A Time-Sensitive Model for Microblog Retrieval

Cunhui Shi, Bo Xu, Hongfei Lin, and Qing Guo

School of Computer Science and Technology,
Dalian University of Technology, Liaoning, Dalian, 116024
{smart,xubo2011,guoqing}@mail.dlut.edu.cn, hflin@dlut.edu.cn

Abstract. Microblog, as a way of online communication, can generate large amounts of information in a very short period. Therefore, how to retrieve the latest relevant information becomes a hot research area. Different from traditional information retrieval (IR), the microblog retrieval emphasizes fresh contents of the information. In order to solve this problem, we extend the traditional IR methods by taking into account the posting time. We propose a time-sensitive retrieval model, which takes the time factor as a prior probability. In the retrieval model, we introduce the pseudo relevance feedback technology as a query expansion approach to improve retrieval performance. Furthermore, we introduce a strategy to filter the initial retrieval results, which takes post quality factors into account including entropy and link features. Experiments on Twitter corpus show that our algorithm is effective to improve the retrieval performance, and the retrieval results can meet the real time retrieval need well.

Keywords: Microblog, Time-Sensitive, Retrieval Model, Entropy.

1 Introduction

In recent years, with the development of microblog, more and more users take part in it to share and obtain information. Statistics shows that until May of 2013, there are totally 288 million active users in Twitter, and they update more than 400 million tweets every day. Meanwhile, microblog becomes more and more popular in China, and many internet companies, such as Sina, Tencent and Sohu, start to provide microblog services. Another statistics shows there are 635 million registered users in Sina microblog until March of 2013, which includes 49.8 million active users and 76.5% of them use microblog on mobile terminals. Since users can update, review and forward microblog posts quickly and easily through mobile terminals, such as cellphones and tablet PCs, microblog generally becomes the first information source of some important issues. Therefore, it poses a challenge to meet the real time requirements for microblog retrieval, and obtaining the most fresh and relevant information becomes increasingly necessary.

In this paper, we propose a method to construct a time-sensitive retrieval model in microblog retrieval, which can retrieve the most relevant and fresh information for microblog users. This model combines the traditional IR methods with a time factor, which is taken as a prior probability. Besides, the pseudo relevance feedback

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 402–409, 2013.

technology is introduced to expand the queries for improving the retrieval performance. Furthermore, we introduce a strategy to filter the initial retrieval results, which takes quality factors into account including entropy and link features. Experimental results on the Twitter corpus show that our algorithm is effective to improve the microblog retrieval performance and meet the real time retrieval requirements well.

The contributions of this paper are as follows: 1) we propose a time-sensitive retrieval model aiming at meeting the real time requirements for microblog. 2) Pseudo relevance feedback technology is introduced to boost the retrieval performance in this process. 3) We take advantage of entropy and link features to get rid of noises in microblog messages in order to obtain more relevant information.

The rest of the paper is organized as follows: Section 2 reviews some related work. In Section 3, we illustrate our time-sensitive retrieval model and some details on our retrieval process. In Section 4, we introduce the experimental settings and present the comparison and analysis of different methods. Section 5 concludes this paper .

2 Related Work

Microblog provides an easier access for us to obtain abundant information and communicate with each other. Along with the data growing rapidly, retrieval in microblog become more and more necessary for users to find the exact information they prefer. There are many researches focusing on microblog retrieval in terms of various aspects. In order to study user behaviors in microblog, Miles Efron [1] gives a general view of microblog retrieval focusing on many important problems to be solved by analyzing semantic features and researching on authority and quality of abstract models and entity retrieval. Besides, the paper is also focusing on the temporal issue, i.e. the influence of the time factor on microblog retrieval, which includes continuous indexing problem and tolerance of information delay in microblog.

Since traditional information retrieval methods like key words matching cannot meet the requirements of massive data very well, some researchers are focusing on real time retrieval, which takes factors, such as timestamp and authority, into account during microblog retrieval to boost the performance [2-9]. For example, Teevan et al. [8] present that microbloggers tend to utilize short queries (always relate to some hot issues) and choose the latest results by analyzing the query logs and searching results. Chen et al. [10] propose an adaptive retrieval framework, which make use of the relationship between users and messages to classify the messages into different categories. They demonstrate that the framework can solve the problem of real time retrieval at the cost of results quality, i.e. it is a trade-off between real time requirements and retrieval performance. Arjumand Youns et al. [11] take twitter as the source of consulting to estimate the popularity of news. Kamran Massoudi [12] improves retrieval performance by using some unique characteristic in microblog, such as forwards and comments. Rinkesh Nagmoti et al. [13] rank the searching results by using information from social network and get good results. By analyzing the structure of social network, Meredith Ringel Morris et al. [14] estimate the reliability of messages on microblog.

