Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2013 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

6-18-2013

ACQR: A Novel Framework to Identify and Predict Influential Users in Micro-Blogging

Wen Chai
Renmin University of China, chaiwen78@ruc.edu.cn

Wei Xu
Renmin University of China, weixu@ruc.edu.cn

Meiyun Zuo
Renmin University of China, zuomy@ruc.edu.cn

Xiaowei Wen
Renmin University of China, elowise0516@ruc.edu.cn

ACQR: A NOVEL FRAMEWORK TO IDENTIFY AND PREDICT INFLUENTIAL USERS IN MICRO-BLOGGING

Wen Chai, School of Business, Renmin University of China, Beijing, China, chaiwen78@ruc.edu.cn

Wei Xu, School of Information, Renmin University of China, Beijing, China, weixu@ruc.edu.cn

Meiyun Zuo*, School of Information, Renmin University of China, Beijing, China, zuomy@ruc.edu.cn

Xiaowei Wen , School of Information, Renmin University of China, Beijing, China, elowise0516@ruc.edu.cn

Abstract

As key roles of online social networks, influential users in micro-blogging have the ability to influence the attitudes or behaviour of others. When it comes to marketing, the users' influence should be associated with a certain topic or field on which people have different levels of preference and expertise. In order to identify and predict influential users in a specific topic more effectively, users' actual influential capability on a certain topic and potential influence unlimited by topics is combined into a novel comprehensive framework named "ACQR" in this research. ACQR framework depicts the attributes of the influentials from four aspects, including activeness (A), centrality (C), quality of post (Q) and reputation (R). Based on this framework, a data mining method is developed for discovering and forecasting the top influentials. Empirical results reveal that our ACQR framework and the data mining method by TOPSIS and SVMs (with polynomial and RBF kernels) can perform very well in identifying and predicting influential users in a certain topic (such as iPhone 5). Furthermore, the dynamic change processes of users' influence from longitudinal perspective are analysed and suggestions to the sales managers are provided.

Keywords: Online social networks, Influential users, Identification, Prediction, Online marketing.

*corresponding author

1 INTRODUCTION

In social network, influential users have the ability to informally influence the attitudes or behaviour of others (Venkatraman 1989; Goldsmith & Horowitz 2006; Amblee & Bui 2011). In information-driven, knowledge-based global market place, enterprises increasingly attach importance to competitive/market intelligence capability (Liebowitz 2006). So many of them rethink their marketing strategies, and want to combine users' influence to implement their marketing actions. Users' influence can be used in word-of-mouth (WoM) marketing which has played an increasingly important role in online shopping (Park et al. 2007; Cheung et al. 2009), and the commercial value will come up with (Culnan, et al. 2010; Stephen & Toubia 2010).

As an important type of online social networking, micro-blogging has experienced spectacular growth since its first debut in 2006. With the fast growing of the number of users, the influence of micro-blogging is becoming more and more widen. When it comes to marketing in micro-blogging, the users' influence should be associated with a certain topic or field on which people have different levels of expertise (Cha et al. 2010). Meanwhile, individuals' online influence is dynamically formed and developed (Cha et al. 2010; Wu et al. 2011) because of the interplay between users' interaction and their social network (Agarwal et al. 2008). If the influentials in micro-blogging can be effectively picked out and forecasted, company could not only directly target them to implement marketing strategy, but also save more marketing cost through fostering and guiding the potential future influentials more effectively. Therefore, how to effectively identify and predict online influential users on a certain topic is a very important issue.

Although the topic relevance and dynamic of influence in e-commerce is highlighted (Cha et al. 2010; Li & Du 2011), the previous efforts more focus on the most influential or popular figures over the global network rather than in a certain topic or field, and most of them analyze the users' influence under static scenarios instead of exhibiting a dynamic change process with time. Hence, here we propose our concerns: 1) How to identify the influential users under a certain topic in micro-blogging with a comprehensive framework? 2) How to predict the influential users on a specific topic in the future? 3) How dose users' influence change with time from longitudinal perspective of topic development?

To answer the research questions above, the remainder of the paper is organized as follows: Section 2 reviews the related literatures and finds the current research gap. Section 3 introduces our research methodology including analysis framework and data mining method for the identification and prediction of influential users. And our proposed framework and data mining method are demonstrated under a selected topic (i.e. iPhone 5). Afterwards, evaluation and further discussion are presented in Section 4. Finally, conclusion and future work are summarized in Section 5.

2 LITERATURE REVIEW

2.1 Definition of Influential Users

In traditional views, a minority of members who are exceptionally persuasive in spreading ideas to others are considered to drive trends on behalf of the majority of ordinary people. The influential individuals are called the opinion leaders in the two-step flow theory (Katz & Lazarsfeld 1955) or innovators in the diffusion of innovation theory (Rogers 1995). By employing the influential roles, a viral marketing can reach a wide audience. These theories of influentials spread well and have been adopted in many businesses. With the emphasis of prevailing culture in a more modern view, scholars begin to reason the choice making based on the opinions of peers and friends (Xiong & Liu 2004; Park et al. 2007). However, the traditional theories of influentials have been criticized because of without taking into account the role of ordinary users. In fact, under a certain topic or field, the ordinary people

also have power or capacity of causing an effect. For example, in micro-blogging, individuals often follow some ordinary users who have special insights or expertise on a certain topic.

Scholars in communication studies and social-psychology define the most influential users of virtual communities (VCs) by utilizing their personality (such as credibility) and the actual content of the communication (O'Keefe 2002). The influential roles have the ability to spark conversations, trigger feedback, or even shape the way of communication in one group (Huffaker 2010). Contributing novel information to the network is emphasized as an essential property of influentials in blogosphere who bring the most representative opinions to their local social network (Song et al. 2007). The online influentials are also described with the words of knowledgeable, communicable, respectable, innovative, and central important to their followers (Li & Du 2011). For the specific context of online social networks (OSN), influential and prestigious nodes are recognized as these nodes with high network status, and commonly identified by several different centrality measures (Heidemann et al. 2010). As a special case of OSN, micro-blogging has distinct characteristics on information diffusion. For instance, unique kind of ordinary users on Twitter who play the role of intermediate layer transmitting almost half of the information broadcasted to the masses are distinguished as an important kind of influentials (Wu et al. 2011).

