CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
*Concurrency Computation: Practice and Experience.* 2018; https://doi.org/10.1002/cpe.5029

# Finding Influential Users of Web Event in Social Media

Qichen Ma, Xiangfeng Luo*, Hai Zhuge*
*School of Computing Engineering and Science, Shanghai University Shanghai, China*
*Guangzhou University, Guangzhou, China*
*System Analytics Research Institute, Aston University, UK*

## SUMMARY

**Different users in social media have different influences to the evolution of a web event. Finding influential users could benefit recommendation, public opinion analysis, product marketing, etc. However, most of the existing methods are only based on social networks (e.g., user follower network or user behavior network), while the role of content discussed by users in social media is often unrecognized. The web event evolves with both user behaviors and semantic information. This paper proposes an approach to find influential users by extracting user behavior network and keyword association link network, and then uses PageRank and HITS to find the influence of users. Experiments on the real-world datasets show the effectiveness of the proposed approach.**

KEY WORDS: *Influential user, Social media, Social network, Web event*

*Corresponding authors:
*Xiangfeng Luo (email: luoxf@shu.edu.cn); Hai Zhuge (email: haizhuge@gmail.com).
All affiliations are equal.

## 1. INTRODUCTION

With the rapid development of internet technology, online social media has become the main platform of information dissemination. Compared to the traditional mass media, online social networks have fewer constraints. For example, people can not only obtain instant information from a range of different social platforms, such as Facebook, Twitter, Snapchat, etc., but also freely share their texts, pictures, videos, and repost the interested information. Merchants advertise their merchandise, terrorists spread illegal information, idols propaganda their works, which makes network information so volume, heterogeneous, and complicated that online social network attracts an increasing number of researchers studying it.

Network information is contagious. For a web event, if the users are very concerned about it and continue to publish microblogs and share their opinions, it will eventually lead to that web event breaking out. On the contrary, if netizens are indifferent to that web event, it may lead to the web event decline and will not cause social impact. We clearly remember the impact of the event ISIS killing the hostages, because the terrorist organization promoted terrorism through social network, post provocative article to delude netizens to join them, and post horror video to make social panics. Also, netizens condemn them and fear them on social network. Like these web events that pollute the environment of social network. Hence, user social influence analysis is a very necessary research direction for finding influential users and controlling the evolution of web events. It also can be applied to viral marketing [1, 2], recommender systems [3, 4], etc.

There is a large number of excellent works on social influence analysis. In the mid of nineteenth century, there already had many studies on social influence of real society. Katz et al [5] studied the pulse of voters during the presidential election, Deutsch et al [6] found the effect of normative social influence, and Granovetter et al [7] studied the strength of social weak ties. Until now, with the development of social media, not only the relationships in real society are mapped to the Internet, but also more and more relationships between strangers are built. The new emerging platforms are more attractive for social influence analysis. [8, 9] found influential users by statistical. Eigenvector centrality [10], pagerank [11], and katz score [12] were also used to compute social influence. [13, 14] also studies the social influence in scientist groups and co-author network. And there are also other applications, such as information diffusion analysis [15, 16] and event detection [17].

Thereinto, much social influence analysis work so far have focused on Twitter, which is a news spreading medium. Twitter users can post and repost their interesting microblogs, and in the microblogs users can @username to get others attention to read their microblogs. Users can comment on other users' microblogs [8]. In our work, we regard repost, comment and @username as three user interaction behaviors. And we analyze social influence based on Weibo, which is Chinese largest social platform analogous to Twitter.

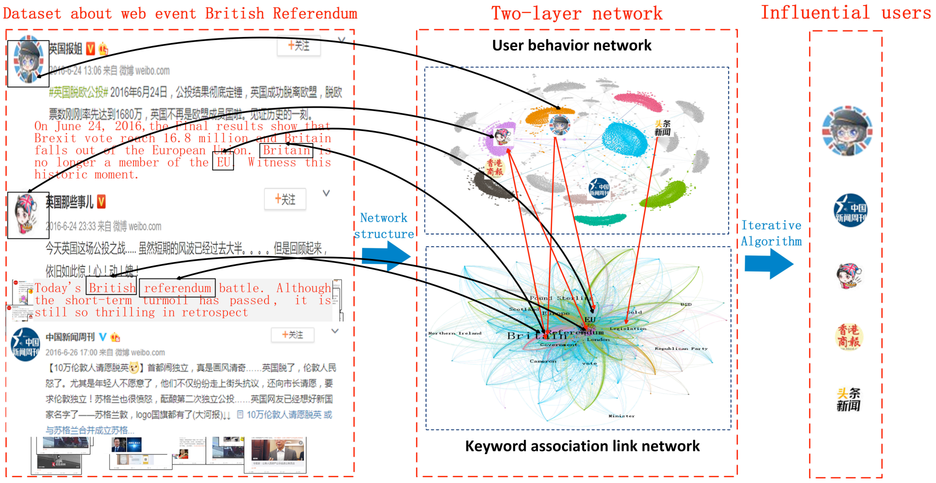
However, existing works are still with following limitations: 1) Not combining user behaviors and user posted semantic information. There is a mutual influence between user behaviors and user posted semantics in a social event. If a celebrity posts a microblog, this microblog may become famous and attract volume followers comment and repost this microblog because of the influence of the celebrity. Like Chinese famous actress Bingbing Fan, she posted a microblog with just one word ”We”, and so far 400 thousand users commented on this microblog and 500 thousand users reposted this microblog. A normal user posts a microblog with much novelty and if this microblog grabs a lot of attention, this microblog also increases this user's influence. So taking into account the semantic information is necessary for social influence analysis. 2) The changing of influence is not considered. A web event may go through latency period, outbreak period, and decline period during its evolution process. In different periods, the influential users may not be the same.