Different from the researches referred above, our research explores the possibility of adding time information to traditional retrieval model, which aims at changing the prior probability between documents and the query. Meanwhile, we take some measures, such as pseudo relevance feedback and quality filters, to improve the retrieval performance further.

3 Microblog Oriented Time-Sensitive Retrieval Model

In this section, we will illustrate some details about our time-sensitive microblog retrieval model. Specifically, there are two steps in our retrieval process. Firstly, we introduce time factors to traditional retrieval models through query expansion techniques. Secondly, we use entropy and short links to filter the result documents. Next, we will show the two steps in details.

3.1 Time-Sensitive Language Model

To construct our retrieval model, we focus on the improvement of traditional language model. Traditionally, documents are ranked based on relevance with the query. According to Bayes rules, the relevance is represented as follows.

$$P(D|Q) = \frac{P(D)P(Q|D)}{P(Q)} \propto P(D)P(Q \mid D) \tag{1}$$

where D is the document. $P(Q)$ is the normalization factor, which can be omitted in calculation. $P(Q|D)$ is the likelihood function w.r.t the document D. $P(D|Q)$ is the probability that query Q is generated from D, which is used to measure the relevance between D and Q. Since $P(D)$, as the document prior probability, is uniform and invariable for documents, it is always omitted in calculation. However, our retrieval model is more focused on the latest microblog documents, i.e., our model is time-sensitive. So $P(D)$ here should reflect something related to the timestamp labeled on the document in order to adapt to microblog retrieval. Therefore, we present our basic retrieval model based on language model as the Eq. (2).

$$P(D|Q) = \frac{P(D)P(Q|D)}{P(Q)} \propto P_{real}(D)P(Q \mid D) \tag{2}$$

where $P_{real}(D)$ is the document prior probability with time information, which distinguish the documents in their timestamps. Specifically, the later the document is delivered, the higher the probability is. In other words, a user takes the time he submitted the query as the baseline. If the timestamp of a document is nearer to the baseline, the document has more probability to be ranked higher in the final ranking list and vice versa.

However, if a user wants to search for information related to a certain period (not the time the query submitted), the model in Eq. (2) may not work well. In order to deal with the problem, we present the concept of the key time point with large

amounts of information delivered, which is denoted as $T_{realQuery}$. For example, a hot issue can cause the information explosion in microblog in the short period. We average the timestamp of the documents delivered in the period as the key time point to modify the time factor in Eq. (2) as shown in Eq. (3).

$$T_{realQuery} = \frac{1}{N}\sum_{i=1}^{N} t(FbDoc(i)) \tag{3}$$

where $t(FbDoc(i))$ is the timestamp of a document i. We introduce pseudo relevance feedback to choose the top-N documents from initial retrieval and average the timestamp to obtain the key time point of a query. Empirically, we set N=10 in our experiments.

$$P_{real}(D) = \frac{P'_{\text{real}}(D)}{\sum_{d \in C} P'_{\text{real}}(d)} \propto P'_{\text{real}}(D) \tag{4}$$

$$P'_{real}(D) = e^{-\frac{|T_D - T_{realQuery}|}{T_{max} - T_{realQuery}}} \tag{5}$$

Eq. (4) and Eq. (5) is the calculation of the prior probability with time factors, where T_D is the timestamp of a document, T_{max} is the timestamp with the max distance with the key time point in a period. We calculate the probability using Eq. (5) instead Eq. (4) in order to embody the time information in document prior more reasonably. The final retrieval model we construct could contribute to microblog retrieval in consideration of the time information.

3.2 Quality Based Results Filter

Since the messages in microblog are always short and limited to a certain length, it is difficult to extract text features as traditional information retrieval methods do. What's more, there are lots of noises in microblog messages including some advertisements and some useless comments. In order to solve the problems above, we introduce information entropy to measure and filter the instant messages.

$$Entropy = -\sum_{i=1}^{m} \frac{n_i}{n} \log \frac{n_i}{n} \tag{6}$$

where n is number of words appeared in a piece of message, in which m is the number of words appeared only once and n_i is the number of words appeared more than once. The entropy indicates the importance of a message. We filter the messages with the entropy below a certain threshold to decrease the noises.