In summary, the concept of influential user in online communication is described in terms of the personality, activity, and status in social network based on the literatures above. Among these three aspects, the personality and status in social network can be regarded as the information disseminating foundation which is the potential influence unrelated with the user's actual communication behaviour on one topic. Meanwhile, user's preference and expertise on a certain topic represented by her/is activity is also essential. On one hand, the potential influence needs to be transformed into actual influence. And on the other hand, the user' willingness and capability for information diffusion related with a specific topic are more pertinent to online marketing and e-commerce.

2.2 Influential Users Identification and Prediction

Currently, a user's influence on Twitter is often measured with the number of followers. The more followers one has, the more influential s/he is. Similarly, the number of retweets and the number of mentions are used as metrics for users' influence calculation (Cha et al. 2010; Kwak et al. 2010). Besides, some other indicators rely on the ratio between the number of one's followers and the number of one's friends and the ratio of attention a user received (retweet, reply and mention) to the tweets s/he posted (Wu & Wang 2012). No matter how different the indicators are, they are all based on the statistical characteristics. Some other studies prefer the approaches of social network analysis that consider direct and/or indirect connections in social networks. Different centrality measures are drawn for identifying important nodes. The three most common measures to quantify the centrality of a node in social networks are respectively degree centrality, closeness centrality, betweenness centrality (Freeman 1979). Since pagerank algorithm was published, scholars try to use and modify this algorithm (such as leader-rank and TunkRank) to extract the online leaders (Song et al. 2007; Heidemann et al. 2010; Wu & Wang 2012).

In the diffusion process of users' influence, connectivity and activity of users in online social network has been emphasized by multiple authors (Heidemann et al. 2010; Huffaker 2010; Zhang et al. 2011). The former can be taken as the potential influence which constructs the network foundation for information diffusion; the latter is used to embody the user's willingness, preference and capability to a certain topic which convert the possible effects into reality. Nonetheless, they are often separately analyzed to represent users' influence. According to the previous research, metrics only involving the users' posting and interaction behaviour but not considering the social link structure are not comprehensive measures (Wu & Wang 2012). As what we know, the number of followers does not always represent the users' influence under a specific context. It is easy to understand that a pop star with high number of followers but never participating in a discussion on the topic of specific

electronic equipment can't be labelled as an influential in this context. That is why we point out that a single index can't comprehensively depict the characteristics of the influentials under a certain topic.

Furthermore, to the best of our knowledge, influential user prediction is an important but rarely researched issue. Bakshy attempted to predict the future influence of seed users on Twitter by using regression tree model, but the fitness of the model on individual level was relatively poor (R^2= 0.34) (Bakshy et al. 2011). With the hope to explore the influential users in the future communication circle, we will take the previous study for reference to propose novel predictors and try other possible methods to improve the performance.

Hence, this paper proposes a comprehensive research framework to cover the multi-dimensional characteristics of an influential, including the activeness (A), centrality (C), quality of post (Q) and reputation (R), by taking activities, social links and properties of sender and receiver into consideration. On the basis of qualitative analysis, a data mining method is contributed to identify and predict influential users in micro-blogging under a certain topic.

3 THE PROPOSED RESEARCH METHODOLOGY

3.1 An ACQR Framework

According to existing literatures and our analysis, personality, activity and connectivity will be used to depict an influential from two essential dimensions. One is the user's actual influential capability on a specific topic, and the other is the user's potential influence for information diffusion.

As shown in Fig. 1, a novel comprehensive framework called ACQR is adopted to depict an influential under a certain topic by the following four attributes: (1) **A**ctiveness (A): s/he actively expresses opinions and interacts with others under a specific topic. (2) **C**entrality (C): s/he occupies the central position in social network which provides structure foundation for the influence diffusion. (3) **Q**uality of Post (Q): her/is posts under a certain topic are widely accepted and reposted based on the high quality of content. (4) **R**eputation (R): her/is good reputation will help establishing the credibility in online world.

In our framework, **A** and **Q** are used to respectively reflect the user's actual preference and expertise on one topic, both of which consist of the user's actual influential capability related with the topic. **C** and **R** represent the user's potential influence scope and credibility, both of which built the user's potential foundation for information disseminating. The proposed framework can pick out the most influential roles by ranking the users' influence based on the four features which are explained in detail as follows.

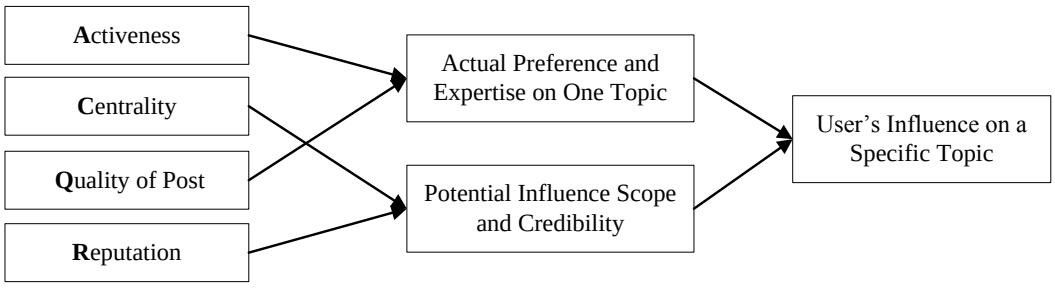

Figure 1. The ACQR Framework

1）Activeness: In micro-blogging, a user can tweet, and retweet/reply others' tweets. Activity on a certain topic converts the user's potential influence into actual impact. More active participation means the user's higher level of preference and contribution on one selected topic (Huffaker 2010). On the other hand, frequent interaction can enhance the trust between each other, which are important

premise of social influence (Lua et al. 2011). Furthermore, the probability of user's tweet read by the others is related to her/is active level. For example, if user A is much more active than B, it is easier to read the tweets of A than the ones of B from their mutual friend C. So the probability that user C reads a tweet from user A is much higher than user B (Wu & Wang 2012). Therefore, often posting and retweeting/replying other's tweets will enhance one's online influence.

