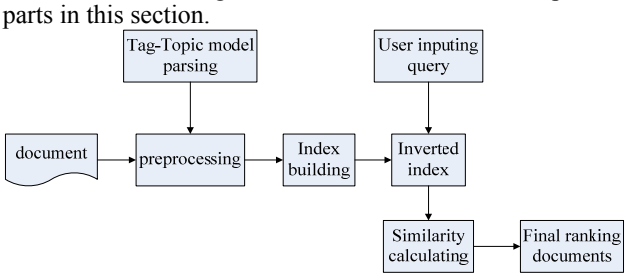

Fig.2 The retrieval system based on tag-topic model

A. Tag-Topic Model parsing

NTCIR-5 as the corpus. First, for every word of each document we get the product of word frequency inverse document frequency, then take corresponding words of the top three products as semantic fingerprint and process all document as Tag-LDA input. We think the more high product of a word , it more represent of this paper. So we take the topic three word as tags of this paper. Second, we obtain two matrices which are tag-topic distribution θ and topic-word distribution Φ by using Tag-LDA to model the documents. We will use the value of these two matrix , they contains quantity of relationship between the semantic fingerprint, topic and word. Third,we get a tag-word matrix through matrix operations. The distribution of this matrix is the tag on the words of a document:

$$Tag - LDA_d(w) = \sum_{j=1}^{T} p(w_i \mid k_i = j)\, p(k_i) \quad (1)$$

Where $p(w_i \mid k_i = j)$ is the probability that the word belongs to the k_i tag. So the value of Tag-LDA(w) for each word reflects the similarity between the word and the document in tag area. And finally we will use it to change the score of document ranking.

B. Documents Ranking by similarity

Take the free/open source example, it simply considered the proportion of the keywords appearing in the document. Its basic similarity scoring formula is introduced as following:

$$Score(q,d) = \sum_{t\ in\ q} \frac{tf(t\ in\ d) \times boost(t.field\ in\ d)}{\times lengthNorm(t.field\ in\ d)} \times coord(q,d) \times queryNorm(q) \quad (2)$$

It computed the score for each document d matching each keyword t in a query q. Tf[14] represents the term frequency. Idf(t)[14] stands for inverse document frequency. Typically, coord(q,d) is the factor based on how many terms found in the document.

In order to consider the inner semantic of a document, we use tag-word matrix in section 4.1 to quantitative descriptions of contribution that each word to the content of document. In our method we set this quantitative value as boost through payload function. Thus, suppose q represents the query, d represents a document, and then score of d derived from its latent semantic degree to express its relevance to query q is define as follows:

$$Score(q,d) = \sum_{t\ in\ q} \frac{tf(t\ in\ d) \times boost(Tag - LDA_d(w))}{\times lengthNorm(t.field\ in\ d) \times coord(q,d)} \times queryNorm(q) \quad (3)$$

IV. EXPERIMENTS AND EVALUATION

A. Comparison of Experimental Results

MAP[16] is Mean Average Precision, or an average accuracy rate method for short, which is defined as seeking out after each relevant document retrieval accuracy average the arithmetic mean. Here on the accuracy required twice the average, so called Mean Average Precision. MAP is reflected in all relevant documents on the system performance of a single-valued index. The system retrieves the relevant documents out of the more forward, MAP should be higher. If the system does not return relevant documents, accuracy rate defaults to 0. It's formula is below.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

R-Precision[17], precision at R-th position in the ranking of results for a query that has R relevant documents. The measure is highly correlated to Average Precision. Also, Precision is equal to Recall at the R-th position. It's formula is below.

$$R - \Pr ecision = \frac{\text{the number of related documents in first R documents}}{R}$$

Precision at k documents(P@K) is still a useful metric, the system returns the first 10 results for accuracy. P@50 are the most important factors to evaluate the performance of information retrieval system.

In our experiments, we compare these three indicators obtained by different methods. We named these three methods LC, LDA[13] and Tag-LDA for short in order to describe it conveniently.

1) Comparison of MAP. This paper shown the results of MAP in Table 1(relax and rigid respectively).

TABLE I: COMPARISON MAP RESULTS ON NTCIR-5 COLLECTION

Standard	LC	LDA	Tag-LDA	
	MAP	MAP	MAP	Change over LDA(%)
Rigid description	0.1867	0.2032	0.2233	+9.89
Relax description	0.2361	0.2465	0.2683	+8.84

We compared with two methods, first is Lucene version 3.4 as the baseline system, named LC method, and then LDA method based on Lucene named LDA. Our method named Tag-LDA method. The table shows the mean average precision for each case. It show that Tag-LDA method the best results of all the three methods.

2) Comparison of the precision at R documents

Precision at R documents represented the precision for a query that has R relevant documents. This measure is highly correlated to Average Precision. Table 2 shows the results of our experiments in the precision at R documents.