Our work is to find influential users in the evolution process of web event. Due to the constantly evolving web event, influential users also change in different periods of time. And capturing the new influential users in time is meaningful for many applications, such as we could adjust strategies for product marketing. Therefore, the influential users we find here is based on different stages of web event rather than based on the life course of web event. We build user behavior network and keyword association link network respectively, and then link the two networks. Not only nodes within the network will interact with each other in homogeneous network, but also the interactions exist across homogeneous networks. Hence, we propose an iterative method to compute the value of social analysis of users based the two networks and finally we obtain an influential user ranking. The overall process is shown in Figure 1. On the left-hand side of figure is our dataset of web event British Referendum, including microblogs content, comments on microblogs, reposted microblogs, @username. In the middle part of the figure shows two-layer network, i.e., user behavior network and keyword association link network. In user behavior network, there is a link from user A to user B if A comments on B's microblog, A forward B's microblog, or B @username of A. In keyword association link network, there will be a link between two keywords if two keywords co-occurring frequently in microblogs. On the right-hand side of figure is the ranking of influential users, which is the output of our algorithm. Our contributions are as follows:

(1) We do not merely consider the interactions between users' behaviors in homogeneous network, but also take account of the impact of semantic information represented by ALN-M (Association link network of microblog) on users. In most studies, social influence computing is just based on homogeneous network, such as follower network and user behavior network. In this paper, our method considers the factors of posted semantic information and link the two networks to form a heterogeneous network, and iteratively compute the influential users in the evolution process of web event.

(2) In the evolution process of web event, influential users are constantly changing. In different period of web event, influential users are not the same. Therefore, we build the time labeled two-layer network, and each time interval of web event has a corresponding two-layer network, so that we can find influential users in different time granularity. In this paper, we take one day as the time interval and find event-based influential users.

The remainder of this paper is organized as follows: Section 2 reviews related works. The details of the proposed approach are described in Section 3, which also introduces the building of the two networks and the link of the two networks. Section 4 presents the experimental setup and evaluation. Finally, we conclude this study in Section 5.



英国报姐 V
2016-6-24 13:06 来自 微博 weibo.com
#英国脱欧公投# 2016年6月24日，公投结果彻底定锤，英国成功脱离欧盟，脱欧票数刚刚率先达到1680万，英国不再是欧盟成员国啦。见证历史的一刻。

On June 24, 2016, the Final results show that Brexit vote reach 16.8 million and Britain falls out of the European Union. Britain is no longer a member of the EU. Witness this historic moment.

英国那些事儿 V
2016-6-24 23:33 来自 微博 weibo.com
今天英国这场公投之战..... 虽然短期的风波已经过去大半。。。。但是回顾起来，依旧如此惊！心！动！魄！

Today's British referendum battle. Although the short-term turmoil has passed, it is still so thrilling in retrospect

中国新闻周刊 V
2016-6-26 17:00 来自 微博 weibo.com
【10万伦敦人请愿脱英】首都闹独立，真是画风清奇......英国脱了，伦敦人民怒了。尤其是年轻人不愿意了，他们不仅纷纷走上街头抗议，还向市长请愿，要求伦敦独立！苏格兰也很愤怒，酝酿第二次独立公投......英国网友已经想好新国家名字了——苏格兰敦，logo国旗都有了(大河报)↓↓ 10万伦敦人请愿脱英 或与苏格兰合并成立苏格...

Figure 1. Our proposed framework for the computing of influential users.

## 2. RELATED WORK

The explosive growth of social media provides the diversity of social platforms, e.g. Twitter, Facebook, Flickr, Instagram, which become our mutual communication tools. Under such condition, research on social media analysis also become diversified. In this section, we introduce the background of related field, especially in the areas of social influence analysis.

Katz et al [5] studied the pulse of voters during the presidential election, they found each social stratum generated its own opinion leaders and these small number of opinion leaders affected most of the ordinary people. Deutsch et al [6] found normative social influence can be exerted to help make an individual be an individual and not just a mirror or puppet of the group. Granovetter et al [7] studied the strength of social weak ties. In general, strong ties guarantees the most communications between friends, but weak ties often play a key role in novel information(such as job information) spread.

