A Retrieval Sorting Approach for Online Forums Based on Domain Topics

Yuyan Zhang
School of Computer Engineering
Weifang University
Weifang 261061, China
jkhou@163.com

***Abstract*—Topical search engine is an extension of general-purpose search engines, which has become an important research subject in Web information retrieval recently. Focusing on the development of Web 2.0 applications, a result ranking approach is proposed on the basis of LDA model to rank the search results from Web forums. Compared with traditional methods, this approach takes up less storage space, and can more quickly and accurately respond to user inquiries. This work has important significance for the research of improving the performance of retrieval results of web forums.**

Keywords- information retrieval; text classification; topic detection; retrieval sorting

I. INTRODUCTION

With the explosion of the information available online, Web searching for full-scale, accurate and high quality information has become increasingly difficult. Online Web-forums and BBS are website which can accommodate the discussion and respond information generated by users for a long time. It has become an important medium for Internet users to interact with each other. It provides a visual platform for users to express their view comments, experiences, thoughts and emotions. There are vast deposits of valuable knowledge and information of online forums for the involvement of thousands of users. The coverage of the topics of these knowledge and information are very extensive, which includes news, entertainment, sports, games, arts, society, science, family and health topics. Many of these topics are closely linked with our lives. However, they are rarely seen in the traditional Web pages. According the survey from the website of comscore.com, 81% of the Internet users will get information from the Web forums with the corresponding topics before they buy the products (such as cars, digital cameras, etc). Quite considerable portion of the users will share their experience in the online forums [1]. In era of the Web2.0, online forums develop at an unbelievable speed, which has become an important source of information [2].

The data in the online forums is the same dynamic growth. Every day, a large number of users create thousands of new posts. Faced with such a large amount of contents, the user can not explore them once over. Therefore, it is necessary for the information retrieval search orienting the forum, and the user can quickly find the information they want. At present, many commercial search engines such as Google and Baidu, are dedicated to provide special retrieval services for the Web forums. Since the posting in the Web forum is free to act, so the quality of posts is often uneven. There are a large number of high-quality posts and low-quality ones. Thereby, how to sort the retrieve results according to their qualities, and make high-quality contents be seen with priority, is a problem worthy of study.

Because the content of web forum posts are often very short, the traditional sorting method based on similarity is often with a weak effect on retrieval of the web forums. There is no explicit link between the forum posts, so the link-based methods, such as PageRank, HITS, etc, it is difficult to obtain good results [3]. The content similarity was introduced into link diagram by some researchers to build a topic hierarchy diagram, and calculated the score of posts based on the topic hierarchy. In the process of creating the topic hierarchy, a method called co-clustering is used, but this method is not suitable for large-scale application [4].

In addition, because the content of posts is often short, it is not very effective just to cluster based on its text. Response diagram can also be constructed through reply relationship of the forum. An approach named PostingRank is proposed which calculates the post score according to the response structure. The final score can be obtained through the combination of the score from the Web site Google and the former score. The disadvantage of this method is that the external data (Google Score) is needed for support [5].There is no explicit response relationship of many web forums, and the reply relationship of some web forums can not fully reflect the relationship between the posts. Implicit response relationship often between posts which is often difficult to reconstruct. The algorithm makes a large number of posts with more response to be with high scores, but the high number of responses is often not equivalent to high quality.

From the viewpoint of practical application, a retrieval sorting approach for online forums is proposed based on domain topics. In the approach, relevant model will be firstly used to integrate with the forum data, thus to infer the topic distribution of the posts, and lastly to give the score according to the characteristics of topic distribution.

II. THE MATHEMATIC MODEL FOR RETRIEVAL SORTING APPROACH

Information retrieval is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. Most of existing sorting algorithms for retrieval results has been largely focused on web pages. However, the content of

Published by Atlantis Press, Paris, France.

online web forums is more short than the contonent of traditional web pages, and there is no explicit link relationship. The traditional sorting algorithm yield good results when used in the web forums.