TABLE II: COMPARISON R-PRECISION RESULTS ON NTCIR-5 COLLECTION

Standard	LC	LDA	Tag-LDA	
	R-Precision	R-Precision	R-Precision	Change over LDA(%)
Rigid description	0.2167	0.2228	0.2499	+12.2
Relax descriptiom	0.2759	0.2772	0.3078	+11.0

3) Comparison of P@. Considering the practical information retrieval system, the majority of users only have patience to check the most N relevant feedback. The value of N generally take 10 to 30. And p@ is an indicator to reflect the precision in the top list documents. Figure 3 and figure 4 show the result of precision in 5, 10, 15, 20 docs in the relax situation and the rigid situation respectively.

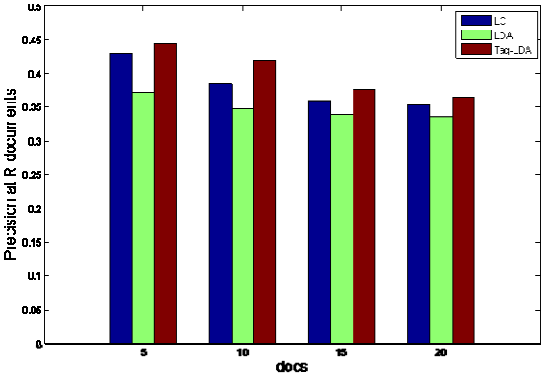

Fig. 3: Comparison of p@ on NTCIR-5 collection in relax situation

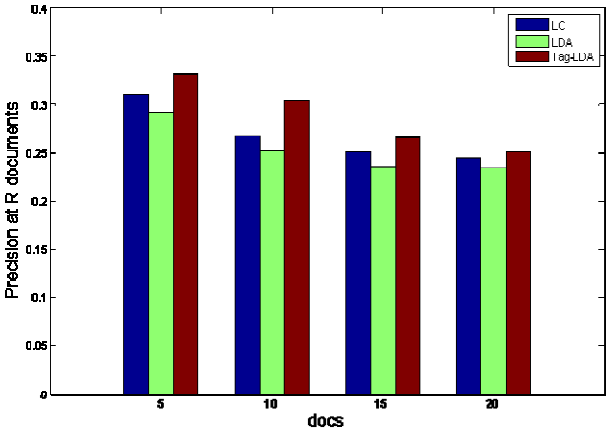

Fig. 4: Comparison of p@ on NTCIR-5 collection in rigid situation

V. CONCLUSIONS

This paper proposes a novel method to improve the performance of information retrieval system by combining the tag-topic model with natural language processing(IR documents reranking). It solves the problem that can't mining well the document with tag for long time. In our paper we use Tag-Topic model to get the semantic information of documents. In practice researchers attempt to fit appropriate model parameters to the data corpus to make our model performs well.

Our ranking algorithm take the semantic of document into account. The experiment shows that our system has a better performance on NTCIR-5.

Furthermore, there are many potential further developments. It would be interesting to further study to dynamically set the number of topics. In addition, the model could be further extended with auto tagging information.

For the future work, we will try to extend the Tag-LDA model, make it have better performance in IR field.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 61003192), the self-determined research funds of CCNU from the colleges' basic research and operation of MOE(No. CCNU13A05014, No. CCNU13C01001) and The Major Project of State Language Commission in the Twelfth Five-year Plan Period ZDI125-1.

REFERENCES

[1] Ponte JM, Croft W B. A Language Modeling Approach to Information Retireval[C]//Proceeding of the 21st International ACM SIGIR Conference on Reasearch and Development in Information Retrieval, 1998: 275-281

[2] A.Letsche, Toward large-scale information retrieval using latent semantic indexing. Master's thesis, University of Tennessee, Knoxville, Tennessee, August (1996)

[3] Thomas Hofmann, Probabilistic Latent Semantic Indexing, Proceedings of the Twenty-Second Annual International SIGIR

Conference on Research and Development in Information Retrieval(SIGIR-99), (1999).

[4] Blei, David M.; Andrew Y. NG, Michael I. Jordan(2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3:993-1022.

[5] HE Ting-Ting,LI Fang. Semantic Knowledge Acquisition from Blogs with Tag-Topic Model[J]. China Communications, 2012, 9(3): 38-48.

[6] Girolami, Mark; Kaban, A. (2003). "On an Equivalence between PLSI and LDA". Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3

[7] Bishop, Christopher M. (2006). Pattern Recognition and Machine Learning. Springer. ISBN 0-387-31073-8.

[8] Casella, George; George, Edward I. (1992). "Explaining the Gibbs sampler". The American Statistician 46 (3): 167–174.

[9] Casella, George; George, Edward I. (1992). "Explaining the Gibbs sampler". The American Statistician 46 (3): 167–174. doi:10.2307/2685208. JSTOR 2685208.(Contains a basic summary and many references.)

[10] Gelfand, Alan E.; Smith, Adrian F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities". Journal of the American Statistical Association 85 (410): 398–409. doi:10.2307/2289776. JSTOR 2289776. MR 1141740.