Social analysis has various application in online social networks, such as emotion influence analysis [18, 19], information diffusion analysis [15, 16], viral marketing [1, 2], recommender systems [3, 4], event detection [17], and social influence analysis [8, 9, 20]. And there are also some commercial services, such as Klout [21], Kred [22], Twitalyzer [23]. Social influence analysis, which is attaching more and more attention by researchers, is also our research field.

Cha et al [8] used indegree, retweet, and mention, these three activities represent the influence of a user. And they have a conclusion that retweets and mentions are the activities more influence than the indegree activity. Pal et al [9] made a more detailed division on statistics. For example, they categorized tweets into three categories: original tweet, conversational tweet and repeated tweet, and for original tweet they extracted several statistical features (e.g. number of original tweets, number of links shared, and number of keyword hashtags used). Then they fund the influence users by these specified features. Embar et al [24] not only considered the comment of tweets, retweet, and the number of tweets posted by a user, but also took centrality of user in graph, response rate into account. They tracked user influences over multiple time-scales, so that they can identify both all-time influential users and recent influential users.

The use of topology of a network is a main method in traditional social influence computing. Closeness centrality [25, 26] needs to calculate the shortest path between all pair of nodes in the network, so the cost of computing is large and the advantage is able to measure the indirect influence of a node. Betweenness centrality [27] could find the connector between two communicties. Eigenvector centrality [10], PageRank [11], and katz score [12] also have a wide applications. And there are also improved method used in Twitter. Weng et al [20] proposed TwitterRank, an extension of PageRank algorithm, to measure the influence of users in Twitter. TwitterRank combined the event model and with the link structure to measure the influence. Feng et al [28] measured user influence in the Twitter network with modified k-shell decomposition algorithm, which is running faster and more effective at identifying a small group of users than the original algorithm.

Wang et al [29] took into account the two parts of social influence: the possibility of impact between two users, and the importance of each user. In addition, they emphasized the effects of common neighbor between two users. There are also some social analysis researches in other types of networks. Jiang et al [13] analyzed social influence of scientist groups, and the proposed model took two general group types (hierarchical and nonhierarchical) and two general collaboration situations (the independently multiplex collaboration relationships and the correlated multiplex collaboration relationships) into consideration. Tang et al [14] proposed Evental Affinity Propagation (TAP) to model the event-level social influence on large networks. And that model applied to the co-author network.

Most of methods study social influence only by user network, e.g. follower and followee network, user behavior network, which is too simple to find the influential users. Semantic information is also very considerable for social influence analysis and we should take into account. Zhuge et al [30, 31] studied the construction of semantic network and sentence ranking, which is instructive to our work. Xuan et al [32, 33] work on matrix also inspire us.

Table 1 Notations in this paper

| Symbol | Description |
|---|---|
| $infl(u)$ | User influence value |
| $ti$ | Time interval |
| $Ev$ | Web event |
| B | User behavior network |
| $ALN-M$ | Association link network of microblog |
| $w_{n_i,n_j}$ | Weight of link from node $n_i$ to node $n_j$ |
| A | bipartite graphs between user and keyword |
| $s(u_i)$ | Influential scores of users |
| $s(n_i)$ | Ranking scores of keywords |
| $\alpha$ | contribution of homogeneous network for the computation of user influence |
| $\beta$ | contributions of heterogeneous network for the computation of user influence |

## 3. OUR METHOD

Before introducing the model of user social influence computing, we present some basic definitions and algorithms used in our model and our viewpoint on defining social influence and notations. The notations are presented in table 1.

In the proposed model, the method of building user behavior network and user content network is firstly introduced; then, two homogeneous networks are associated as a heterogeneous network; Finally, we iteratively compute the social influence of users through two-layer network. Our method is based on two assumptions [34].

**Assumption 1**: A user has a high influential score if she/he is linked by many users and by few but very influential users, and a keyword has a high rank if it is linked by many other keywords and by few but highly ranked keywords.

In this assumption, links between users represent their interactions in behavior network, and links between keywords represent keywords co-occurrence in user content network. This assumption is similar to PageRank of which description is a page has high rank if a page has many backlinks and when a page has a few highly ranked backlinks.

**Assumption 2**: A user should have a high influential score if her/his microblog contains many highly ranked keyword, and a keyword would have a high rank if it appears in the microblog from the influential user.

This assumption is similar to HITS if we consider keywords and users as authorities and hubs respectively. We give the following intuitive description of assumption 2: if a celebrity posts a microblog, because of the influence of the celebrity, this microblog may become influential. And if a microblog is famous and has a lot of users to comment and repost this microblog, this microblog also increases this user’s influence. This is similar to the rules that important words appear in important paragraphs as discussed in [30, 31].

### 3.1 User social influence

In previous studies, social influence is defined in a number of different ways and influence user is also called opinion leader, influential user, influencers, etc. And social influence of users also has different explanations, for example, it can represent user influence in the entire social network, or in a particular field, or in a specific event; it also can be the capability to be opinion leader of an event. In this paper, we give our own definition of user social influence.