2）Centrality: Different with the activeness, centrality in a social network is structural characteristic which has been extensively studied to understand and explain human behaviour in multiple social networks. Previous literatures suggest that a user's online social network status play a critical role for influential identification. Key individuals' connectivity or position in the network triggers a cascade of influence (Resnick et al. 2000), and indicate how individuals become influential through the relationships that they build (Wasserman & Faust 1994). It is pointed out that a user's connectivity in the network could be a significant factor for advertising effectiveness. Well-connected users with many connections to others can be highly relevant for the promotion of brands, products, and viral marketing campaigns (Heidemann, et al. 2010).

3）Quality of Post: Quality of post represents the author's expertise and capability on one selected topic. Posts with high quality are more influential than those with low quality. When users search for information under some certain topics in micro-blogging, they want to be provided higher quality content, rather than the spam containing the same key words. In fact, the level to which a tweet contains high quality information varies dramatically. Some scholars describe interesting tweets with Well-formedness, factuality and navigational quality (Vosecky et al. 2012). Emoticons, post length, shouting and the existence of hyperlinks are also considered as quality indicators for micro-blogging posts (Weerkamp & Rijke 2008). No matter judged by semantic or form, high quality tweets will be accepted and retweeted by readers. So the number of retweeted by the others can be taken as a measurement of the quality (Massoudi et al. 2011).

4）Reputation: With rapid popularization, OSN has unfortunately been employed as channels to diffuse spam. The undesirable content with camouflage existing in OSN could potentially attract a large number of users. If the information is irrelevant or spam, the receiver cannot accept or trust it. To avoid the spam and browse desirable content, scholars draw attention to users' reputation (Lua et al. 2011; Jung 2012). Generally, the information from the user with good reputation is credible. Similar with centrality in social network, good reputation is unrelated with the actual behaviour under a certain topic, and is also an important aspect of potential influence.

3.2 A Data Mining Method

Under the proposed framework, this subsection presents a data mining method to pick out and forecast the influential users. As shown in Fig. 2, it consists of six steps:

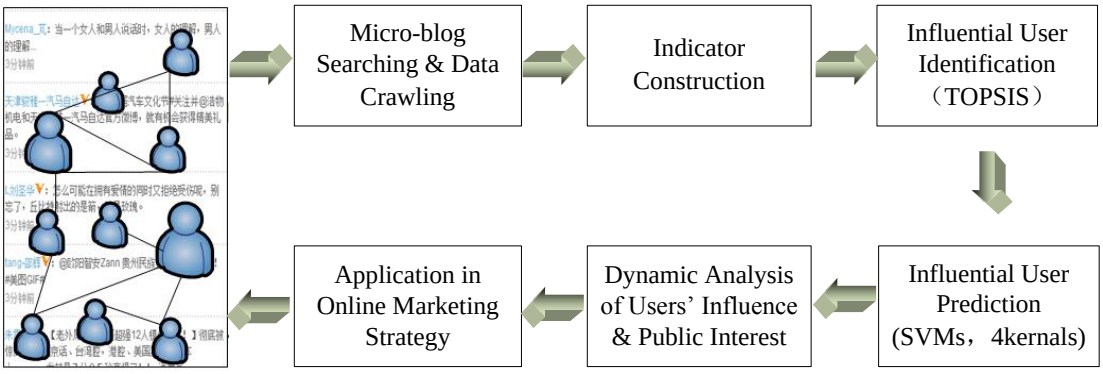

Figure 2. The Data Mining Method for Identifying and Predicting Influential Users

(1) Keyword searching and data crawling. This step uses some hot keywords to locate seed users in micro-blogging platform. This can be carried out by entering keyword (in this paper, iPhone 5 is taken as the keyword) into search engine provided by the micro-blogging platform. Then the micro-blogging that contains the keywords and the corresponding information about the sender and receiver can be located. After that, open API and crawler developed by ourselves will be used to crawl the data.

(2) Indicator construction. As the discussion above, activeness (A), centrality (C), quality of post (Q) and reputation (R) of a user will be used to indicate one's influence under a specific topic. Referring to the suggestion that the interactions in an online social network are affected by shared message, involvement, and relationships (Yang & Ng 2007), different sources of data are drawn to mine influential roles, covering activities, properties (marked with "V" or not), social network relationships of senders and receivers (as shown in Fig. 3). Finally, 5 indicators are extracted as shown in Table 1.

Feature	Indicator	Description
A	$F_1 : N_{tweets} + N_{reposting} + N_{replying}$	N_{tweets} is the number of the user's original tweets; $N_{reposting}$ is the number of the user reposting others' tweets; $N_{replying}$ is the number of the user replying others' tweets.
C	F_2 : In-degree centrality	The number of links to a node
	F_3 : $PR(i) = \frac{(1-d)}{N} + d \bullet \sum_{j \in B_i} \frac{PR(j)}{O_j}$	N is taken as the total number of users in the network, B_i as the set of users follow the user i, and O_j as the number of outgoing links from user j.
Q	$F_4 = \sum_{i=1}^{N_{reposts}} w_i + \sum_{i=1}^{N_{replies}} w_i$	w_i is the reputation value (shown in F_5) of reposter and replier; $N_{reposts}$ is the number of reposted by the others; $N_{replies}$ is the number of replied by the others.
R	F_5 : Reputation value	Level 1: Reputation value =0.25 (non-marked with "V", and followers non-marked with "V"); Level 2:Reputation value =0.5 (non-marked with "V", but followers marked with "V"); Level 3:Reputation value =0.75 (marked with "V", but followers non-marked with "V"); Level 4: Reputation value =1.0 (marked with "V", and followers marked with "V").