The model of LDA (Latent Dirichlet Allocation) is a multi-layer generative probabilistic model, which includes three-tier structure: the words, the themes, and the document [6]. Each document of LDA is represented as a mixed topic, and at the same time, each topic is a polynomial distribution of a fixed term table. All of these topics are shared by all the documents of the collection. A word of LDA is assumed to be generated by a topic mixture. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. The efficient approximate inference techniques are presentted based on variational methods and an EM algorithm for empirical Bayes parameter estimation. The results is reportted in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model. Each document has a specific topics ratio, which is generated from the the sample of Dirichlet distribution. As a production model, the extraction of latent semantic structure and presentation document using the LDA model, have been successfully applied to many text-related fields. The graphics model of LDA is shown in Figure 1.

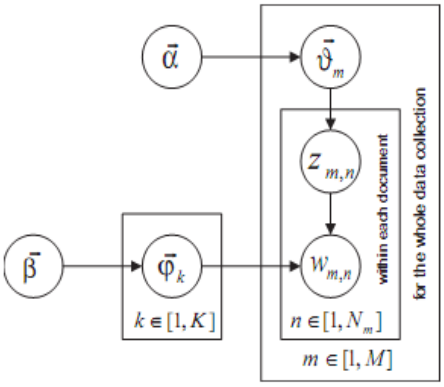

Figure 1. The graphics model of Latent Dirichlet Allocation

There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis, and technologies. The goal of our approach is to create clusters which are internally coherent, but different clearly from each other. Clustering algorithms group a set of documents into subsets or clusters. The process of the retrieval sorting approach is shown as follows.

Given a document collection D, which contains M documents and V different words. Each document d_m contains a word sequence $\{w_1, w_2,..., W_N\}$. In the LDA model corresponding to the collection D, if the number of topics is assumed as a fixed mumber K, then the production of a document d_m can be expressed as the following two processes:

Randomly select a K-dimensional vector θ_m from the Dirichlet distribution $p(\theta|\alpha)$, which represents the the topic mixing ratio of the document d d_m;

Randomly select a V-dimensional vector φ_k from the Dirichlet distribution $p(\varphi|\beta)$, which represents the word distribution of the topic k.

For each word wm in d_m, n:

1 Randomly select a topic zm,n from the multinomial distribution $p(z_m, n|\theta_m)$;

2 Extract a word w_m, n from the multinomial distribution $p(w_m, n|z_m, n, \beta)$.

Given the Dirichlet parameters, the joint distribution of all the known variables and the hidden variables can be achieved through the follwing expression:

$$p(w_m, z_m, \theta_m, \phi \mid \alpha, \beta) =$$

$$p(\phi \mid \beta) \prod_{n=1}^{N_m} p(w_{m,n} \mid \varphi_{z_{m,n}}) p(z_{m,n} \mid \theta_m) p(\theta_m \mid \alpha)$$

The similarity degree of a document wm can be achieved by the integral of Φ and θ_m, and the summation of z_m, which is shown as the fllowing expression:

$$p(w_m \mid \alpha, \beta) = \iint p(\theta_m \mid \alpha) p(\phi \mid \beta) \cdot$$

$$\prod_{n=1}^{N_m} p(w_{m,n} \mid \theta_m, \phi) d\phi d\theta_m$$

Finally, the similarity degree of the entire data set is the product of the similarity degree of all documents:

$$p(W \mid \alpha, \beta) = \prod_{m=1}^{M} p(w_m \mid \alpha, \beta)$$

The experiment by Phan et al have proved that is more effective to determine the similarity degree between sparse texts using topic distribution than that using the similarity degree on the basis of word frequency. Therefore, we firstly use LDA model to fit the data on the web forums, and then predict the topic distribution of each post θ_p.