**Definition 1** Social influence of a user. User social influence is the ability to influence disseminate information in the evolution course of web event. User social influence is reflected in reposted and comment number of user’s microblogs, and the influence of microblogs, etc. This is similar to PageRank, a user has a highly influential score if she/he is linked by many users and by few but very influential users. The social influence of user is represented by a normalized value into the range of [0,1], and for a specific user, her(his) influence may change in different time interval. So the social influence of a user can be represent by,

$$infl(u) =< v, ti, Ev > \tag{1}$$

where v denotes the social influence value of user u, ti denotes the time interval of event Ev. If a user’s social influence value is 1, then this user could be of great ability to affect the development of web event.

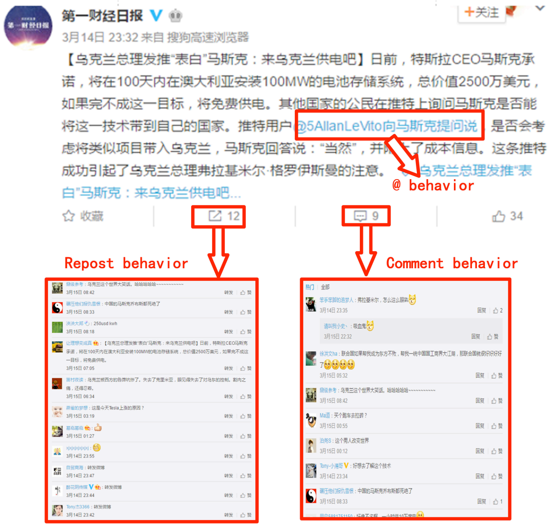
### 3.2 User behavior network

When an event occurs, online users can post their views by these newly emerging social media at any time and in anywhere. Users can post and repost their interesting microblogs, and in the microblogs users can @username to get others attention to read their microblogs. Users can also comment on other people's microblogs. We regard repost, comment and @username as three user interaction behaviors. If many celebrities participate in the discussion of the event, online users will produce large quantity of interaction behaviors. Under the interaction behaviors, information evolved in the process of propagation and development of the event may become uncontrollable. There is no doubt that the perception of interaction behaviors is crucial for evaluating the development situation of web event. Building user behavior network based on interaction behaviors and analyzing structure of behavior network is a valid way to perceive the interaction behaviors. According to [5], a small number of opinion leaders affect most of the ordinary people, so it is meaningful to find out the most influential users to control the information dissemination and the development of web event. First of all, we should build user behavior network [35] which is the startpoint to social influence analysis.

**Definition 2** User behavior network. User behavior network is constructed by user nodes and their interactive behaviors, including repost, comment, and @username behaviors. User interactive behaviors promote information dissemination and event evolution so that user behavior network can reflect user influence in a certain extent. User behavior network is modeled as a directed graph

$$B_{Ev} =< V, E, t > \tag{2}$$

where V denotes users in the network, E denotes edges which represents interaction behavior relationship which contains users, and t denotes the occurrence time of the relationship.


第一财经日报
3月14日 23:32 来自 搜狗高速浏览器
【乌克兰总理发推"表白"马斯克：来乌克兰供电吧】日前，特斯拉CEO马斯克承诺，将在100天内在澳大利亚安装100MW的电池存储系统，总价值2500万美元，如果完不成这一目标，将免费供电。其他国家的公民在推特上询问马斯克是否能将这一技术带到自己的国家。推特用户@5AllanLeVito向马斯克提问说，是否会考虑将类似项目带入乌克兰，马斯克回答说："当然"，并附上了成本信息。这条推特成功引起了乌克兰总理弗拉基米尔·格罗伊斯曼的注意。乌克兰总理发推"表白"马斯克：来乌克兰供电吧...

鼎级参考：乌克兰这个世界大笑话。哈哈哈哈哈~~~~~~~~~~~
3月15日 08:42

碾压他们报仇雪恨：中国的马斯克乔布斯都死绝了
3月15日 08:33

泱泱大邦：250usd kwh
3月15日 08:18

让理想变成真：【乌克兰总理发推"表白"马斯克：来乌克兰供电吧】日前，特斯拉CEO马斯克承诺，将在100天内在澳大利亚安装100MW的电池存储系统，总价值2500万美元，如果完不成这一目标，将免费供电。
3月15日 07:05

