Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 2008 – 2011

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Mining Hot Topics from Twitter Streams

Jing Guo[a,b,1,*], Peng Zhang[b,*], Jianlong Tan[b,*], Li Guo[b,*]

[a]*School of Computer Science, Beijing University of Posts and Telecommunications*
[b]*Institute of Information Engineering, Chinese Academy of Sciences*

Abstract

Mining hot topics from twitter streams has attracted a lot of attention in recent years. Traditional hot topic mining from Internet Web pages were mainly based on text clustering. However, compared to the texts in Web pages, twitter texts are relatively short with sparse attributes. Moreover, twitter data often increase rapidly with fast spreading speed, which poses great challenge to existing topic mining models. To this end, we propose, in this paper, a flexible stream mining approach for hot twitter topic detection. Specifically, we propose to use the Frequent Pattern stream mining algorithm (i.e. FP-stream) to detect hot topics from twitter streams. Empirical studies on real world twitter data demonstrate the utility of the proposed method.

Keywords: Data stream mining, Hot topic mining, Frequent pattern mining, Twitter streams

1. Introduction

As an important platform for message sharing and view dissemination, there are unprecedentedly large and complex data on the Internet. How to extract meaningful public opinion from the massive Internet data, without being submerged in the knowledge ocean, is an open challenge. To address this challenge, many public opinion analysis models have been proposed to extract important opinions from the Internet [1-4]. For example, the TDT (Topic Detection and Tracking) method combines both NLP and data mining to automatically recognize and continuously track important opinions from text streams [1], the SDA method [2] in the social science computing combines the sociology and data mining to automatically search, analyze and report the public opinions based on the Web page data.

Hot topic mining represents one of the most important research areas in the public opinion analysis. The key technology includes the text classification and clustering, topic detection and tracking, opinion tendency identification, and multi-document automatic summarization. The fundamental modules in a hot topic mining system are illustrated in Fig. 1.

Now we will explain the system framework given in Fig. 1. Obviously, the *data acquisition* module plays a fundamental role in the system. It continuously downloads web pages from the Internet, which are then fed into the *Web text preprocess* module for data cleansing. Then, these data will be converted into numeric vectors in the next *text vectorization* module. After that, all the text data will be categorized into predefined classes in the *text clustering*

*Email addresses: guojing@software.ict.ac.cn, {zhangpeng, tjl, guoli}@ict.ac.cn

[1]This research was supported by the National Science Foundation of China (NSFC) under Grant No. 61003167, National High Technology Research and Development Program 863 under Grant No. 2011AA01A103.

doi:10.1016/j.procs.2012.04.224

module, where each class represents a topic. Note that different clustering algorithms will yield different topics. The *hot topic evaluation* module ranks the above topics by comprehensively studying various parameters, such as the report number, comment number, click number, to name a few. At the last step, the *hot topic output* module generates the above hot topics from different viewpoints.

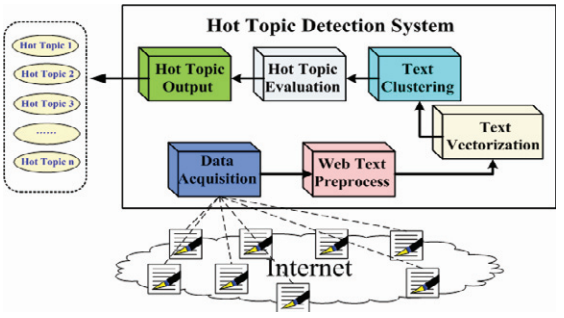

Figure 1: An illustration of the hot topic detection system

Twitter texts are very different from the generic Internet Web page texts. Twitter texts are often very short and sparse, spreading rapidly. These new characteristics of twitter texts pose great challenge to the topic mining efficiency, resource consumption and mining accuracy. To address these problems, we propose, in this paper, a flexible stream mining approach for hot twitter topic detection. Specifically, we propose to use the Frequent Pattern stream mining algorithm (i.e. FP-stream) to detect hot topics from twitter streams. Technically, we take each twitter text as a transaction, and each word in a transaction as an item. Thus, detecting the twitter hot topics is tantamount to mining frequent patterns from the twitter streams. The mining results (i.e. the frequent word-sets) are the hot twitter topics. Empirical studies on real world twitter data demonstrate the utility of the proposed method.

2. Mining Hot Topics from Twitter Streams

2.1. Hot Twitter Topic Detection Approach

Generally, a hot topic is defined as a popular or frequent twitter topic in a given time interval. Fig. 2 describes the key components for hot twitter topic detection system, among which the frequent twitter pattern mining module is the most important part. After processing the twitter texts in the data acquisition module and the twitter text preprocess module, each twitter text can be taken as a transaction, while each single word in a twitter text can be taken as an item. This way, detecting the twitter topics is tantamount to mining stream frequent patterns from the twitter data. Consequently, the results of the frequent word-sets are the hot twitter topics. The hot twitter topics will be displayed in the topic output module for better understanding. Stream frequent pattern mining is an important research branch in data mining research community [5-10]. Representative work includes: Lossy Counting algorithm [5] and FP-Stream algorithm [6]. Specifically, Lossy Counting algorithm [5] cannot distinguish impacts between the old and new transactions. While FP-Stream algorithm can obtain time-sensitive results. Indeed, distinguishing the old and new transactions plays a key role in hot twitter topic detection. Therefore, we use the FP-Stream algorithm as the basic mining algorithm for hot twitter topic detection.