Next, we give the relationship between the quality of the post and its topic distribution:

First of all, Loulwah AlSumait et al have shown clearly that, LDA model itself would produce some low-quality topics, which mainly contains the following three kinds:

The topic of the word Uniform distribution: the production probability of each word is very even, which can be judged with the similarity of the following benchmarks formula:

$$p(w_i \mid T) = \frac{1}{V}$$

Herein, i={1, 2, …, V}, and V is the total numbers of words.

Empty semantic topics: there is no obvious meaning, and the word distribution is similar to the empirical distribution:

$$p(w_i \mid T) = \sum_{k=1}^{K} p(w_i \mid k) p(k)$$

$$p(k) = \frac{\sum_{m=1}^{M} \theta_{m,k}}{M}$$

Background topics: the generation probability of each document is more evenly, which can be judged with the similarity of the following base formula:

$$p(d_m \mid T) = \frac{1}{M}$$

We find the topics with low quality using the above three criteria. Then, the topic quality score can be calculated according to the the proportion of topics with low-quality of the topic distribution of the post *i*, which is shown as fllows.

$$S_{tq}(i) = 1 - \frac{\sum_{j \in I/J} p(z_j \mid post_i)}{\sum_{k=1}^{K} p(z_k \mid post_i)} = 1 - \frac{\sum_{j \in I/J} \theta_j^{(i)}}{\sum_{k=1}^{K} \theta_k^{(i)}}$$

Typically, the higher the quality of the posts, the more clearly the topic will be. That is to say, in the topic distribution, the probability of a small amount of the topics are more prominent, so we use the Shannon entropy to calculate the topic outstanding score of the post *i*.

$$S_{to}(i) = -L \sum_{k=1}^{K} p(z_k \mid post_i) \log p(z_k \mid post_i)$$

$$= -L \sum_{k=1}^{K} \theta_k^{(i)} \log \theta_k^i$$

The quality of a post is also reflected in the similarity with the query. Therefore, we also using LDA model obtained by fitting approach to predict the topic distribution for querying. Then, the similarity scores of the topic can be calculated through querying the KL distance from *q* to the post *i*, which is shown as fllows.

$$S_{tm} = \sum_{k=1}^{K} \theta_k^{(i)} \log \frac{\theta_k^{(i)}}{\theta_k^{(q)}}$$

Finally, the topic distribution based score (TDBS) of the post is achieved through the fllowing expression:

$$TDBS = aS_{tq} + bS_{to} + cS_{tm}$$

Herein, *a*, *b*, *c* is three parameters, and *a* = *b* = *c* = 1/3.

III. Analysis about the Application of the Retrieval Sorting Approach

We have collected 26,396 posts from the section of "news and current affairs" of the CDC web forums, and the obtained data were fitted using GibbsLDA. The number of topics *K* is setted to 50 ~ 200, α is setted as 0.5, β is setted as 0.1.

Next, 600 queries was constructed, and the relevant search results was sorted by artificial scoring. The first 10 results with highest score obtained by using TDBS, will be tested the the number of the first 10 highest scores in the manual results, which is noted as P@10. It will be combined with the MAP in the test to measure the performance of the proposed approach.

Figure 2 shows the comparison relations between the average performance of four methods (the TDBS, BM2500, PageRank and PostingRank) when *K*=200. It can be seen that our approach is superior to the other three methods. Figure 3 and Figure 4 respectively shows the performance of the proposed method changes with the topic numbers of LDA and the sampling numbers of Gibbs in the fitting process of LDA model. It can be seen that the performance of the method is relatively stable.

