Mining micro-blogging users' interest features via fingerprint generation

Dong Liu, Quanyuan Wu, Weihong Han
School of Computer
National University of Defense Technology
Changsha, China
nudtld@yahoo.cn

***Abstract*—Nowadays, micro-blogging is widely used as a communication and information sharing social network service, therefore mining micro-blogging users' behavior features is very important both in the economic and social fields. A framework for the analysis of user's interest features is proposed in this paper. After data cleaning, word segmentation, POS(part of speech) filtering and synonym merging, the keywords that called terms of all the tweets posted by a typical user in 2011 are extracted. Then VSM(vector space model) is used to generate the feature vector of the tweets from these terms. Furthermore, a *k*-bit binary called *fingerprint* is generated from the high dimensional feature vector of the tweets by use of Simhash algorithm. The micro-blogging user's interest features and change patterns could be detected by analyzing the fingerprint sequences and the distance between the adjacent two fingerprints. Taking Sina micro-blogging as background, a series of experiments are done to prove the effectiveness of the algorithms.**

Keywords-micro-blogging;interest feature; tweet; fingerprint

I. INTRODUCTION

As a communication and information sharing social network service, micro-blogging has spread rapidly over the world. Chinese micro-blogging platforms have developed extremely rapidly as well. By the end of November 2011, the number of Chinese micro-blogging accounts has grown to 320 million and the number of posted messages which are called *tweets* reached over 150 million each day [1]. Along with the overall popularization of the Internet as well as the continuous development of the mobile communication network, the user coverage of micro-blogging will grow much larger.

Mining users' features plays an important role in the micro-blogging data mining, because it is very useful not only to conduct the patterns of users' behaviors, but also to provide them with personalized services, such as business recommendation. Besides, when analyzing public sentiment, the feature analysis could help to mine the behavior patterns and provide powerful support to the trace of sensitive topics and key users. One of the methods to mine users' features is to analyze their posted tweets. Different from traditional internet media, the features of micro-blogging bring severe challenges to the analysis. In micro-blogging, a tweet could be posted within the limit of 140 characters, and the contents can be anything that the users want to share, such as what they are doing or seeing, hot news, videos, audios, pictures, et al. Users could post and share tweets instantly through a variety of network clients such as computers, mobile phones, and so on. Consequently, there are many features of tweets such as short length, non-standard format, large quantity, that restrict the analysis by use of traditional methods. In this paper, we propose a framework to mine the micro-blogging users' interest features via the analysis of posted tweets.

The rest of the paper is organized as follows. Section II introduces the related work. Section III analyses the pre-processing module of tweets. Section IV makes analysis on generation of tweets feature vectors and user's interest fingerprint. Section V evaluates the proposed algorithms and proves their effectiveness. Section VI concludes this paper.

II. RELATED WORK

Micro-blogging, Along with its rapidly development, has been widely studied in many fields. Twitter, the most popular micro-blogging platform over the world, has been studied with regard to the topological characteristics [2], the information dissemination patterns [2] and so on. The researchers also compare it to traditional media and social networks [3]. Sina micro-blogging, the largest micro-blogging platform in China, also attracts many studies. Fan Pengyi et al. launched an active measurement of Sina micro-blogging and investigated on topological characteristics and user behavior features [4].

In micro-blogging user feature analysis, researchers study on measuring user influence [5], recommendations [6], Hashtag analysis [7] and et al. They also analyze the behavior features of users such as sentiment, interest and motivation, by use of text processing including feature extraction [8], part-of-speech tagging [9], named entity recognition [10].

The present studies of user feature analysis mostly base on the user groups and emphasis on finding the relationships among users, such as "following", "retweeting", "recommending" and so on. In this paper, from the individual's point of view, we analyze all the tweets posted by a typical user and propose fingerprint to represent his interest features. By analyzing the fingerprint sequences, we finally mine the user's interest features and change patterns.