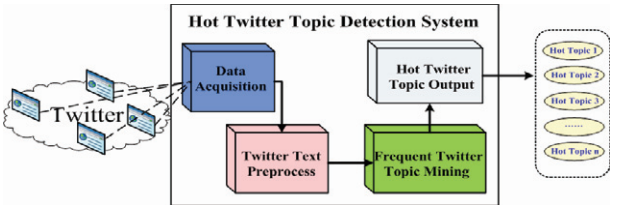

Figure 2: Hot Twitter Topic Detection System

2.2. The FP-Stream Algorithm

There already exists work to show that FP-Stream algorithm is capable of maintaining time-sensitive frequent patterns in data stream environments in limited main memory. However, it cannot be used to our problem directly because of two reasons.

First, the updating of the FP-Stream structure is based on data chunks. It works only when enough incoming transactions have arrived to form a new data chunk and the number of each data chunk share the same chunk size. However, in reality, delaying the mining process until a given number of data is obtained is impractical, as people tending to analyze public opinion based on the time stamps, but not on the batch number. For example, we are likely to ask what the hot twitter topics in the past hour are, instead of what the hot twitter topic in the past 100 twitters are.

Therefore, we revise the FP-Stream structure for a better practical application. The improved FP-Stream structure is shown in Fig.3. The major difference between the improved and original FP-Stream structure is the tilted-time window table embedded. The tilted-time window table in improved FP-Stream structure is extended and consists of three different parts. The tilt-window records time-point information in a stream, sup_num records the corresponding item-set number in corresponding time, and batch_num records the batch number in corresponding time. This structure allows FP-Stream algorithm to proceed twitter streams batch by batch, and the batch number can be changing with time, which makes the FP-Stream algorithm more practical and flexible.

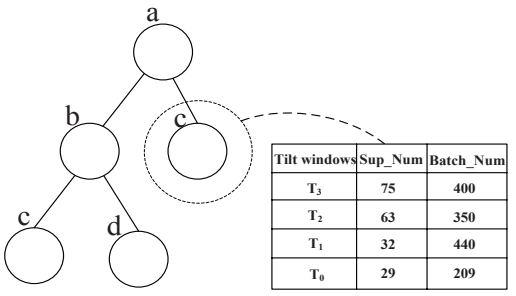

Figure 3: Hot Twitter Topic Detection System

Second, we all know that it is unrealistic to hold all streaming data in the limited main memory. Thus, FP-Stream algorithm divides patterns into three categories: frequent patterns, sub-frequent patterns, and infrequent patterns. Let the support threshold be σ, the error bound be ϵ. A pattern is frequent if its support is no less than σ. It is sub-frequent if it s support is less than σ, but no less than ϵ. Otherwise, it is infrequent. The frequent and sub-frequent patterns are the interest patterns in a stream.

However, as we mentioned above, twitter data is quite different from the traditional Internet Web page data. The twitter data is of short content and spreading rapidly. FP-Stream algorithm is not well addressed yet in terms of its bounds of memory consumption. Hence, we have to develop FP-Stream algorithm for effectively mining frequent patters from rapidly spreading twitter stream. In this paper, we extend the definition of the three pattern categories. A pattern is frequent if its support is more than σ. It is sub-frequent if it s support is less than σ but no less than $(\sigma - \epsilon)$. Otherwise, it is infrequent. The frequent and sub-frequent patterns are what we are interested in. This way, the developed FP-Stream algorithm will have high accuracy, require less memory, and consume less CPU time[10] for mining in a twitter stream.

Generally, we are more interested in the latest information in the public opinion analysis tasks, as news is time-sensitive. In order to discover practical and sensitive hot topics, we will design certain damping-factor for sup_num in the new FP-stream algorithm. By doing so, we can quartette that the latest information has the highest significant impact in the mining results.

3. Experiments

In our experiments, we use the twitter dataset [10] for testing. The dataset contains twitter short texts regarding the swine flu topic between April 22 and October 13, 2009. Our experiments are based on the texts collected from April 26 to May 3, 2009.

After twitter data preprocessing, cleaning and stop-word removing, all the twitter texts can be transformed to well-formatted transactions. We treat twitter texts generated in one day as a batch data. Note that the size of each batch may be different. By setting the support threshold to be 0.03, and the error bound value to be 0.001, we use the FP-Stream algorithm to mine frequent patterns in twitter dataset. The frequent patterns is the hot twitter topics. The mining results are shown as follows.

From the results in Table.1, we can come to the conclusion that the developed FP-Stream algorithm is a practical and effective approach for hot twitter topic detection.

Performance Study w.r.t. Time Consumption. Time consumption is an important measure for stream mining. The corresponding results of the proposed algorithm are shown in Fig.4. We can observe that the time consumption of the proposed algorithm is in the magnitude of second, which reduces with the decreasing support threshold. Therefore, we can safely say that the time consumption of the proposed algorithm scales well with the support threshold value.