笛村夜读：乌克兰被西方的鲁莽坑惨了。失去了克里米亚，眼见得失去了对乌东的控制。割肉之痛，还得忍着。
3月15日 06:34

燕雀的梦想：这是今天Tesla上涨的原因？
3月15日 03:19

幂赢幂赢：👍
3月15日 01:27

xjxjxjxjxjxj：😁
3月14日 23:55

自贸商海：转发微博
3月14日 23:47

醉花阴传媒：转发微博
3月14日 23:44

Tony杰3366：转发微博
3月14日 23:42

笨手笨脚的追梦人：弗拉基米尔，怎么这么眼熟
3月14日 23:35

请叫我小史、：吸血鬼
3月15日 22:32

徐洪文ha：联合国如果帮我成为东方不败，帮我一统中国国工商界大江湖，那联合国就很好好好好了😳😳😳😳
3月15日 05:32

鼎级参考：乌克兰这个世界大笑话。哈哈哈哈哈~~~~~~~~~~~
3月15日 08:42

Ma道：买个跑车去拉砖？
3月15日 00:55

泊尧S：这个男人改变世界
3月15日 00:12

Tony-小潮哥：好想去了解这个技术
3月14日 23:34

碾压他们报仇雪恨：中国的马斯克乔布斯都死绝了
3月15日 08:33

用户5881751150：好像不多啊，一小时供10万度电

Figure 2. An illustration of three main user behaviors, e.g. @ behavior, repost behavior, and comment behavior.

As shown in Figure 2, for each microblog, we can see that it comments, reposts and @ another user, where all these interaction behaviors have their corresponding timestamps. If Bob comments a microblog posted by Tom at time t, then there is an edge from Bob to Tom in the user behavior network. According to the timestamp t of user behavior, we can get different user behavior networks in different time interval. In our user behavior network, each node denotes a user. When there is a link between two users, it means they have an interactive behavior which can be comment, repost or @ behavior. The weight of the link $w_{u_i,u_j}$ presents the strength of this interaction behavior between users, and $w_{u_i,u_j}$ is evaluated by the number of interactions.

### 3.3 User content network

When a hot event breaks out, explosive growth of microblogs take place in social network. Event-related content is discussed by large number of users through posts, reposts, and comments. A great deal of repeated semantic content cause cognitive complexity of users and hinder the research of event evolution.

Definition 3 Association link network of microblog(ALN-M). ALN-M, which is constructed by words and their associations, aims to establish associated relation among keywords from various microblogs. ALN-M is represented by,

$$ALN - M_{Ev} =< W, L, t > \tag{3}$$

where W is a set of keyword nodes, and L denotes a set of weighted links belong to W×W. Each short microblog contains some keywords and these keywords may co-occur in different microblogs, so the semantic content of web event can be expressed by the keywords and their association rules [30, 36], which eventually forms user content network, namely association link network of microblog.

Every link in ALN-M has a timestamp t. ALN-M is similar to user behavior network that can be divided into different networks according to timestamp. $w_{n_i,n_j}$, which is the weight of link between node i and node j, denotes the strength of association relations. And it is computed by,

$$w_{n_i,n_j} = supp(n_i, n_j) \tag{4}$$

where $w_{n_i,n_j}$ denotes the support of node i and node j. In ALN-M, each node represents a word. When there is a link between two words, it means these two words frequently co-occur in one microblog. Unlike ALN, ALN-M is improved by adapting the characteristic of short text of microblog. While mining the semantic information, we take one microblog as a transaction. And each microblogging or commentary has a time tag, so we can build ALN-M in different time interval.

In data processing phase, word segmentation is an indispensable part of it. We collect the network terms as user dictionary, which ensure the accuracy of word segmentation and the integrity of semantic information.

### 3.4 The association of two networks

After building the user behavior network and user content network, we get two sets of nodes. $\mathrm{U} = \{u_i | 1 \le i \le V\}$ and $N = \{n_j | 1 \le j \le W\}$ represent user set and keyword set, respectively. In order to associate the two networks, we build the bipartite graphs between user set U and keyword set N. The association rule is as follows: if user $u_i$ post a microblog and this microblog contains the keyword $n_i$, then we create a link $L_{ij}$ between $u_i$ and $n_j$. And the network is represented by,

$$\mathrm{A}_{Ev} =< V, W, L, t > \tag{5}$$

where V and W denote user set and keyword set, respectively. L denotes link set and t denotes the occurrence time of the link.

After the association of two networks, we get the two-layer network. This two-layer network can be represented by three matrixes,

$$B_{matrix} = \begin{bmatrix} n_{11} & \cdots & n_{1v} \\ \vdots & \ddots & \vdots \\ n_{v1} & \cdots & n_{vv} \end{bmatrix} \tag{6}$$

$$ALN - M_{matrix} = \begin{bmatrix} n_{11} & \cdots & n_{1w} \\ \vdots & \ddots & \vdots \\ n_{w1} & \cdots & n_{ww} \end{bmatrix} \tag{7}$$

$$A_{matrix} = \begin{bmatrix} n_{11} & \cdots & n_{1w} \\ \vdots & \ddots & \vdots \\ n_{v1} & \cdots & n_{vw} \end{bmatrix} \tag{8}$$

where $B_{matrix}$ denotes user behavior matrix, $ALN - M_{matrix}$ denotes ALN-M matrix, and $A_{matrix}$ denotes association matrix of behavior network and user content network. These three matrixes are normalized to make sure each value ranging from 0 to 1. And for $A_{matrix}$, each value is 0 or 1 as follows:

$$n_{ij} = \begin{cases} 1, & if\ there\ is\ a\ link\ between\ i\ and\ j \\ 0, & else \end{cases} \tag{9}$$