Table 1. Indicators and Measurement

"A" is ascertained from the number of tweets and the number of retweeting and replying others tweets (F_1). "C" in social network is determined by the in-degree centrality (F_2) (Hanneman & Riddle 2005) and pagerank value (F_3) (Brin & Page 1998; Langville & Meyer 2004) in this paper. In micro-blogging, how broadly the user's influence can diffuse more depend on her/is followers and their followers' connectivity in online social network, and not all connections are equal. "Q" (F_4) is calculated by the number of reposted and replied (Massoudi et al. 2011) which is modified by considering the impact of receivers' reputation. The higher the receiver's reputation value is, the higher the quality score is. "R" (F_5) is judged by user's reputation value which is classified into 4 levels according to whether the user and her/is followers are marked with "V". Once a user has been marked with "V", her/is identity has been verified by the micro-blogging platform, meaning that the user's reputation in real life is acknowledged in virtual world.

(3) Influential user identification. In this part, the indicators introduced above will be used to rank top K (e.g. top 0.5%, top 1%, top 5% and top 10%) influential users in micro-blogging under a specific topic. As influential user identification is a multi-features decision problem, Technique for

Order Preference by Similarity to Ideal Solution (TOPSIS) (Ju & Wang 2012) is selected as our approach. TOPSIS summarizes the Euclidean distance between measurements and the ideal solution, and the chosen alternative should have the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution.

All indicators are positively associated with user's influence in our study. For the indicator j, the positive ideal solution A_j^+ is determined as $A_j^+ = max\left(F_{ij}\right),(j=1,...5)$, and the negative ideal solution A_j^- is determined as $A_j^- = min\left(F_{ij}\right),(j=1,...5)$, where F_{ij} is a crisp value indicating the performance rating of each user i with regard to each indicator j.

Then, the Euclidean distance between user i and the positive ideal solution is calculated by $d_i^+ = \sqrt{\sum_{j=1}^{5}(A_j^+ - F_{ij})^2}$, and the Euclidean distance between the user i and the negative ideal solution is calculated by $d_i^- = \sqrt{\sum_{j=1}^{5}(A_j^- - F_{ij})^2}$. The standard TOPSIS is described as $s_i = d_i^- / (d_i^+ + d_i^-)$.

(4) Influential user prediction. This stage employs the users' activities, social links and properties in the past to forecast whether s/he can be the top K influential users in the next period of interaction. As the purpose of classification is to establish a model which can predict an attribute based on some other attributes. One of the most useful classification algorithm, support vector machines (SVM), can be employed for forecasting issue (Varol et al. 2009). Support vector machines (SVMs), proposed by Vapnik (1995), is based on Statistical Learning Theory (SLT), and has been widely applied in classification and prediction.

A classification task usually separate data into training and testing sets. Each instance of training set contains several "attributes" and one "target value" (i.e. the class labels).The goal of Support Vector Machines (SVMs) is to develop a model (based on the training set) which predicts the target value of the testing set given the testing data attributes. For better understanding our procedure, more details of SVMs can be found in (Schölkopf 2000).

(5) Dynamic analysis of users' influence and public interest. Based on the results of identification and prediction, the dynamic of users' influence with the topic development will be further discussed. Two questions are going to be answered: 1) how does a user's influence change? 2) whether the public will change their focus according to the influentials' ranking adjustment?

(6) Application in online marketing strategy. First, the dynamic process of users' influence exhibited in our research will provide evidence for online marketing strategy adjustment. Because, at the different phases of topic development, the most influential users' ranking changes with time. It is more important that our proposed framework and data mining method can help companies to discover and forecast who the most influential users are. And then the special individuals can be guided and motivated by companies to deliver product information, give personal comments provide recommendations, and supplement professional knowledge for product improvement.

3.3 Demonstration

To evaluate the effectiveness and efficiency of the proposed ACQR framework and data mining method, search engine, open API and crawler were used to obtain data from a famous micro-blogging platform in China, including users' social links, activities and properties. An application was built and compiled using C++ and Python for demonstration purposes. As regard to the prediction stage, a software package named "LIBSVM" is introduced, and it has gained wide popularity in classification and forecasting, machine learning and many other areas (Chang & Lin 2011).

Our crawler was launched from August 17th to September 27th, in 2012, during which period the new iPhone (5rd generation) was released. The keyword "iPhone 5" was entered into the search engine

which is provided by the micro-blogging platform itself. Finally, 576,946 posts and 441,899 users were located, and among them, 83,679 reposts and replies generated by 36,646 users were identified. Social links matrix was gained according to the virtual "following" relationship. As an important property, user's status of whether marked with "V" was crawled, including not only the users who participated in the topic, but also their followers.

Before implementing the data mining method, the initial data needs to be preprocessed. Firstly, the indicators based on the ACQR framework were extracted, and then they were normalized using logarithmic transformation (shown in Eq. (1)) for future TOPSIS calculating. In order to predict the future top k influential users, the dataset was divided into different cross-section subsets in form of weekly data.

$$x_i' = log(1+x_i) / log(1+x_{max}) \tag{1}$$

To avoid identifying a user with high activeness but low centrality as an influential, a coefficient of dispersion was employed to modify the standard TOPSIS (Li & Du 2011) (shown in Eq. (2)). The similarity of user i to the ideal solution is shown as follows,

$$s_i' = \left(1 - \frac{SF_{ij}}{\overline{F_{ij}}}\right) \bullet \left[d_i^- / (d_i^+ + d_i^-)\right] \tag{2}$$

where $\overline{F_{ij}}$ is the average of 5 indicators $\overline{F_{ij}} = \frac{1}{5}\sum_{i=1}^{5} F_{ij}$, and SF_{ij} is the standard deviation of these indicators $SF_{ij} = \sqrt{\sum_{i=1}^{5} (\overline{F_{ij}} - F_{ij})^2} / 5$. Finally, the influential users is ranked according to their value of s_i'.The greater s_i' is, the higher the rank is.

