http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (18): 4534-4538, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4534.4538

Topic Thread Extraction from Search Results of Microblogs

[1]Lin Li, [1]Shi Qiao and [2]ShiliXiong
[1]School of Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070, Hubei
[2]Department of Advertising, UniversityofIllinoisatUrbana-Champaign, Urbana, USA

Abstract: Accompanied by the gradual integration of the microblogging in life, people would like to focus on the subjects they are interested in so that the technique of microblog retrieving becomes more and more popular. Currently, the engine of microblog retrieval uses list-view to show the retrieval results. Although considering the forwarding account and comment time it is still inconvenience for users to know the retrieval content in generally. This study puts forward the method that we organize the results according to the topic threads in order to improve the quality of microblog retrieval. Firstly, we made some earlier stage processing to the microblogs and then we gave out the Manual Sampling based Dynamic Incremental Clustering Algorithm. Finally we utilized the algorithm to extract the topic thread and show the result to the users. The data source is from Sina Weibo and it contains 14 topics and 74662 microblogs in total. The results show that, the algorithm is effective. Compared with the traditional k-means clustering it is 5 times faster and its precision is near.

Key words: Topic thread, microblogs, microblog retrieval, list view, clustering

INTRODUCTION

Microblogging service is a platform that users can share, spread and obtain information based on the relationships. Users can build his/her own individual community through Web, wap or other clients. One microblog contains at most 140 Chinese characters, once we send it out, the message would be shared to others. In August 2009, the first mainstream microblog, Sina Weibo, was opened up. By the end of the first half of this year, the user of the Sina Weibo had been 536 million. The speed of development of Chinese microblog has exceeded its ancestor Twitter which created Chinese style Internet myth. As the users are increasing, there are large amount of microblogs every second. People hope to find the topics they are interested in through the platform, therefore, the technique of microblog retrieval has highlighted its importance.

At present, the mainstream of the microblogging platform provides some basic retrieval services. Twitter's search engine displays the microblogs in list according to the sorting result of posting time and popularity. Sina Weibo and Tencent Weibo also provide some similar search function like Twitter. Overall the sorting algorithm of the retrieval functions in the mainstream microblogging platform considers forwarding number, comments, posting time, etc. However it lacks some deep analysis in semantic and sociality. Therefore, there result presentation has no regard for its hidden structure. Here, we come up with a word "topic thread" which means a series posted messages in a given topic. We use the topic thread extraction method to get several topics and it is convenient for users to get the topic they care about.

In this study, in order to deal with the textual data of microblog, we first did some preliminary work, such as segmented words, built term index and wiped out stop words and so on. And then we took the parameter selection in clustering into account. Finally, we utilized our algorithm to extract the topic threads. Through the experiment, we get the topic threads and correspondent microblogs.

The contribution of this study:

- We put forward the idea that organizes retrieval results according to the topic threads in order to improve the quality of microblog retrieval
- We give out the manual selected based dynamic incremental clustering algorithm
- The algorithm is much faster than the traditional clustering algorithm and the precision is comparable.

The study is organized as follows. We list several related work in the section 2. Then, in section 3, we describe the problem and give out our topic thread extraction method. Section 4 presents the data and the evaluation method and we also analyze our experimental

Corresponding Author: Lin Li, School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, 430070

results in this part. At last, we draw our conclusion and give out some topics for future research.

RELATED WORK

In recent years with the continuous development of social network, research on the microblog processing technology becomes hot. Compared with the traditional text retrieval in webpage, the messages in microblog have their own characteristics. A lot of research uses text mining technology to deal with microblog message. For example, Pervin *et al.* (2013) describe a method and implementation for extracting trending topics from a high velocity real-time stream of microblog posts. Hu *et al.* (2013) present a mathematical optimization formulation that incorporates the sentiment consistency and emotional contagion theories into the super-vised learning process; and utilize sparse learning to tackle noisy texts in microblogging. Lin *et al.* (2012) explore the problem of generating storylines from microblogs for user input queries. Efron *et al.* (2012) present experimental results using a corpus of microblog (Twitter) data and a corpus of metadata records from a federated digital library. But they have not yet been fully considered the relationship between attributes, such as time, content, phenomenon, location, user's social relations, etc.