### 3.5 Iterative algorithm of two-layer network for user social influence computing

PageRank [37] computes the importance of nodes in homogeneous network: a page has high rank if the sum of the ranks of its backlinks is high. Similar to World Wide Web, user behavior network and ALN-M are also homogeneous networks. A user has high rank if the sum of the ranks of his/her neighbors is high, the same to the keyword. HITS [38] is another algorithm for analyzing homogeneous network and it divides the pages into hubs and authorities, which likes a bipartite graphs. Despite our two-layer network is a heterogeneous network, the user and the keyword can be considered as hubs and authorities. Therefore, combing the PageRank and HITS into one model is meaningful for social influence analysis [31, 34, 39].

We use two vectors $\mathrm{u} = [s(u_i)]_{v\times1}$ and $\mathrm{n} = [s(n_i)]_{w\times1}$ to denote the influential scores of users and ranking scores of keywords, respectively. In this paper, influential score $\mathrm{u} = [s(u_{i)}]_{v\times1}$ is the vector we want instead of keyword ranking vector $\mathrm{n} = [s(n_{i)}]_{w\times1}$, but keyword ranking vector is a very important part in our model.

According to assumption 1 and assumption 2, $\mathrm{u} = [s(u_{i)}]_{v\times1}$ can be computed as follows:

$$s(u_i) = B_{matrix} \cdot s(u_i) + A_{matrix} \cdot s(n_i) \tag{10}$$

and $\mathrm{n} = [s(n_i)]_{w\times1}$ can be computed as follows:

$$s(n_i) = ALN - M_{matrix} \cdot s(n_i) + A_{matrix} \cdot s(u_i) \tag{11}$$

Based on the above two formulas (10) and (11), we still consider two factors. One is that two parameters should be presented to indicate the contributions of networks for two vectors. And the other is that the computation of two vectors is an iterative process. Therefore, we improve the formulas above:

$$s(u_i)^{t+1} = \alpha B_{matrix} \cdot s(u_i)^t + \beta A_{matrix} \cdot s(n_i)^t \tag{12}$$

$$s(n_i)^{t+1} = \alpha ALN - M_{matrix} \cdot s(n_i)^t + \beta A_{matrix} \cdot s(u_i)^t \tag{13}$$

where $\alpha$ and $\beta$ denote the contributions of homogeneous and heterogeneous network for the computation of two vectors, and $\alpha + \beta = 1$. In this paper, we consider homogeneous and heterogeneous network and they have the equally influence for two vectors' computation, i.e., we set $\alpha = 0.5$ and $\beta = 0.5$, respectively.

When the difference of vector $s(u_i)^t$, $s(u_i)^{t+1}$ and $s(n_i)^t$, $s(n_i)^{t+1}$ are less than a certain threshold, the iteration is considered to achieve convergence. In this paper, the threshold is 0.0001.

The whole procedure is summarized in Algorithm 1.

**Algorithm 1** computation of user influence

```
Input: two-layer networks(user behavior network $B$, user content network $ALN-M$,)
in
Output: influential scores $u = [s(u_i)]_{v\times 1}$ and keyword ranking vector $n = [s(n_i)]_{w\times 1}$
initializing two vectors
for $T_i$ in time series $T$ of event $Ev$
    tag=true
    while tag do
        update $s(u_i)$ by Eq.12
        update $s(n_i)$ by Eq.13
        if |dis($s(u_i)^t$,$s(u_i)^{t+1}$) + dis($s(n_i)^t$,$s(n_i)^{t+1}$)|< 0.0001
            tag=false
        end if
    end while
end for
```

## 4. EXPERIMENTS

In this section, we introduce the datasets we used and discuss the evaluation of our model by two metrics.

### 4.1 Datasets

Our datasets are extracted from Sina Weibo(weibo.com) which is a biggest social platform in China analogous to Twitter. We crawled microblogs on six events with totally 18 thousand microblogs, 230 thousand reposts and 115 thousand comments. For each event, we crawled the microblogs for three days. More detailed information about datasets are listed in the table 2.

Table 2 Details of dataset in our experiment

| Topic | Microblog number | Comment number | forward number |
|---|---|---|---|
| Shanghai Disneyland | 1,496 | 7,349 | 18,497 |
| British Referendum | 8,871 | 45,593 | 81,047 |

| | | | |
|---|---|---|---|
| Hangzhou G20 Summit | 813 | 1,947 | 5,460 |
| Samsung Note7 exploded | 1,734 | 23,251 | 46,642 |
| New IPhone released | 3,129 | 29,915 | 47,119 |
| US election | 2,063 | 9,809 | 34,125 |
| Total | 18,106 | 117,864 | 232,890 |

### 4.2 Evaluation metrics

At present, there is no general data set for social influence analysis, and different studies focus on different aspects of social influence, e.g. some study the social influence in a given domain, some study the social influences through scientists' collaboration on scientific activities. These two points increase the difficulty of finding a good comparison method. At last, we choose [24] as our contrast method, which is similar to our study, e.g. finding influential users in the evolution process of web event through microblogging short text, updating influence score at real-time.