Besides, there are always some hyperlinks in the messages, which enrich the contents of messages. So we take the link information as another way to measure the importance of the messages, i.e., if a message contains some hyperlinks, the score of the message will be multiplied by a constant as the following equation shows.

$$score_{new} = \alpha * score \tag{7}$$

In our experiments, we set α=1.2 since relative good performance can be achieved under this setting. We remove the useless information using the two measures above to obtain the documents with abundant information in microblog retrieval.

3.3 The Overall Retrieval Process

In this section, we will give more details about our retrieval process. We examine our retrieval model on TREC datasets from Microblog Track in our experiments. Table 1 shows the overall retrieval process of our algorithms.

Table 1. The time-sensitive retrieval model of microblog

Algorithms	Time-Sensitive Retrieval Model
Input	An original query
Output	documents ranking list
Step 1	Original Query=constructQuery(words[]); PRFDocs[1,2,⋯,10]=Retrieval(Original Query);

Step 2	For i=1 to10
	For each term in PRFDocs[i]
	If(term !=stopper) Weight(term)= tf * idf ;
	End for
	TermsWeight[i]=Combine(Score(PRFDocs[i]),weight(PRFDocs[i]));
	End for

Step 4	RealQueryTime=avgTime($\sum_{i=1}^{10}$ Time(PRFDocs[i]))
Step 5	For each doc computeP(Q|D) = $\frac{P(D)P(Q\|D)}{P(Q)} \propto P_{real}(D)P(Q\|D)$) End for
Step 6	If(HasLink) $score_{new} = \alpha \cdot score$ Else $score_{new} = score$

Step 7	For each doc in the runs
	If(Entropy>=EntropyThrethold&&$score_{new}$>=ScoreThrethold)
	Resultdocs.add(doc)
	End for

Step 8	Return docs[1,2,⋯,N]

Initial Retrieval. A user submits a query and then our system conducts initial retrieval based on traditional language model and vector space model.

Query Expansion. We choose the top 10 documents in the initial ranking list as feedback documents. Then, we select expansion terms from these documents as follows. Firstly we stem the words in documents and remove stop words. Secondly, we

use traditional TF-IDF model to score each term. Thirdly, we choose *k* terms with the highest scores as expansion terms for expanding the query.

Further Retrieval Using Time-Sensitive Model. We retrieve again using query expansion. Meanwhile, entropy and link information is utilized to filter documents and obtain the final ranking list.

4 Experiments

4.1 Experimental Settings

Corpus. We evaluate our method on TREC dataset from Microblog Track based on Twitter. The dataset contains 50 topics with relevance judgments, which is one of the standard dataset in Microblog research. Tweets in the dataset are stored in a standard format, i.e., <*tweetid*, *username*, *status*, *time*, *text*>, where *tweedid* is the identifier of each tweet, *username* is the name of user who delivers the tweet, *time* is a number indicating the timestamp and *text* is the content.

Evaluation Methods. We take Indri as our basic search environment, and MAP (Mean of Average Precision) is adopted as the evaluation measure. In evaluation, the higher the value of MAP is, the better the performance is.

Baselines. To measure the effectiveness of our method, we compare our method with some traditional and state-of-the-art methods on the dataset. Tfidf is the retrieval model using vector space models twice, where the terms are weighted using TFIDF. TfidfFb is the retrieval model using vector space model in initial retrieval and time-sensitive model in second retrieval. LM is the retrieval model using language models twice. LMFb is the retrieval model using language model in initial retrieval and time-sensitive model in second retrieval. MixSp is the linear interpolation of Tfidf and LM methods. MixFb is the linear interpolation of TfidfFb and LMFb methods.

4.2 Experimental Results and Analysis

Table 2 presents the evaluation results of the 6 retrieval models. From the table we can see that no matter what retrieval models is used in consideration of time factor, the performance can be improved. In comparison, language model performs better than Tfidf, indicating that language model is more suitable and stable for microblog documents in extracting key information. What's more, the performance of LMFb and MIXFb is increased over LM and MixSp by nearly 10%, while TfidfFb is increased over Tfidf by only 2%. This phenomenon also indicates that the time-sensitive retrieval model is more effective when used with language model.

We also conduct experiments under 50 topics in the dataset. Figure 1 shows the results of time-sensitive models with different initial retrieval methods. From the results we can see that the performance of two methods linear interpolation is better compared with other methods. Notably, we find that TfidfFb performs the best among all methods in some query topics, which indicates the time-sensitive Tfidf model is not stable on all the queries.