The retrieval technology on BBS and blog platform has considered the structure connection between comments. Xi *et al.* (2004) explore the problem of creating an effective ranking function to predict the most relevant messages to queries in community search. Elsas and Carbonell (2009) compare methods that utilize thread structure to a naive method that treats a thread as a single document. Sun *et al.* (2008) study the popular queries collected over one year period and compare their search results returned by a blog search engine (i.e., Technorati) and a news search engine (i.e., Google News). Smith *et al.* (2000) discuss Threaded Text Chat, a program designed to address some of the deficiencies of current chat programs. In addition the quantity of the microblog is large and the text is very short. It's difficult to do the semantic analysis, so we need some unique methods on microblogs. Because of these problems, Qureshi *et al.* (2012) propose a technique for extracting important key terms/phrases in a considered topical domain. Vitale *et al.* (2012) waive the common practice of expanding the feature-space with new dimensions derived either from explicit or from latent semantic analysis. For long text, list view is a much more appropriate scheme. However, our object of study is short-text microblog which contains at most 140 Chinese characters and it is not convenient to get information for users. The contribution of this study is a novel approach to recovering thread structure in discussion forums where this explicit meta-data is missing (Wang *et al.*, 2008).

Topic thread is a structured information organization. The research in extracting structured information was focus on the application fields like BBS, blog and instant messaging (Seo *et al.*, 2009; Qamra *et al.*, 2006; Shen *et al.*, 2006) and the related research is rare in the field of microblogging information retrieve (Luo *et al.*, 2012). Luo *et al.* improved the quality of retrieving through mining the structure of the microblog's text. In this study, we focus on the expression of the results and we extract topic threads from the results. Therefore the users can understand the topic they searched in different points of view which helps users to know the development of the topic.

OUR APPROACH

In this section, we describe the problem we faced and give out an algorithm of topic thread extraction. We discuss some issues at last.

Problem description: In China, the user of the microblog has been more than 0.5 billion nowadays, we create huge amount of data at all times and the topics change very fast in a day. However, the shortcoming of the display for the retrieve is that we need to scan the results one by one. This kind of display (list-view) is acceptable for the traditional webpage, because the user can see many contents in the web. Microblog is very short and it is not easy to know the concrete content reading one by one. So we present a theory that we organize the results according to the topic threads. Users can realize several topics belonging to the retrieve results on the whole, as well as its main content. If the user interested in one topic, he/she can enter that topic to read one by one.

Manual sampling based dynamic incremental clustering algorithm: We propose a topic thread extraction algorithm for the search results of microblog retrieval, so called Manual Sampling based Dynamic Incremental Clustering Algorithm (MS-DICA). Its details are described as follow.

Input:

- k: No. of clusters
- l: No. of the samples (In this study we select 10 samples in each topic thread, 140 samples in total)
- D: Chinese microblog dataset which contains N microblogs
- I: Dataset D after preliminary data processing

- M: The samples

Output:

- k topic threads

Method:

- Get I through preliminary data processing such as segment D, build term index, filter stop words, calculate TF and TF*IDF
- Select M from I manually
- Use k-means clustering on M, get the central points
- For every item in I, calculate the distance with every central points, appoint the item to the cluster which nearer to it
- Get the term index that in top 5 of the TF in the cluster, they are the labels of the topic thread

The basic idea of the whole algorithm is that we first use the k-means cluster algorithm on the samplings which selected manually to get the central point. Then, we calculate the distance between every microblog and each central point to decide the cluster it belongs. Compared with random sampling, manual sampling guarantees the quality in clustering. The microblog is a data stream which grows rapidly along with time, so the method that calculates distance item by item makes quick cluster possible and the results show that the precision is acceptable.

DISCUSSIONS

As the microblog is very short and only has 140 Chinese characters, we just considered the textual feature and we did not expand the other features in the text, we would do the research in the future. Moreover, the formula we used in k-means cluster and algorithm (4) were all Euclidean distance, other distance formula, such as Cosine, had a lot of error.

EXPERIMENTS

In this part, firstly we introduce the source of the data and the way we deal with it. Secondly we give out the evaluation method of the precision. At last, we display the results and make some analysis on them.