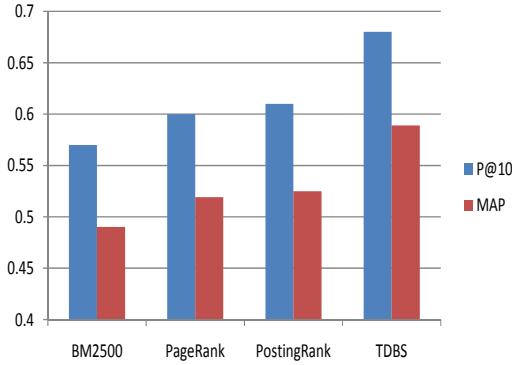

Figure 2. Comparison of the performance of several sorting methods

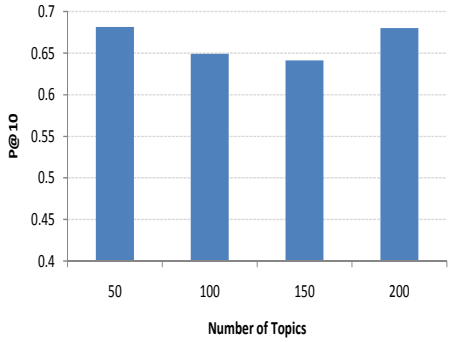

Figure 3. The performance of TDBS changes with the numbers of topics

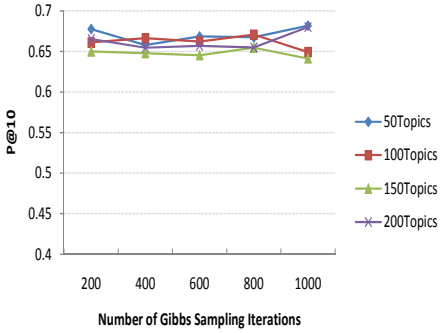

Figure 4. The performance of TDBS changes with the sampling numbers of Gibbs

IV. CONCLUSION

With the development of Web2.0 technology, online forums and Blogs and other network applications have become an important source of information, on which the information retrieval and mining has become an important issue of information retrieval. In this paper, focusing on the extensive application of Web 2.0 technology, a ranking approach is proposed on the basis of LDA model to rank the search results from Web forums. It is more effective than the traditional ones in terms of running time and size of index space. This approach obtained a higher score in the Scales of MAP and P@N. This work has important significancy for the research of result ranking approach for new web forums and the research of improving the performance of search results of web forums.

ACKNOWLEDGMENT

The author is most grateful to the anonymous referees for their constructive and helpful comments on the earlier version of the manuscript that helped to improve the presentation of the paper considerably. This research was supported by the foundation of science-technology development project of Shandong Province of China under Grant No. 2011YD01042 and No. 2011YD01043.

REFERENCES

[1] X. Han, J. Ma. “Hot Research Topic Extraction in Digital Libraries”. Journal of Computational Information Systems, Vol. 6, No. 3, 2009. pp.318-325

[2] Z. Chen, J. Ma, X. Han. „An Effective Relevance Prediction Algorithm Based on Hierarchical Taxonomy for Focused Crawling”. Proceedings of the 2008 Asia Information Retrieval Symposium, Springer LNCS. Vol. 4993. 2008, pp.613-619.

[3] Y. Li, J. Ma, Y. Sun. „Applying Dewey Encoding to Construct XML Index for Path and Keyword Query”. Proceedings of 2009 International Workshop on Database Technology and Applications, 2009, pp.553-556.

[4] D. Blei, A. Ng, and M.. “Jordan. Latent Dirichlet Allocation”. Journal of Machine Learning Research, Vol. 9, No. 3: 2003. pp. 993–1022,

[5] L. Song, J. Ma, J. Lei. “A Semantic Method of Deep Web Classification”. Journal of Information & Computational Science, Vol.5, No. 5, Nov. 2008, pp.2017-2025.

[6] X. Wei, and W. B. Croft. “LDA-based document models for ad-hoc retrieval”. Proceedings of the 29th SIGIR Conference, 2006, pp. 178-185

[7] J. Koehler, R. Hauser, S. Kapoor, F.Y. Wu, S. Kumaran, “A model-driven transformation method”. Proceedings of Seventh IEEE International conference on Enterprise Distributed Object Computing, IEEE Computer Society, 2003, pp. 186-197