A Graph-based Reliable User Classification

Bayar Tsolmon, Kyung-Soon Lee*

Division of Computer Science and Engineering, CAIIT, Chonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do, 561-756 Republic of Korea
bayar_277@yahoo.com, selfsolee@chonbuk.ac.kr

Abstract. When some hot social issue or event occurs, it will significantly increase the number of comments and retweet on that day on Twitter. However, as the amount of SNS data increases, the noise also increases synchronously, thus a reliable user classification method is being required. In this paper, we classify the users who are interested in the issue as "socially well-known user" and "reliable and highly active user". "A graph-based user reliability measurement" and "Weekly user activity measurement" are introduced to classify users who are interested in the issue. Eight of social issues were experimented in Twitter data to verify validity of the proposed method. The top 10 results of the experiment showed 76.8% of performance in average precision (P@10). The experimental results show that the proposed method is effective for classifying users in Twitter corpus.

Keywords: Graph-based user metric, User classification, Timeline analysis

1 Introduction

With the rapidly increasing amount of data on the internet lately, the research on social user classification attracts more and more attention. Compared with the news and blog data, the Social Network Service (SNS) data are more widely used in the real-time event extraction and recommendation system. However, as the amount of SNS data increases, the noise also increases synchronously, thus a reliable user classification method is being required.

Since the existing user classification methods [1,2,3] only depend on the statically behavior of the user, there was a problem that some important user who has fewer numbers of followers might be missed. Especially the frequency of a retweet of tweets that is irrelevant to the event such as rumor or advertisement is high in the SNS environment.

There are recent works for social user classification based on user behavior analysis and timeline analysis. Social media has become indispensable to users recently. A rich set of studies has been conducted in various forms of social media. T. Tinati et al. [4] developed a model based upon the Twitter message exchange which enables us to analyze conversations around specific topics and identify key players in a conversation. Kwak et al. [5] compared three different measures of influence-

* Corresponding author.

T. Herawan et al. (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7_7,

number of followers, page-rank, and the number of retweets-finding that the ranking of the most influential users differed depending on the measure.

In this paper, we propose a graph based reliable user classification based on timeline and social user behavior analysis, and a user activity measurement method to extract reliable and highly active users who are needed for recommendation system or event extraction. The proposed method classifies the Twitter users as socially well-known user, reliable and highly active user, normal user and low active user. *Reliable and highly active user* is defined as the "user who writes a lot" about the issue, and that is "user who is being re-tweeted for several times". When the writing about an issue is being mentioned and being retweeted by people, it's become a reliable user.

The rest of the paper is organized as follows: Section 2 presents user classification; Section 3 describes the proposed method of a graph-based reliable user extraction method. Section 4 shows our experimental results on a Korean tweet collection. We conclude the paper in Section 5.

2 User Classification

1) User classification using follower and following ratio. It could be seen that the characteristic that makes Twitter a social network service is following and follower. Measuring follower and following ratio (*FFRatio*) that shows how much does the user do Twitter activity by using a Twitter user`s number of following and follower.

$$FFRatio(p) = \frac{\#\ of Follower(p)}{\#\ of\ Following(p)} \quad (1)$$

where, *Follower(p)* represents the number of followers of user *p* and *Following(p)* is the number of following of the user *p*.

2) User classification using a retweet and tweet ratio. The retweet ratios tend to be most meaningful when they are used to compare users within the same issue. The retweet ratio (*RTRatio*) of Twitter user who mentioned about an issue is calculated as follows.

$$RTRatio(p) = \frac{Total\ RT(p)}{Total\ Tweet(p)} \quad (2)$$

where, *TotalTweet(p*) represents how many tweets have been posted by the user *p* about the issue. *TotalRT(p)* is the number of retweets for all the tweets posted by the user *p*. The formula shows the average retweet ratio when user *p* writes a tweet about an issue.

A graph based on *FFRatio* and *RTRatio* is as in the Figure 1. The left side of Figure 1 shows the distribution of followings and followers. Following and follower ratio is almost same. The higher the number of the follower can be considered as a socially well-known user. Based on the *FFRatio* value of each user, the Twitter users can be categorized into three groups as follows (Table 1).