Table 2. Comparision of different retrieval methods

Methods	MAP	
Tfidf	0.2185	
TfidfFb	0.2219	1.56%
LM	0.2603	
LMFb	0.2851	9.53%
Mix	0.2674	
MixFb	0.2936	9.80%

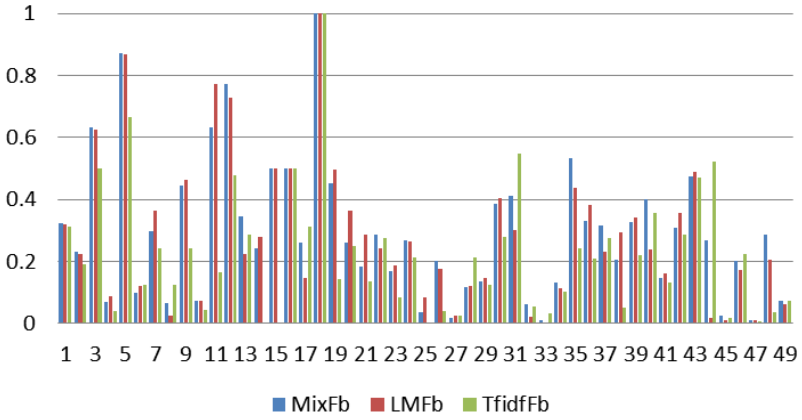

Fig. 1. MAP histogram of 50 topics

5 Conclusion

In this paper, we propose a microblog retrieval model by adding time factor to the traditional language model. Specifically, we use query expansion technique, pseudo relevance feedback, to boost retrieval performance and filter the result documents using entropy and link information. Experiments on TREC datasets show that our method outperforms traditional language model and vector space model. Our retrieval model can meet users' information need better.

Acknowledgements. This work is partially supported by grant from the Natural Science Foundation of China (No.60673039, 60973068, 61277370), the National High Tech Research and Development Plan of China (No.2006AA01Z151), Natural Science Foundation of Liaoning Province, China (No.201202031), State Education Ministry and The Research Fund for the Doctoral Program of Higher Education (No.20090041110002).

References

1. Efron, M.: Information search and retrieval in microblogs. Journal of the American Society for Information Science and Technology 62(6), 996–1008 (2011)
2. Cheong, M., Lee, V.: Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In: Proceeding of the 2nd ACM Workshop on Social web Search and Mining, pp. 1–8 (2009)
3. Dong, A., Zhang, R., Kolari, P., et al.: Time is of the essence: improving recency ranking using Twitter data. In: Proceedings of the 19th International Conference on World Wide Web, pp. 331–340 (2010)
4. Efron, M.: Hashtag Retrieval in a microblogging environment. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 787–788 (2010)
5. Evans, M., Chi, E.H.: Towards a model of understanding social search. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, pp. 485–494 (2008)
6. Geer, D.: Is It Really Time for Real-Time Search? Computer 43(3), 16–19 (2010)
7. Horowitz, D., Kamvar, S.D.: The anatomy of a large-scale social search engine. In: Proceedings of the 19th International Conference on World Wide Web, pp. 431–440 (2010)
8. Teevan, J., Ramage, D., Morris, M.R.: TwitterSearch: A Comparison of Microblog Search and Web Search. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 35–44 (2011)
9. Weng, J., Lim, E., Jiang, J., et al.: TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270 (2010)
10. Chen, C., Li, F., et al.: TI: An efficient indexing mechanism for real-time search on tweets. In: Proceedings of the 2011 International Conference on Management of Data, pp. 648–660 (2011)
11. Younus, A., Qureshi, M.A., Ghazi, A.N., et al.: Ins and Outs of News: Twitter as a Real-Time News Analysis Service. In: Proceedings of the Workshop on Visual Interfaces to the Social and Semantic Web (2011)
12. Massoudi, K., Tsagkias, M., Rijke, M.D., et al.: Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. In: The 33rd European Conference on Information Retrieval, pp. 362–367 (2011)
13. Nagmoti, R., Teredesai, A., Cock, M.D.: Ranking Approaches for Microblog Search. In: International Conference on Web Intelligence and Intelligent Agent Technology, pp. 153–157 (2010)
14. Meredith Ringel, M., Scott, C., Asta, R., Aaron, H., Julia, S.: Tweeting is Believing? Understanding Microblog Credibility Perceptions. In: The 12th Computer Supported Cooperative Work, pp. 441–450 (2012)