III. PRE-PROCESSING OF TWEETS

A. Data cleaning

When mining micro-blogging user's interest features, we mainly analyze the text contents of the posted tweets and don't concern the multimedia information such as pictures, videos and so on. Therefore, we first delete all the

Published by Atlantis Press, Paris, France.

multimedia information of the tweets. Furthermore, the user identifiers of retweeted tweets, emotion icons, URLs, et al. are not concerned in this paper when analyzing user's interest features, so they also need to be removed.

B. Word segmentation and POS filtering

After data cleaning, we perform word segmentation and POS(part of speech) tagging for the tweets by use of ICTCLAS [11] (Institute of Computing Technology, Chinese Lexical Analysis System).

The key words which represent the user's interest are extracted, and their POS usually are noun(n), verb (v), idiom(i), and the proper noun could be subdivided into location name(nl), institute name(ni), place name(ns), person name(np) and time(nt). Therefore, we filter the words according to their POS, and only retain words with the above POS. The frequencies of the retain words are calculated at last.

C. Synonym merging

We build a synonym lexicon based on the "Synonym Lexicon-expanded version" [12] of HIT-SCIR(Harbin Institute of Technology-Social Computing and Information Retrieval), and then make the synonym merging to the tweet keywords . The synonym merging is to filter the synonyms according to the lexicon and kept the frequent one as the feature keyword of the tweet. The frequency of the keyword is the sum of its synonyms' frequencies.

D. Example

Take a tweet for example; the pre-processing procedure is illustrated as follows.

The original tweet is: "北京这几天的天气实在不好[可怜]，沙尘气候得防治，这天气太影响身体了[抓狂]。//@旋转音符：北京出现今年最强沙尘天气，http://v.youku.com/v_show/id_XMjYzMTk4NzUy.html";

After removing the user identifier(//@旋转音符:), emotion icons([可怜] [抓狂]), URL, the cleaned tweet data is: "北京这几天的天气实在不好，沙尘气候得防治，这天气太影响身体了。北京出现今年最强沙尘天气";

Then, After word segmentation and POS filtering: "北京/ns 天气/n 沙尘/n 气候/n 防治/v 天气/n 影响/v 身体/n 北京/ns 出现/v 今年/nt 沙尘/n 天气/n";

The tweet keywords and frequencies: "北京 2，天气 3，沙尘 2，气候 1，防治 1，影响 1，身体 1，出现 1，今年 1";

"天气" and "气候" is considered to be synonyms. The two words need to be merged，and "天气" which is more frequent is kept as the keyword.

The final result: "北京 2，天气 4，沙尘 2，防治 1，影响 1，身体 1，出现 1，今年 1".

IV. Generation of Tweet Set's Feature Vector and User's Interest Fingerprint

A. Generation of tweet set's feature vector

In this paper, we use VSM(vector space model) to represent the features of the processed tweets data.

Without considering other factors such like sudden incidents, we can assume that there exist relationships among the posted tweets. Therefore, we make aggregational statistics analyses about the tweets based on a certain time granularity (such as 1 hour or 1 day). At last, we finally get a tweet set sequence (D_1, D_2,…, D_m ,…), D_m is a tweet set formed by the tweets posted in the i[th] time-granularity period. After data pre-processing, D_m was separate into several keywords which are called *terms*. After statistic of these terms, we finally got an *n*-dimensional vector ($<t_1,tf_1>,<t_2,tf_2>,\cdots,<t_n,tf_n>$), t_i ($i\in[1,n]$) is the i[th] term and it is different from any other term. tf_i is the frequency of the t_i, and *n* is the total number of terms. We define some related concepts as follows.

Definition 1: *Average Frequency AF* means the average of the terms' frequency:

$$AF = \sum_{i=1}^{k} tf_i \Big/ k \tag{1}$$

k is the number of tf_i which are different from each other.