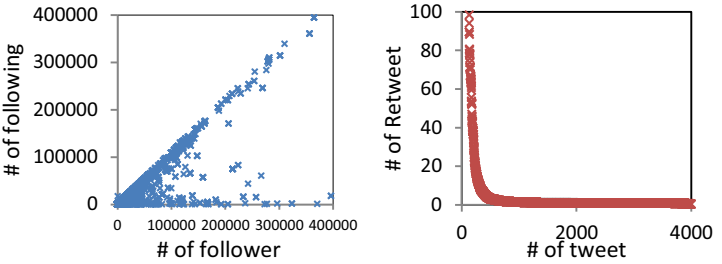

Fig. 1. Distribution of number of followings and followers and *RTRatio*

Table 1. User distribution based on the *FFRatio* value

User group	*FFRatio*	Description
CLB: Celebrity	Greater than 1.5	Follower >> Following
NRM: Normal	0.5 ~ 1.5	Follower ≈ Following
SPM: Spam	Less than 0.5	Follower << Following

The description of each group users is follows.

- CLB user: As a user who has many followers, classify as a celebrity
- NRM user: The case that number of following and follower is similar, classify as a normal
- SPM user: As a user who has more following than followers, classify as a spam or user with low activity

The right side of Figure 1 represents the retweet distribution (*RTRatio*). There is not much users with higher retweet frequency compared as entire users. From this figure it can be seen that users are divided by 4 big groups as follows (Table 2).

Table 2. Table of user distribution through the *RTRatio* value

User group	*RTRatio*	Description	
		Total RT	Total Tweet
A: Popular	Greater than 80	High	Low
B: Active	2.5 ~ 80	High	High
C: Normal	0.5 ~2.5	Low	High
D: Inactive	Less than 0.5	Low	Low

In this paper, we classify a Twitter user who is an important property for the event extraction based on *FFRatio* and *RTRatio* to four groups as follows.

User classification:

1) Socially well-known user (A-CLB & B-CLB)
2) Reliable and highly active user (A-NRM & B-NRM)
3) Normal user (C-NRM)
4) Low active user (D-SPM)

If the socially well-known users mention about the certain event, it indicates that a big social event happened. The reliable and highly active users are valuable users because they post important information every time an event occurs. The user who belongs to A-CLB (combination of group A and CLB) is the user who has a number of followers than following, and has 80 or more of retweet of the tweet wrote about an issue.

3 A graph-based reliable user extraction method

3.1 Extracting socially well-known users

It is hard to analyze user reliability based on the number of Twitter user's followers and tweets. However, highly active users tend to have a lot number of tweets and retweets. To extract socially well-known users, we adapted a HITS (Hyperlink-Induced Topic Search) [6] algorithm to extract "a user who is being retweeted several times" and "a user who is active in Twitter" by analyzing the social network among the Twitter users.

The directed graph is constructed as G = (V, E) on the each issue: *V* represents a user group and *E* represents a linkage group. The directed edge (p, q) $\in$ E is created when the tweet of a user *p* mentions a user *q*. Additionally applying mention, RT, Retweet value with edge weight between nodes to the existing HITS algorithm is as the same as in figure 2.

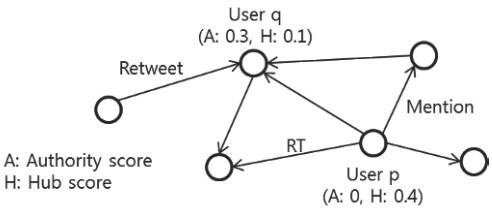

Fig. 2. Example of the weighted HITS graph illustration: Edge creation for Twitter users

Each user has an authority score and a hub score in HITS algorithm [6]. When out-link from the high authority node, it becomes higher hub, and when it is in-linked from high hub nodes, it becomes high authority. The formula that calculated by giving additional an edge weight on the original HITS algorithm to classify users by analyzing network among Twitter users in this paper is as follows.

$$HubScore^{(0)}(p) = FFRatio(p) \tag{3}$$

$$AuthScore^{(0)}(p) = RTRatio(p) \tag{4}$$

The weighted Hub score and Authority score are calculated as formula 5, 6.

$$HubScore^{(T+1)}(p) = \sum_{p \to q} w_{pq} \times AuthScore^{T}(q) \tag{5}$$

$$AuthScore^{(T+1)}(p) = \sum_{q \to p} w_{qp} \times HubScore^{T}(q) \tag{6}$$

The edge weight w_{qp} is as follows:

$$w_{qp} = \sum_{q \to p} FreqRT(q,p) + \sum_{q \to p} Mention(q,p) \tag{7}$$

The effectiveness of the HITS algorithm depends on the initial value and edge weight. The top ranked 100 users with high *AuthScores* are selected as socially well-known users.