In prediction stage, "0/1" was used to label the users (denoted as "I") in different groups which are identified by improved TOPSIS. If a user was identified as a top k (e.g. top 0.5%) user, s/he would be labelled with "1", otherwise labelled with "0".The attributes of first three weeks are used to predict the influentials' label in the next week. In our context, 18 attributes with time delay are chosen as the attributes of SVMs for predicting the target value I^t, including $I^{(t-1)}$, $I^{(t-2)}$, $I^{(t-3)}$, and $F_j^{(t-1)}$, $F_j^{(t-2)}$, $F_j^{(t-3)}$, $j = 1,2,...5$. Specifically, the formula is shown in Eq. (3):

$$I^t = f(I^{(t-1)}, I^{(t-2)}, I^{(t-3)}, F_1^{(t-1)}, F_1^{(t-2)}, F_1^{(t-3)}, F_2^{(t-1)}, F_2^{(t-2)}, F_2^{(t-3)}, ..., F_5^{(t-1)}, F_5^{(t-2)}, F_5^{(t-3)}) \tag{3}$$

4 kernels (Linear kernel, Polynomial kernel, Radial basis function (RBF) kernel, and Sigmoid kernel) of SVM were employed to train the prediction model for forecasting top 0.5%, top1%, top2% and top5%. Then the results was evaluated and compared to verify our proposed framework and data mining method.

4 EVALUATION AND COMPARISON

4.1 Influential Users Identification

The influential users of 6 weeks were respectively identified based on the TOPSIS value. To compare our novel framework with the traditional users ranking measures based on statistical index and social network analysis, several segments of top K identified users by different ranking methods were created, including ACQR framework, number of followers, number of posts (including number of tweets, reposting and replying others), number of reposted (including reposted and replied by the others), and the value of pagerank. The user ranking by ACQR is taken as the benchmark temporarily, and the degree of consistency of 4 methods with it are shown in Table 2.

Top K Users	Week	Posts	Followers	Pagerank	Reposted
100	1	8%	21%	26%	63%
	2	8%	52%	66%	13%
	3	5%	51%	59%	27%
	4	6%	37%	44%	70%
	5	7%	41%	45%	46%
	6	3%	35%	39%	52%
500	1	17%	62%	73%	20%
	2	3%	67%	77%	7%
	3	3%	65%	78%	11%
	4	7%	58%	68%	31%
	5	3%	64%	71%	29%
	6	4%	59%	68%	26%
1000	1	10%	70%	77%	13%
	2	3%	71%	81%	6%
	3	2%	68%	79%	10%
	4	6%	65%	77%	21%
	5	4%	70%	76%	22%
	6	5%	66%	75%	20%
5000	1	6%	67%	36%	6%
	2	6%	68%	37%	4%
	3	7%	68%	36%	7%
	4	10%	64%	37%	15%
	5	9%	66%	36%	16%
	6	11%	65%	36%	17%

Table 2. The consistency of different ranking measures with ACQR

Obviously, user ranking result by ACQR is different with the most popular single indices which are also employed or modified in our framework. As the range extension from top 100 to top 5000，the indicator with highest consistency value switch from "Reposted" to "Followers" and "Pagerank", and finally "Followers" outperform all the others. Specifically, for the top100, the preponderance of "Reposted" is apparent which means that the difference of influence is mainly depicted by the quality of post. From top 500 to top 1000, the "Followers" and "Pagerank" have striking advantage over the others and reflect the most disparity of influence. For the top 5000, the performance of "Followers" is still stable, which equal to emphasize the significant impact of followers number on the influential users' ranking. In contrast, the consistency of our ranking results with "Posts" is lowest from top 100 to top 5000. That's because no apparent difference on posts quantity exists between the influentials and non-influentials. In other words, only a small group excluding spammers performs very actively, and the others are very close in posts number.

On the longitudinal perspective, the consistency values of "Followers" and "Pagerank" reduce moderately after a slight climbing-up accompanying with the reverse process of "Reposted". The disparity in "Reposted" among the participants increase with the topic reaches the hottest. The phenomenon shows that the consistency between the identification results by ACQR and by other commonly used measures vary with the range extended (from top 100 to 5000) and the topic development (from preliminary stage to the hottest stage). Therefore, it is suggested that our ACQR-framework reflects the dynamic of users' involvement and interaction under a certain topic, and embodies the complexity and variety of users' influence on certain topic in the micro-blogging world.

In order to further verify the performance of our framework, the attributes and rank results by ACQR of three typical users from the samples are listed in Table 3. As users' online influence is dynamically formed and developed, the top influentials of these six weeks vary widely, which have been well demonstrated by ACQR framework. Seen from Table 3, based on our concerns both on the actual

influential capability on one topic and potential influence foundation, user A was not ranked (denoted as "/" in the last column) in the first three weeks because of no posting or getting reposted. Although owning the relative smallest amount of followers among these three typical ones, user B list on the peak from 1st to 3rd week. Moreover, B has not been marked with "V", so s/he can be recognized as a star from grassroots. In contrast, user C has a great number of followers, but s/he ranks behind B according to low activeness and quality of posts. If C wants to enhance influence on this topic, s/he needs to obtain more acceptances in form of being reposted, and only taking high centrality in online social network is not enough. It is a good instance that A listed on the peak in 4th and 6th week not only because of the most great number of followers but also the high activeness and quality of post. To sum up, the comparative analysis suggests that the diligent grassroots can perform outstandingly depending on activeness and high quality of post at the preliminary stage of a certain topic when most of potential influential users are inactive. However, with the topic development, more and more individuals participate in the online discussion. Especially, the ones with high central position and expertise will become the top influential at the hottest stage of topic. Finally, as an important advantage of ACQR, no spammers were founded in our top 1000 list when checked manually. So it is suggested that our proposed ACQR framework can identify influential users effectively.