Performance Study w.r.t Memory Consumption. Memory consumption is another important measure for stream mining. The results w.r.t. different memory consumptions are shown in Fig.5. We can observe that the number of nodes we reduced in the proposed algorithm decreases with the support threshold. Therefore, we can safely say that our method scales well to the support threshold.

Table 1: Mining Results With Different Time

Mining Result	Time	Mining Result	Time	Mining Result	Time
Flu	4.26-5.3	Flu Confirmed	5.1-5.3	Flu Symtoms	5.3
Swine	4.26-5.3	Swine confirmed	5.1-5.3	Swine Symtoms	5.3
Swine Flu	4.26-5.3	Swine Flu Confirmed	5.1-5.3	Swine Hysteri	5.3
Mexico	4.30-5.3	Pandemic	5.3	Flu Hysteri	5.3
Cases	4.30-5.3	pigs	5.3	Swine Flu Pandemic	5.3
Flu Mexico	4.30-5.3	Symtoms	5.3	Swine Flu pigs	5.3
Swine Mexico	4.30-5.3	Hysteri	5.3	Swine Flu Symtoms	5.3
Swine Cases	4.30-5.3	Swine Pandemic	5.3	Swine Flu Hysteri	5.3
Flu Cases	4.30-5.3	Flu Pandemic	5.3	Flu Confirmed Cases	5.3
Swine Flu Cases	4.30-5.3	Flu pigs	5.3	Swine Confirmed Cases	5.3
Confirmed	5.1-5.3	Swine pigs	5.3	Swine Flu Confirmed Cases	5.3

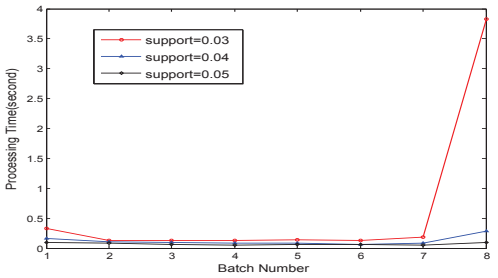

Figure 4: Performance Study w.r.t. Time Consumption

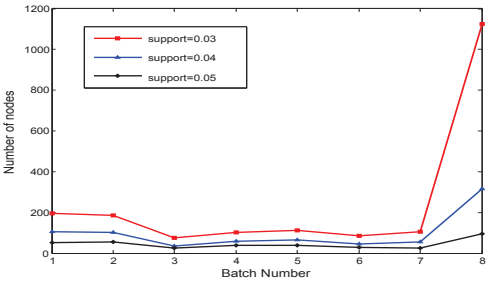

Figure 5: Performance Study w.r.t. Memory Usage

4. Conclusions

Hot topic detection from twitter streams is an challenging research problem in data stream mining community. Traditional hot topic detection approaches from Internet Web page data are based on clustering algorithms. However, twitter data is quite different from Web page data because twitter data (text) is often very short, sparse, and spreading rapidly. To mine hot topics from twitter streams, we propose a more flexible and practical approach based on data stream frequent pattern mining. Specifically, a new stream frequent pattern mining algorithm is proposed based on the FP-stream algorithm. Experiments have demonstrated its effectiveness in mining the real-world twitter data.

References

[1] J.Allan,J.Carbonell, G.Doddington,et al. Topic Detection and Tracking Polot Study Final Report. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.(1998)
[2] http://sda.berkeley.edu/
[3] NesstarUnlocking data-Creating knowledge[EBOL] http://www.nesstar.com/
[4] information available to the public[EBOL] http://www.freepatentsonline.com/4930077.html
[5] G. Manku and R. Motwani. Approximate frequency counts over data streams. In Proc. of VLDB. (2002)
[6] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu. Mining Frequent Patterns in Data Streams at Multiple Time Granularities. AAAI/MIT. (2003)
[7] M. Charikar, K. Chen, and M. Colton. Finding frequent items in data streams. Theoretical Computer Science, 312, 3–15 (2004)
[8] P. Zhang, J. Li, P. Wang, B. Gao, X. Zhu, and L. Guo. Enabling fast prediction for ensemble models on data streams. In Proc. of KDD. (2011)
[9] P. Zhang, X.Zhu, Y. Shi, L.Guo, and X. Wu. Robust ensemble learning for mining noisy data streams. Decision Support Systems, 50, 469–479 (2011).
[10] J. Guo, P. Zhang, J. Tan, and L. Guo. Mining Frequent Patterns across Multiple Data Streams. In Proc. of CIKM. (2011).
[11] Yu, J.X, Chong, Z., Lu, H., Zhang, Z., Zhou, A.. False positive or false negative: mining frequent itemsets from high speed transactional data streams. In: Proc, VLDB, 204–215 (2004).
[12] http://www.datatang.com/datares/go.aspx?dataid=601994
[13] P. Wang, P. Zhang, and L. Guo, Minining Multi-label Data Streams Using Ensemble-based Active Learning, In Proc. Of SIAM SDM 2012