3.2 Extracting reliable and highly active users

In the HITS algorithm, the generally reliable users in society are extracted without considering the timeline of each user. Since there is a user who writes a lot of tweets about the issue and actively write whenever the event related to the issue occurs, these users regularly have relatively higher activity than other users. In figure 3, user 2 and user 3 write tweets about the issue every week. User 1 and user 4 do Twitter activity on the issue only in a certain period.

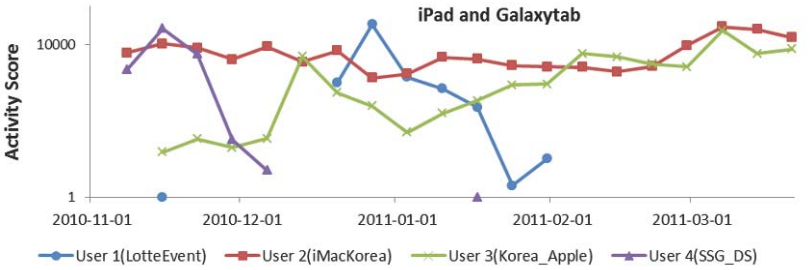

Fig. 3. A Graph of a user who has a high activity score

For cases like user 2 and user 3, the following formula calculates the average weekly activity score with user's weekly tweets and frequency of the retweet for measuring "user who writes often" as an average value.

$$Activity\ Score(u) = \frac{1}{W}\sum_{i=1}^{W} TweetFreq(u, d_i) \times RTFreq(u, d_i) \quad (8)$$

where w shows the number of weeks; $TweetFreq$ shows the sum of tweets d that a user u wrote in the each ith week; $RTFreq$ represents the number of retweets d of the tweets written by a user u in the each ith week.

4 Experiments and Evaluation

We have evaluated the effectiveness of the proposed method on tweet collection. Eight issues are chosen and tweet documents for the issues are collected, spanning from November 1, 2010 to March 26, 2011 by Twitter API (all issues and tweets are written in Korean). Table 3 shows the number of users who wrote tweets on each issue.

Table 3. Twitter user set

Category	*Issue*	*# of user*
Product	Canon & Nikon	21,369
	iPad & Galaxy tab	115,022
People	Park Ji-Sung	29,568
	Kim Yu-Na	10,563
Company	Apple	71,878
	Samsung	108,800
Natural Disaster	Earthquake	110,345
Terrorism	Chonanham	19,473

The methods of comparative experiment to extract reliable user is as in the following. In this paper, the method based on Twitter user's *FFRatio* and *RTRatio* is used as a baseline for the extraction of reliable users.

Table 4. Comparative Methods & User classification

User classification	Baseline	Proposed method
Socially well-known user	A-CLB & B-CLB	HITS AuthScore: High
Reliable and active user	A-NRM & B-NRM	Activity Score
Normal user	C-NRM	HITS HubScore: High
Low active user	D-SPM	HITS Auth & Hub Score: Low

The reliable user extraction method using *FFRatio* and *RTRatio* in the static behavior analysis of Twitter user and using dynamic behavior analysis on the issue are proposed. The result of calculation of precision (P@10) for top 10 extracted reliable users from each method is as follows.