User ID	Week	Post	Followers	Reposted	Marked with "V"	Rank by ACQR
User A	1	0	2021585	0	Y	/
	2	0	2022767	0	Y	/
	3	0	2025223	0	Y	/
	4	8	1957697	208	Y	1
	5	1	1960489	6	Y	31
	6	3	1959711	271	Y	1
User B	1	109	13703	224	N	1
	2	145	13880	864	N	3
	3	497	14087	322	N	1
	4	340	13939	11	N	91
	5	442	13863	7	N	68
	6	368	13861	14	N	50
User C	1	5	836350	0	Y	31
	2	5	800266	0	Y	9
	3	7	765961	0	Y	19
	4	11	736004	4	Y	58
	5	6	705484	0	Y	81
	6	12	675386	0	Y	60

Table 3. Three typical users from samples

4.2 Influential Users Prediction

For the stage of prediction, the influential users in the 6th week are predicted. Whether the predicting results match the identifying results will be determined by some standard accuracy measures. Here, three kinds of measures, namely, the accuracy, precision and recall rates (Olson & Delen 2008) are introduced to compare which SVM-kernel performance better than the others.

Precision present the fraction of retrieved instances that are relevant, while recall means the fraction of relevant instances that are retrieved. Both of them are based on a definition and understanding of relevance. More simply, high recall rate equals to return most of the relevant results. High precision means that returning more relevant results than irrelevant. If any one of them is overemphasized, the rest one will be negatively impacted. In this context, the precision and recall rate are more suitable to be the evaluation measures.

The effect of prediction using 4 kernel functions of SVMs is offered in Table 4. In general, prediction accuracies of different kernel functions are desirable, here the precision and recall need to be further compared. Particularly, liner and sigmoid kernels outperform on the recall rate which means that how much percent of real influential users are founded out by SVMs. While polynomial and RBF kernels perform better on precision which means how much percent of influential users founded by SVMs are real ones. In our context, precision is more important because our target is to find the real influentials but not find more ones. Therefore, in this stage, Polynomial and RBF kernels are recommended. Generally speaking, our prediction effectiveness is ideal.

Kernel	Top K	Accuracy	Recall	Precise
Linear	0.5%	99.56%	81.63%	93.69%
	1%	99.15%	77.44%	90.57%
	2%	98.25%	69.36%	79.31%
	5%	95.67%	70.78%	81.82%
Polynomial	0.5%	99.60%	78.55%	98.71%
	1%	99.24%	74.87%	98.71%
	2%	98.69%	62.23%	95.36%
	5%	96.24%	64.05%	85.22%
RBF	0.5%	99.57%	79.91%	93.71%
	1%	99.24%	74.83%	98.71%
	2%	98.70%	62.11%	95.38%
	5%	96.25%	64.02%	85.27%
Sigmoid	0.5%	99.49%	86.29%	84.52%
	1%	98.99%	80.54%	79.94%
	2%	97.85%	68.16%	63.31%
	5%	94.39%	61.24%	54.60%

Table 4. Predication Accuracy (Comparison of 4 kernel functions)

4.3 Further Discussion

The evolution of user involvement under this selected topic (iPhone 5) during 6 weeks is shown in Fig. 3. As can be seen, the number of post and repost sharply grows from 9th Sep, and the increase of post number is more obvious than repost number. That is because more and more ordinary people (can be named grassroots) join in this topic to express their opinions and thoughts especially on the eve of iPhone 5 released. These grassroots' posts are almost exhibited to their family members or friends, and the receivers may be lack of motivation to repost the information in terms of the content. However, as an important information flow contributed to this topic, the information from grassroots cannot be trifled with. Peer to peer influence will be exerted by some posts which reflect the ordinary people's attitudes and actual actions. Social sharing information from friends is considered valuable for online purchasing, which play an important role in social commerce (Liang et al. 2011).

Meanwhile, the ratio of users non-marked with "V" in different top k list increase evidently with time as shown in Fig. 4. It can be declared that more and more ordinary people take part in this topic of iPone5, they edged into the influential lists depend on their good performance. In our dataset, there are 5674 verified users (marked with "V") all told, accounting for 1.3% of total users crawled. Among the top 0.5% (=2208), almost 70% (=1546) influentials are verified users, and the rest 662 influentials are ordinary people non-marked with "V". Even though it comes to the top 5%, not all the verified users list on the influentials. The growth in the proportion of non-verified users in the top k lists implies that some ordinary users are experiencing the process from grassroots to stars as the topic reaches the hottest.