To evaluate our method, we consider two metrics:

1. **Metric 1**: influential user ranking. Comparing the ranking of influential users of two methods. We select top 10, 20, 30, 40, 50 influential users for experimental analysis, respectively.

2. **Metric 2**: semantic information measuring. We want to find the influential users in the evolution process of web event, so it is necessary to consider the semantic information of microblogs of influential users. If a user posts a microblog, which has a lot of comments and forward and the content is so important that affect the evolution of the event as well. Then, this kind of user is the influential user we want to find. Herein, we propose semantic coverage and semantic contribution to measure the semantic information of influential users. The following equations are the proposed metrics:

$$cov(u_1 \dots u_k) = \frac{\sum_{i=1}^{k} \emptyset_i}{\sum_{j=1}^{N} \emptyset_j} \tag{14}$$

$$cont(u_1 \dots u_k) = 1 - \frac{\sum_{i=1}^{l} \emptyset_i}{\sum_{j=1}^{N} \emptyset_j} \tag{15}$$

Where $cov(u_1 \dots u_k)$ denotes semantic coverage of top-k users, $\emptyset_i$ denotes semantic information of $u_i$ computed by [40], N denotes user set. $cont(u_1 \dots u_k)$ denotes semantic contribution of top-k users, and $k \cup l = N, k \cap l = \emptyset$.

### 4.3 Experimental results and discussions

We achieved the method of contrast test, taking into account the number of microblog, comment, and forward of users, and PageRank value of users in behavior network, etc. User influential score will be obtained by contrast method. Due to the different calculation methods, we just compare the user influential ranking instead of comparing user influential score.

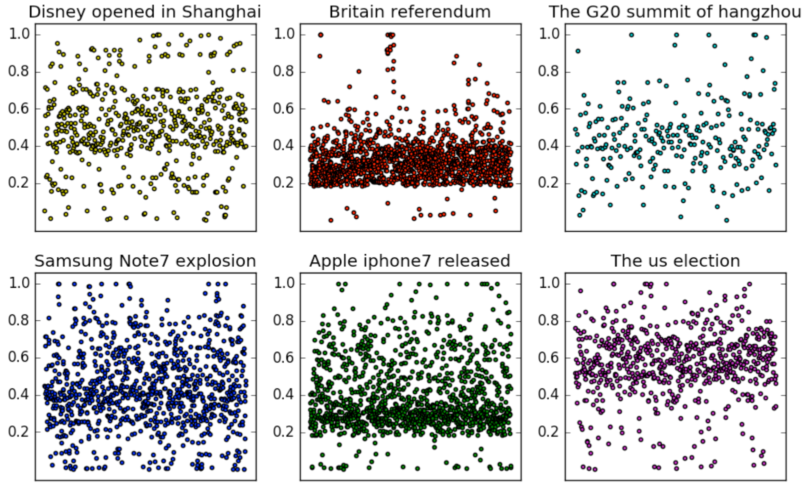


Figure 3. Visualization of the user score in six events.

In figure 3 we plot user score in six events presented by six subfigures. Each point in the plot denotes the user and his/her corresponding influential score in y-axis. We can see most of influential score range from 0.2 to 0.7, and few users get the score of over 0.9 or less than 0.1 point. In other words, very few users have a very high social influence or have no social influence.

### 4.4 Ranking of influential user

In order to compare the user ranking of the two methods, we select top 10, 20, 30, 40, 50 users respectively to compute the number of overlapping user. As shown in the table, there are only three in top ten users, which less than the half. But with the increase of the influential users, the proportion of overlapping users increased. When we select top-50 influential users of each method, there are 29 overlapped users. The detailed influential user ranking of event British Referendum is showed in appendix.

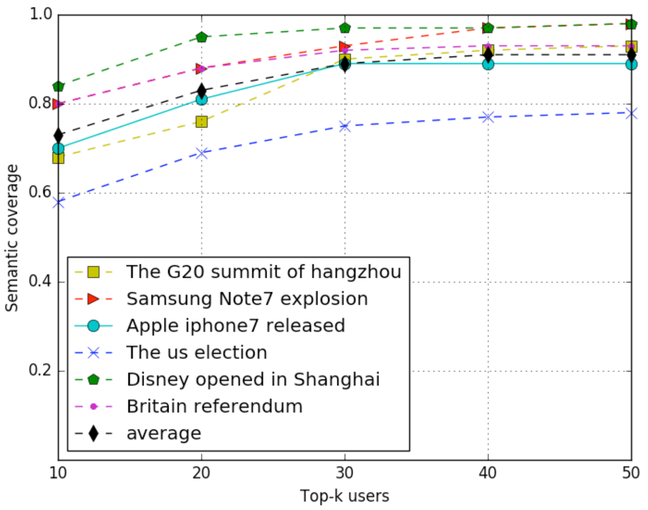
Table 3. The average overlapped number of top-k users of six events in different time interval

| Top-k users | Number of overlapped user |
|---|---|
| 10 | 3 |
| 20 | 7 |
| 30 | 14 |
| 40 | 22 |
| 50 | 29 |