Definition 2: *Term frequency weighting coefficient* $\Phi(tf_i)$ represents the importance of the term t_i whose frequency is tf_i in D_i. It is obviously that the frequent terms represent the tweet features mostly and the non-frequent ones are almost useless. $\Phi(tf_i)$ is defined as follow.

$$\Phi(tf_i) = \frac{1}{1+2^{(AF-tf_i)}} \tag{2}$$

Finally, we define the weight of term t_i as follow:

$$w_i = \Phi(tf_i) \times tf_i \tag{3}$$

The tweet set D_m is represented as vector ($<t_1,w_1>,<t_2,w_2>,\ldots,<t_n,w_n>$). w_i is the weight of t_i, and it represent the importance of t_i in the tweet set.

B. Generation of user's interest fingerprint

Because the feature vector of tweet set usually has large dimensions, it needs large storage space and is not appropriate for real-time data processing as well. Therefore, the dimension reduction is essential. Based on the Simhash algorithm [13] proposed by Charika, the high-dimensional feature vector of tweet sets is transformed into a *k*-bit vector.

The algorithm is show in Table I. The *n*-dimensional tweet set feature vector is represented as ($<t_1,w_1>,<t_2,w_2>,\ldots,<t_n,w_n>$). Firstly, we initialize a *k*–dimensional vector *V* with each dimension as zero (line 1). Then, for each t_i, it is hashed into a *k*−bit hash value. These *k* bits increment or decrement the *k* components of the vector *V* by the weight w_i based on the value of each bit of the hash value calculated. If the i[th] bit of hash value is 1, the i[th] component of *V* adds w_i; otherwise, if the i[th] bit is 0, the i[th] element of *V* deducts w_i (line4-8). The vector *V* is called *interest fingerprint vector*. Finally, a *k*-bit binary *S* is

generated from V (line 10-12), and it is called *interest fingerprint*.

TABLE I. THE ALGORITHM OF INTEREST FINGERPRINT GENERATION

Input: tweet set feature vector $(<t_1,w_1>,<t_2,w_2>,\ldots,<t_n,w_n>)$
Output: interest fingerprint vector $V(v_1,v_2,\ldots,v_k)$,
interest fingerprint $S(s_k....s_2s_1)$

1. $V(v_1,v_2,\ldots,v_k)$ initialized to 0
2. for each t_i
3. token_hash = k-bit_hash(t_i);
4. for i =1 to k do
5. if (i^{th} bit of token_hash == 1)
6. $v_i = v_i + w_i$
7. else
8. $v_i = v_i - w_i$
9. $S(s_k....s_2s_1)$ initialed to 0
10. for i =1 to k
11. if ($v_i > 0$)
12. $s_i = 1$
13. return V, S

User's interests could be concluded from his posted tweets which are represented by interest fingerprint vectors, so the interest fingerprint could represent the user's interest concisely. Hamming distance is used to represent the difference of two fingerprints, and it also reflects the difference of the tweet sets features which represent the interests of user.

V. EXPERIMENT

Taking Sina micro-blogging as background, a verified user is chosen at first. This user always pays attention to many hot topics continuously. By use of OpenAPI [14], we collect all the 4236 tweets posted in 2011. There are 1132 original tweets and 2510 retweeted tweets which were commented by the user as well. Because most of the tweets include the user's own viewpoints, they are very useful when analyzing user's interest features.

A. Hamming distance between similar tweet set

As described in section IV, the terms whose frequencies are bigger than the average could represent the tweet set's features.

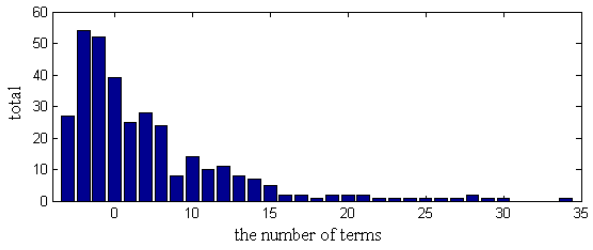

Figure 1. Distribution of the number of frequent terms

Fig. 1 shows the distribution of the number of terms whose frequencies are bigger than average frequency. Obviously, the number of frequent terms in a tweet set is mostly smaller than 15. Therefore, we construct 3 test sets each of which has 6 categories of test texts according to its similarity to the standard text. The similarities between test texts and the standard text are 100%, 80%, 60%, 40%, 20%, 0%. Texts of the three test sets are formed by 5, 10 and 15 terms respectively, and each term's frequency is set to 1. The Hash function of Simhash algorithm is 64-bit FNV. The experiment result is shown in Fig. 2. Test set 1, 2 and 3 represent the three test sets whose texts are formed by 5, 10 and 15 terms respectively.