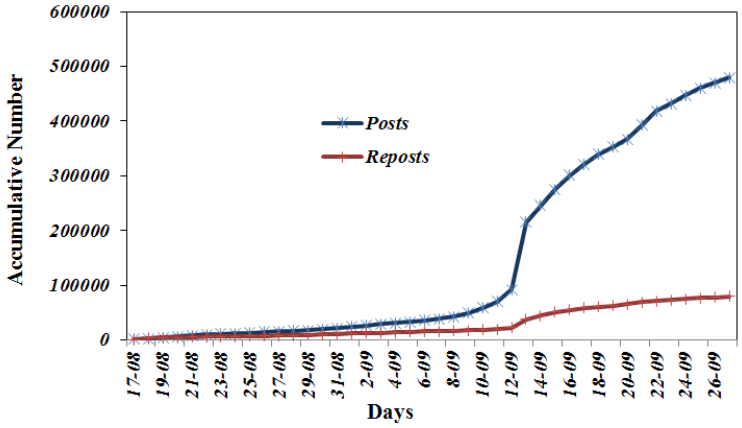

Figure 3. Accumulative number of posts and reposts during 6 weeks

In addition, based on our identification and prediction result, the users' expertise and the public interests will be further analyzed. For instance, user A mentioned in table 3 is an IT professional (judged by the verified user's information), whose insights are easily diffused. However, due to the drawbacks in "Posts", her/is followers gradually cancel their "following". As a grass root non-marked with "V", user B depends on high quality posts with relevance and professionalism (number of reposted) to obtain more and more followers. User C (a media representative shown in the verified information) is marked with "V" but not an expert on smart phone or the relative electronic products, and her/is posts under this topic seldom get reposted by the followers. Similar with A, user C's followers are dwindling. Even though in micro-blogging which is considered to be different from traditional blog, the content also plays an important role. Celebrity effect does not always work well.

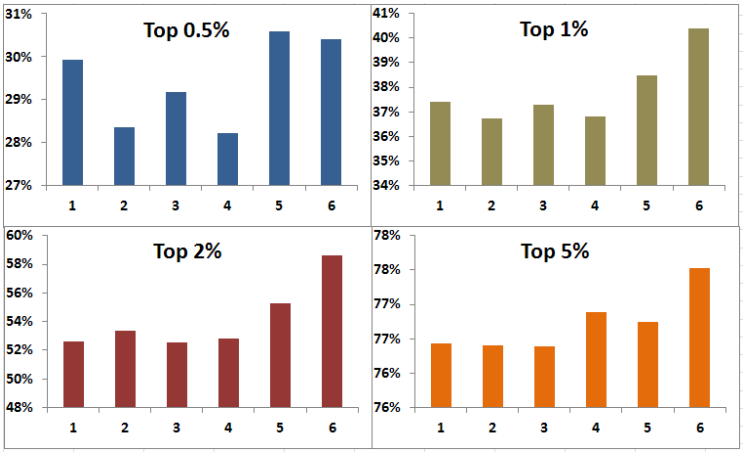

Figure 4. Ratio of users non-marked with "V" in identified influentials during 6 weeks

5 CONCLUSIONS AND FUTURE WORK

Different from the previous literatures just adopting single dimension or a few indicators to identify the most influential users over the whole network, this paper integrated the users' actual influential capability on one selected topic and the potential influence for information dissemination to discover and forecast the influentials under a certain topic. A novel ACQR framework is proposed to depict influential users' attributes, including activeness (A), centrality (C), quality of post (Q) and reputation (R), by taking activity data, social links and properties of senders and receivers into account. Based on the ACQR framework, a data mining method is introduced to identify and predict the influential users. In terms of qualitative comparison and quantitative evaluation, the empirical results reveal that our

ACQR framework and the data mining method by TOPSIS and SVMs (with polynomial and RBF kernels) have good performance in discovering and forecasting influential users under a specific topic (e.g. iPhone5) in micro-blogging.

Due to more attention to application of online influence in WoM marketing, the dynamic change process of users' influence and public interest were further analyzed in our research from longitudinal perspective with the topic development. Several findings are as follows: 1) At the preliminary stage, it is relative easy for diligent and competent grass-root stars to perform outstandingly. Yet, as the topic reaches the hottest, high centrality users with the relative expertise are more easily exert greater influence. 2) The level of expertise and active performance are very crucial for users' influence and their followers' loyalty. If a user overlooks any one of them, the public will switch their focus onto other influentials. 3) The growth in the proportion of non-verified users in the top k lists implies that some ordinary users are experiencing the process from grassroots to stars as the topic reaches the hottest. Our novel tools and findings can be utilized by sales managers to determine which kinds of influentials should be targeted in different periods of the online product marketing.

Here we also acknowledge several limitations of our research which need for future work. At first, circumscribed by the keyword in micro-blogging searching and the sampling time span, the datasets are only a small volume which cannot reflect the complete picture of the true condition in micro-blogging. Secondly, our TOPSIS method can be further improved by considering the weights of potential influence and actual influential capability on one selected topic. The third, according to the performance on different attributes and phases of topic development, the identified influential users should be further classified. The fourth, given the temporary nature of topics, it is more consequential to apply our framework and method in some selected fields. At last, the growth process of grassroots' influence (e.g. user B) is worthy of our tracking in the future.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 71273265, 70971130 and 71001103, part by Beijing Natural Science Foundation under Grant 9112009 and 9122013, part by Program for New Century Excellent Talents in University, part by Program for Excellent Talents in Beijing, and part by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China under Grant 13XNH168, 10XNJ026.

References

Agarwal, R.; Gupta, A. K. and Kraut, R. (2008). Editorial overview−The interplay between digital and social networks. Information Systems Research, 19(3), 243-252.

Amblee, N. and Bui, T. (2011). Harnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital microproducts. International Journal of Electronic Commerce, 16(2), 91-114.

Bakshy, E.; Hofman, J. M.; Mason, W. A. and Watts, D. J. (2011). Everyone's an influential: Quantifying influence on Twitter. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, China, pp. 65-74.

Brin, S. and Page, L. (1998). The anatomy of a large-Scale hyper textual web search engine. Computer Networks and ISDN Systems, 30(1-7), 107-117.

Cha, M.; Haddadi, H.; Benevenuto, F. and Gummadi, K.P. (2010). Measuring user influence in Twitter: The million follower fallacy. In 4th International Association for the Advancement of Artificial Intelligence on Weblogs and Social Media, Washington, DC, pp. 65–74.

Chang, C. C. and Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), p. 27.