Data and its processing method: The data we use in our study was crawled from Sina Microblog which was from the end of March 2012 to the end of June 2012. There were 14 topics and 74662 tweets in total and all of them were IT/technology topic. We used the Chinese words segmentation tools (MyTxtSegTag) to segment these tweets and then we utilized Chinese words frequency statistics tools (MyZiCiFreq) and stop words table to get 13167 term index. After that, we figured out the term frequency and term frequency*inverse document frequency of every tweets. In the end, we selected 10 samples from every topic (140 samples in total) and we used these samples to get the cluster points.

Evaluation method: The precision we talk is whether the tweets are correctly divided into corresponding topic threads under our algorithm. So our evaluation criterion is as follows:

$$\text{Precision} = \frac{\sum_{Ci}}{N} \quad (1)$$

Equation 1 represents the tweet that is correctly divided into its topic, N means the total number included in one topic thread.

In order to better illustrate our algorithm, we use average precision to evaluate the algorithm, as defined in Eq. 2:

$$\text{Average precision} = \frac{\text{Precision}}{M} \quad (2)$$

In Equation 2 M means the total number of all the topic thread.

RESULTS AND DISCUSSION

What we need to consider in the experiment are K and term weight selection when we use MS-DICA clustering and how to confirm the label of the topic thread after the MS-DICA clustering.

Hard clustering and soft clustering: Our data source comes from 14 topic threads. In order to calculate the precision, we need to know the topic of the clusters. We first get the label of each cluster from the central points, then, we use the equation 1 to calculate the accuracy of each topic.

One cluster may contain multi-topics and two or more clusters may also belong to one topic in this method. Soft Clustering presents above two cases, however, hard clustering only can include the latter one.

Here, we took "K=18, Term Weight= TF" in consideration.

When we use soft clustering strategy, the average precision is higher than the hard clustering (46.35% vs

16.16%). For this case, we believe it is reasonable. Because in some cases, several topics contain the same key word, for example "iPad", cluster1 and cluster3 both have this key word, so we think they are different subtopics in one central topic.

Parameter study of k and term weighting: Our data contains 14 topic thread, so we set the value range of K between 10~20. Here, we only give several K values in our table. We also considered the term weight and we display them in the table below.

We can see from the Table 1, the average precision comes to a peak when K value is 18 and the number is 46.63%. It isn't equal to the number of the topic thread because there are several noisy points. Therefore, K value is a little bigger than the number of topic thread. But the K can't be too large, if so it will take too much time and the precision will fall.

On the other hand, the TF term weight is better than TF*IDF term weight in each K value. So the TF term weight is more ideal than TF*IDF term weight in extraction of the microblog's topic thread.

It is the most ideal when K=18 and term weight is TF in our experiment.

Clustering labelling: Labels are the key words of a topic thread which can represent the meaning of the cluster exactly. For instance, in Table 2, we took the top 5 TF of the term index of cluster1 and cluster11. We can see from cluster1, the labels are "HTC", "banned", "reasonable", "USA", "China" which we can infer the cluster1's topic thread is "#HTC is banned in USA.Is it reasonable? #". It is the same for the cluster11.

Table 1: Parameter select and its precision, soft clustering.

K value	Term weight	Average precision (%)
10	TF	27.99
	TF*IDF	32.71
15	TF	41.37
	TF*IDF	26.97
18	TF	46.35
	TF*IDF	35.57
20	TF	42.64
	TF*IDF	33.29

Table 2: Top 5 TF of the term index of cluster1 and cluster11

Cluster1	Cluster11	Label
0.0245	0.0087	HTC
0.0197	0.0072	Banned
0.0197	0.0070	Reasonable
0.0197	0.0081	USA
0.0088		China
	0.0058	Prefer

Table 3: Compares with k-means clustering

Algorithm	Run time	Precision (%)
K-means	30 min	23.32
MS-DICA	6 min	20.26

Comparisons with K-means clustering: Here in Table 3, we took 1500 tweets as an example, k-means clustering need 30 min, however, MS-DICA clustering only took 6 min. We can say that MS-DICA has a great improvement in time consuming on microblog data. On the other hand, the precision did not descend too much. The most important point is that when the data becomes huge, k-means clustering can't accomplish the job. But our MS-DICA can add tweet and topic thread dynamically which only need select the sample and calculate the central point again.