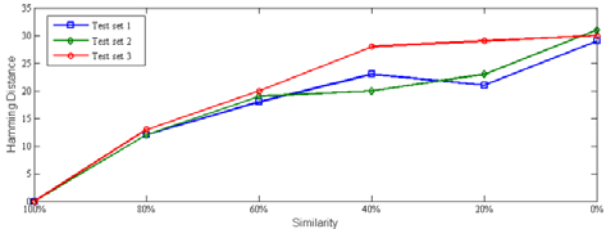

Figure 2. Hamming distance between the test set and standard set

We could conclude from the experiment that the tweet set's features are similar whose fingerprints' hamming distances are smaller than 15, and the tweet set's features are almost completely different if their fingerprints' hamming distances are larger than 25.

B. Analysis of user's interest fingerprint

We analyze the interest fingerprint in this section. According to the universal behavior of micro-blogging users, the time granularity for tweets aggregation is set to be 1 day. First of all, we select the tweets posted in three different days. After processing, we obtain all the eligible word terms as shown in Table II. The number after each term is its frequency.

TABLE II. THE TERMS AND FREQUENCIES

Date	Terms and their frequency (Top-10, sorted by frequencies)
2011-12-31	(方舟子 38), (打假 12), (中国 11), (抄袭 11), (文章 7), (网络 6), (菊花 5), (深圳 4), (发表 4), (教授 4)
2011-12-30	(方舟子 19), (打假 12), (菊花 7), (抄袭 7), (没有 6), (海宁 6), (中国 6), (上访 5), (农民工 4), (工资 4)
2011-10-31	(临沂 6), (警察 5), (可能 5), (做客 5), (没有 4), (摩托车 4), (诬陷 4) (肇事者 4), (媒体 4), (疏忽 3)

TABLE III. DISTANCE BETWEEN EACH TWO INTEREST FINGERPRINTS

	2011-12-31	2011-12-30	2011-10-31
2011-12-31	0		
2011-12-30	5	0	
2011-10-31	30	31	0

The hamming distance between each two interest fingerprints of tweet sets is shown in Table III. We can conclude from Table II and original tweets that the user mainly focused on the incidents "方舟子打假及抄袭事件/Fang Zhouzi's exposing of academic fraud and plagiarizing others' papers" in December 31, 2011. The use not only focused on the incidents about "Fang Zhouzi" but also paid attention to other incidents like "海宁农民工上访事件/petitioning of migrant workers in Haining" in December 30,

and he mainly focused on the incidents about "临沂警察/policemen of Linyi ", "摩托车肇事/Motorcycle accident" in October 31.

Because the user mainly focused on the incidents about "方舟子/Fang Zhouzi" in December 30 and 31, the interest fingerprints of tweet sets posted in those two days have close hamming distance which is smaller than 15. In October 31, the user completely focused on the other incidents, so the hamming distances of fingerprints between October 31 and December 30, 31 are larger than 25. Obviously, the tweet set fingerprints can represent the user everyday interest.