Table 5. Comparative experiment for reliable user extraction P@10

User classification	*Socially well-known user*		*Reliable and active user*		*Average*	
Issue	A-CLB B-CLB	Auth Score	B-NRM C-NRM	Activity Score	Baseline	Proposed method
1. Canon & Nikon	**0.5**	0.4	0.2	**0.4**	0.35	0.40
2. iPad & Galaxy tab	0.7	**0.8**	**0.9**	0.8	**0.80**	**0.80**
3. Park Ji-Sung	**0.7**	0.6	0.6	**1.0**	0.65	**0.80**
4. Kim Yu-Na	0.5	**0.9**	0.6	**0.9**	0.55	**0.90**
5. Apple	0.5	0.5	0.5	**0.7**	0.50	0.60
6. Samsung	0.5	**1.0**	0.3	**0.9**	0.40	**0.95**
7. Earthquake	0.4	**0.7**	0.2	**0.8**	0.30	0.75
8. Chonanham	0.8	**0.9**	0.5	**1.0**	0.65	**0.95**
Average	0.57	**0.72**	0.47	**0.81**	0.52	**0.76**

In the result of experiment, the proposed method showed better performance than baseline. The method *AuthScore* and *Activity Score* achieved 72% and 81% respectively. Since an experiment carried out by targeting users who are interested in the issue, a user who is socially popular about each issue, a user who has a reliable and direct correlation with an event could be extracted.

The distribution of reliable users' tweet about the earthquake issue is shown in Figure 4.

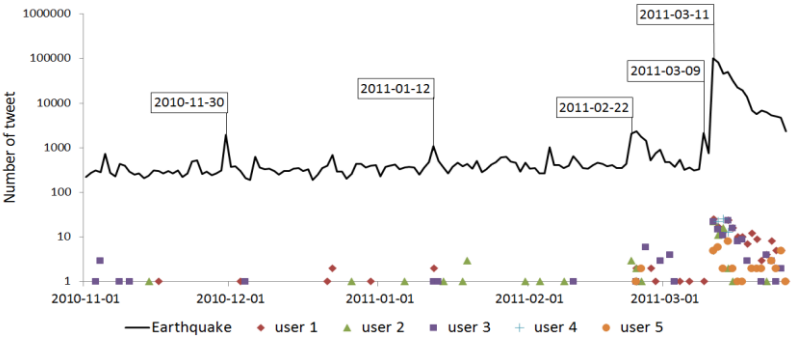

Fig. 4. The number of tweets containing the "*Earthquake*" issue word and distribution of reliable users.

Table 6. Top 5 reliable users about earthquake issue

User ID	User screen name	Description
User 1	KoreanRedCross	Official Twitter account of the Korean Red Cross
User 2	parknews9	KBS Nine O'clock News anchor
User 3	kbsnewstweet	Official Twitter account of KBS News
User 4	Russa	Blogger
User 5	mofatkr	Ministry of Foreign Affairs

In the Figure 4, the higher picks with date label represent the occurrence of earthquakes. From here, it can be seen that when the number of daily tweet frequency become higher, all reliable users wrote about an issue on the same day. Reliable and highly active top 5 users related to earthquake issue are shown in Table 6. In the case of user 1, tweets are not only about earthquake occurrence but also about donation and help. In other hand, when an earthquake occurs, user 5 posts tweet about guidelines for South Korean citizens who live in abroad. A user 2 and user 3 write the earthquake news.

5 Conclusion

In this paper, a Twitter user classification method through graph-based reliability measurement metrics and user activity metrics using timeline analysis were proposed in a network of users who are interested in the issue. Reliable user can be used for other use such as recommendation system, as well as event extraction. Eight of social issues were experimented in Twitter data to verify validity of the proposed method. The top 10 results of experiments achieved 76.8% of performance in precision (P@10). Discovering methods for better user behavior analysis and less dependent on the number of tweets are future works.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2044811).

References

1. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1633-1636, ACM (2010)
2. Boyd, D., Golder, S., & Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In System Sciences (HICSS), pp. 1-10. IEEE (2010)
3. Mendoza, M., Poblete, B., Castillo, C.: Twitter Under Crisis: Can we trust what we RT?. In Proceedings of the first workshop on social media analytics, pp. 71-79, ACM (2010).
4. Tinati, R., Carr, L., Hall, W., Bentwood, J.: Identifying communicator roles in Twitter. In Proceedings of the 21st international conference companion on World Wide Web, pp. 1161-1168, ACM (2012)

5. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World Wide Web, pp. 591-600, ACM. (2010)
6. Kleinberg J. M.: Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, 46(5) pp. 604-632, (1999)