CONCLUSION AND FUTURE WORK

In this study, we first analyzed the current problems in microblog retrieving. To solve these problems, we put forward the MS-DICA to extract the topic threads in microblog. Then, for this algorithm, we considered several variables in our experiment. Finally, we completed the experiment and compared with the traditional k-means algorithm in precision and velocity. Considering the final results, our algorithm has an advantage in velocity and the accuracy is very close contrasted to the k-means clustering. Due to the number of the microblog is growing fast, so the MS-DICA is very meaningful. In our future study, we will extend other features in the text, such as time, semantics and sociality, etc.

ACKNOWLEDGMENT

This research was undertaken as part of Project 61003130 funded by National Natural Science Foundation of China and Project 135210001 supported by the Fundamental Research Funds for the Central Universities.

REFERENCES

Efron, M., P. Organisciak and K. Fenlon, 2012. Improving retrieval of short texts through document expansion. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, August 12-16, 2012, Portland, OR., USA., pp: 911-920.

Elsas, J.L. and J.G. Carbonell, 2009. It pays to be picky: An evaluation of thread retrieval in online forums. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2009, Boston, MA., USA., pp: 714-715.

Hu, X., L. Tang, J.L. Tang and H. Liu, 2013. Exploiting social relations for sentiment analysis in microblogging. Proceedings of the 6th ACM International Conference on Web Search and Data, February 4-8, 2013, Rome italy, pp: 537-546.

Lin, C., C. Lin, J.X. Li, D.D. Wang, Y. Chen and T. Li, 2012. Generating event storylines from microblogs. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, October 29-November 2, 2012, Maui, HI., USA., pp: 175-184.

Luo, Z.C., M. Osborne, S. Petrovic and T. Wang, 2012. Improving twitter retrieval by exploiting structural information. Proceedings of the 26th AAAI Conference on Artificial Intelligence, July 14-18, 2012, Bellevue, USA.

Pervin, N., F. Fang, A. Datta, K. Dutta and D.E. Vandermeer, 2013. Fast, scalable and context-sensitive detection of trending topics in microblog post streams. ACM Trans. Manage. Inf. Syst., Vol. 3 10.1145/2407740.2407743

Qamra, A., B. Tseng and E.Y. Chang, 2006. Mining blog stories using community-based and temporal clustering. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, November 5-11, 2006, Arlington, VA., USA., pp: 58-67.

Qureshi, M.A., C. O'Riordan and G. Pasi, 2012. Short-text domain specific key terms/phrases extraction using an n-gram model with wikipedia. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, October 29-November 2, 2012, Maui, HI., USA., pp: 2515-2518.

Seo, J.W., W.B. Croft and D.A. Smith, 2009. Online community search using thread structure. Proceedings of the 18th ACM Conference on Information and Knowledge Management, November 2-6, 2009, Hong Kong, China, pp: 1907-1910.

Shen, D., Q. Yang, J.T. Sun and Z. Chen, 2006. Thread detection in dynamic text message streams. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006, Seattle, Washington, USA., pp: 35-42.

Smith, M., J.J. Cadiz and B. Burkhalter, 2000. Conversation trees and threaded chats. Proceedings of the ACM Conference on Computer Supported Cooperative Work, December 2-6, 2000, Philadelphia, PA., USA., pp: 97-105.

Sun, A., M. Hu and E.P. Lim, 2008. Searching blogs and news: A study on popular queries. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, 2008, Singapore, pp: 729-730.

Vitale, D., P. Ferragina and U. Scaiella, 2012. Classification of Short Texts by Deploying Topical Annotations. In: Advances in Information Retrieval, Baeza-Yates, R., A.P. de Vries, H. Zaragoza, B.B. Cambazoglu, V. Murdock, R. Lempel, F. Silvestri (Eds.). Springer-Verlag, Berlin, Germany, pp: 376-387.

Wang, Y.C., M. Joshi, W.W. Cohen and C.P. Rose, 2008. Recovering implicit thread structure in newsgroup style conversations. Proceedings of the 2nd International Conference on Weblogs and Social Media, March 30-April 2, 2008, Seattle, USA.

Xi, W., J. Lind and E. Brill, 2004. Learning effective ranking functions for newsgroup search. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, Sheffield, UK., pp: 394-401.