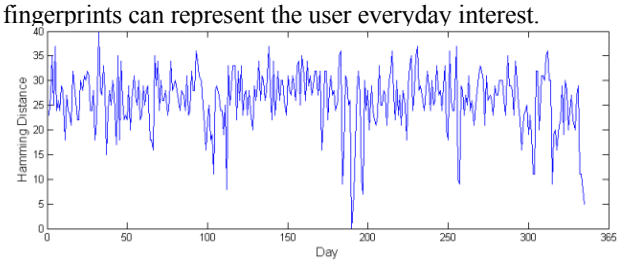

Figure 3. Changing of hamming distance between each two fingerprints

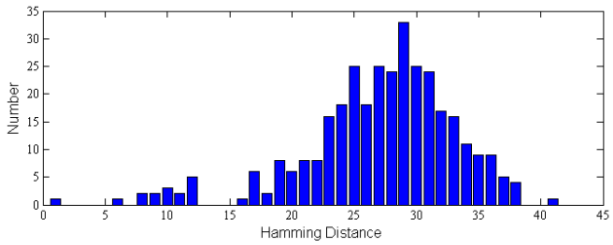

Figure 4. Distribution of the hamming Distances

Fig. 3 shows the changing curve of hamming distance between fingerprint of tweet sets posted each day and the day before, Fig. 4 shows the distribution of hamming distances. We can conclude that the hamming distances are grouped mainly between 20 and 35(nearly 85%), and it means that the tweets posted each day were almost different from those posed the day before. There still exists a few hamming distances which are smaller than 15(nearly 7%), and it means that the user paid attention to some hot topics in those days.

VI. CONCLUSION

In this paper, we propose the micro-blogging user's interest fingerprint to represent his interest features. After data pre-processing, feature extraction and fingerprint generation, we finally obtain the interest fingerprint which is proved very useful when analyzing the features and change patterns of user's interest by our experiments.

However, we only consider the synonym semantics of terms when analyzing the tweet set's features. To analyze the tweets more accurately by analyzing the tweet's context is our further work.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for a careful checking and for helpful comments that improved this paper. This work was supported by National Key Technology R&D Program (No.2012BAH38B-04), "863" Program (No.2010AA012505, 2011AA010702, 2012AA01A401, 2012AA01A402) and NSFC (No. 60933005)

REFERENCES

[1] Jiang Peng, The reduction of micro-bloggging users' activity and stability,http://whb.news365.com.cn/jkw/201112/t20111221_3209369.htm.

[2] Huberman B A, Romero D M and Wu Fang, "Social networks that matter: twitter under microscope," First Monday, 2009, pp. 1–5.

[3] Kwak H, Lee C, Park H, et al., "What is twitter, a social network or a new media," In: Proceedings of the 19th Int Conf on World Wide Web. New York: ACM, 2010, pp. 591–600.

[4] Fan Pengyi, Wang Hui, Jiang Zhihong, et al., "Measurement of Microblogging Network," Journal of Computer Reasearch and Development, 2012, pp. 691–699.

[5] Cha M, Haddadi H, Benevenuto F, et al., "Measuring user influence in Twitter: the million follower fallacy," In: Proceedings of the 4th International AAAI Conference on Weblogs and Social, 2010, pp. 10–17.

[6] Qu Z and Liu Y, "Interactive group suggesting for twitter," In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 2, 2011, pp. 519–523.

[7] Huang J, Thornton K M and Efthimiadis E N. "Conversational tagging in Twitter," In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 2010, pp. 173–178.

[8] Wu W, Zhang B and Ostendorf M, "Automatic generation of personalized annotation tags for twitter users," In: Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association, 2010, pp. 689–692.

[9] Gimpel K, Schneider N, Oonnor B, et al., "Part-of-speech tagging for twitter: annotation, features, and experiments," In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 2, 2011, pp. 42–47.

[10] Liu X, Zhang S,Wei F and Zhou M, "Recognizing named entities in tweets," In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Human Language Technologies, vol. 1, 2011, pp. 359–367.

[11] ICTCLAS. http://ictclas.org/.

[12] HIT-SCIR. "Synonym Lexicon-expanded version," http://www.datatang.com/data/42306.

[13] Moses s Charikar, "Similarity Estimation Techniques from Rounding Algorithms," Annual ACM Symposium on Theory of Computing. USA: ACM, 2002, pp. 380–388.

[14] OpenAPI of Sina micro-blog. http://open.weibo.com/.