Cheung, M. Y.; Luo, C.; Sia, C. L. and Chen, H. (2009). Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. International Journal of Electronic Commerce, 13(4), 9-38.
Cho, Y.; Hwang, J. and Lee, D. (2011). Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach. Technological Forecasting and Social Change, 79, 97-106.
Choi, S. M.; Ko, S. K. and Han, Y. S. (2012). A movie recommendation algorithm based on genre correlations. Expert Systems with Applications, 39, 8079-8085.
Culnan, M. J.; McHugh, P. J. and Zubillaga, J. I. (2010). How large U.S. companies can use Twitter and other social media to gain business value. MIS Quarterly Executive, 9(4), 243-259.
Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. Social Networks, 1(3), 215-239.
Goldsmith, R. E. and Horowitz, D. (2006). Measuring motivations for online opinion seeking. Journal of Interactive Advertising, 6(2), 1-16.
Han, Y. S. and Choi, S. M. (2013). Representative reviewers for Internet social media. Expert Systems with Applications, 40, 1274-1282.
Hanneman, R. and Riddle, M. (2005). Introduction to Social Network Methods. University of California, Riverside.
Heidemann, J.; Klier, M. and Probst, F. (2010). Identifying key users in online social networks: A PageRank based approach. In Proceedings of the 31th International Conference on Information Systems, p. 79.
Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. Human Communication Research, 36, 593-617.
Ju, Y. and Wang, A. (2012). Emergency alternative evaluation under group decision makers: A method of incorporating DS/AHP with extended TOPSIS. Expert Systems with Applications, 39(1), 1315-1323.
Jung, J. J. (2012). Computational reputation model based on selecting consensus choices: An empirical study on semantic wiki platform. Expert Systems with Applications, 39, 9002–9007.
Katz, E. and Lazarsfeld, P. (1955). Personal Influence: The Part of Played by People in the Flow of Mass Communications. Free Press.
Kwak, H.; Lee, C.; Park, H. and Moon, S. (2010). What is Twitter, a social network or a news media? In Proceedings of the 19th International Conference on World Wide Web, pp. 591-600.
Langville, A. N. and Meyer, C. D. (2004). Deeper Inside Pagerank. Internet Mathematics, 1(3), 335-380.
Li, F. and Du, T. C. (2011). Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs. Decision Support Systems, 51(1), 190-197.
Liang, T. P.; Ho, Y. T.; Li, Y. W. and Turban, E. (2011). What drives social commerce: The role of social support and relationship quality. International Journal of Electronic Commerce, 16, (2), 69-90.
Liebowitz, J. (2006). Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management. Auerbach Publications, Boca Raton, FL.
Lua, E. K.; Chen, R. and Cai, Z. (2011). Social trust and reputation in online social networks. 2011 IEEE 17th International Conference on Parallel and Distributed Systems, pp. 811-816.
Massoudi, K.; Tsagkias, M.; Rijke, D. M. and Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. Advances in Information Retrieval: 6611, 362-367.
O'Keefe, D. J. (2002). Persuasion: Theory and Research (2nd ed.). Thousand Oaks, CA: Sage.
Olson, D. L. and Delen, D. (2008). Advanced Data Mining Techniques, Springer Verlag.
Park, D.; Lee, J. and Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. International Journal of Electronic Commerce, 11(4), 125-148.

Resnick, P.; Kuwabara, K.; Zeckhauser, R. and Friedman, E. (2000). Reputation systems. Communications of the ACM, 43(12), 45-48.
Rogers, E. M. (1995). The Diffusion of Innovation (4th ed.). New York: Free Press.
Schölkopf, B.; Smola, A. J.; Williamson, R. C. and Bartlett, P. L. (2000). New support vector algorithms. Neural Computation, 12(5), 1207-1245.
Song, X.; Chi, Y.; Hino, K. and Tseng, B. (2007). Identifying opinion leaders in the blogosphere. In Proceedings of the sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisboa, Portugal, pp. 6-8.
Stephen, A. T. and Toubia, O. (2010). Driving value from social commerce networks. Journal of Marketing Research, 47(2), 215-228.
Vapnik, V. (1995). The Nature of Statistical Learning Theory, Springer Verlag, New York.
Varol, Y.; Oztop, H. F.; Koca, A. and Avci, E. (2009). Forecasting of entropy production due to buoyant convection using support vector machines (SVM) in a partially cooled square cross-sectional room. Expert Systems with Applications, 36 (3, Part 2), 5813-5821.
Venkatraman, M. P. (1989). Opinion leaders, adopters, and communicative adopters: a role analysis. Psychology and Marketing, 6(1), 51-68.
Vosecky, J.; Leung, K. and Ng, W. (2012). Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links. Database Systems for Advanced Applications, Springer. 7238, 397-413.
Wasserman, S. and Faust, K. (1994). Social Network Analysis: Methods and Applications, Cambridge University Press, New York.
Weerkamp, W. and Rijke, M. (2008). Credibility improves topical blog post retrieval. In Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 923-931.
Wu, S.; Hofman, J. M.; Mason, W. A. and Watts, D. J. (2011). Who says what to whom on Twitter. In Proceedings of the 20th International Conference on World Wide Web, pp. 705-714.
Wu, X. and Wang J. (2012). Micro-blog in China: Identify influential users and automatically classify posts on Sina micro-blog. Journal of Ambient Intelligence and Humanized Computing, 1-13.
Xiong, L. and Liu, L. (2004). Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. IEEE Transactions on Knowledge and Data Engineering, 16(7), 843-857..
Yang, C.C. and Ng, T.D. (2007). Terrorism and crime related weblog social network: link, content analysis and information visualization. In Proceedings of 2007 IEEE International Conference on Intelligence and Security Informatics, pp. 55-58.
Zhang, M.; Sun,C. and Liu, W. (2011). Identifying influential users of micro-blogging services: A dynamic action-based network approach. In proceeding of Pacific Asia Conference on Information Systems 2011, p. 223.