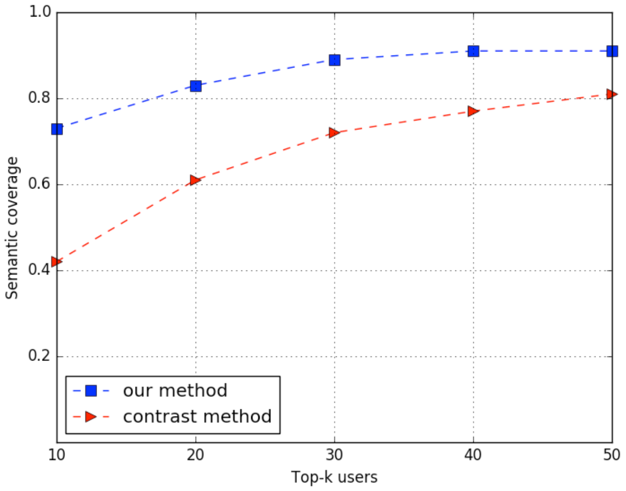
The results indicates that most of the influential users selected by the two methods are the same, and the difference is the ranking of influential users.

### 4.5 Semantic information measuring

**Semantic coverage**
According to the influential user ranking of two methods, we can see that the more user we choose, the higher the proportion of the overlapping user. In next, we measure the two methods in the aspect of semantic coverage. We want to find the influential users in the evolution process of web event, so that the semantic information of microblog posted by user is another essential measure for evaluating influential users. When a user is very popular and semantic coverage of his/her microblogs is high at the same time, and then this user is the influential user can affect the evolution of web event.

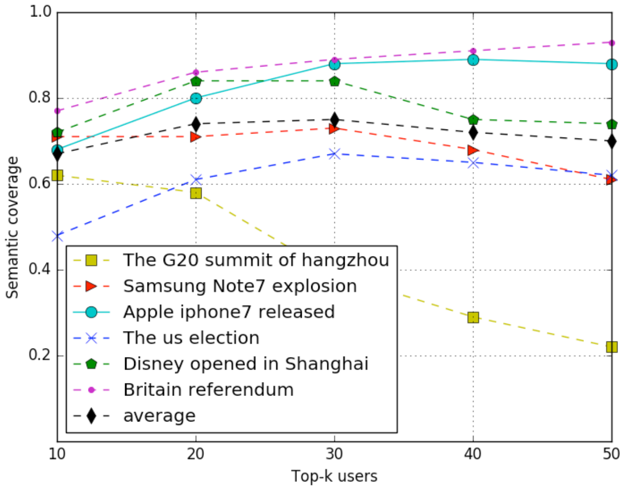


(a) Semantic coverage of six events and average

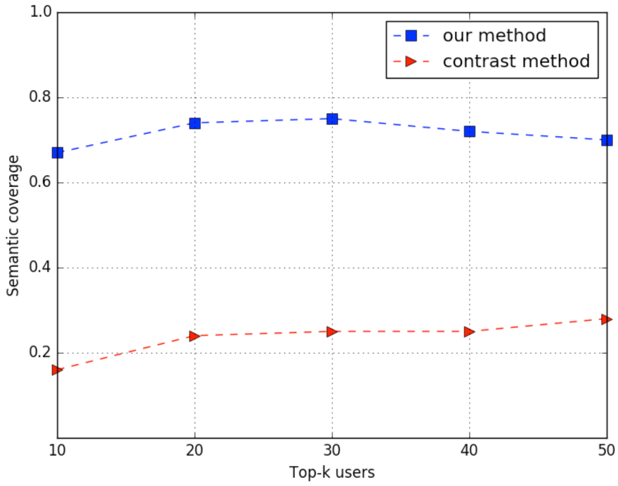


(b) Average semantic coverage of two methods
Figure 4. Semantic coverage of top-k users

Figure 4(a) shows the semantic coverage of six web events and their average semantic coverage (the black line in figure 4(a)) of our method. It can be seen that the x-axis is the number of user(top-k), y- axis represents the percentage of semantic coverage. The semantic coverage of top10 user is already up to 0.7, and with the increase of users, the curves become converge. Figure 4(b) shows the average semantic coverage of six web events of two methods. From the graph we can see that the semantic coverage of our method is higher than the contrast method in different coordinates, and the rate of semantic convergence of our method is faster.



(a) Semantic coverage of top-k users of six events and their average excluding the overlap users

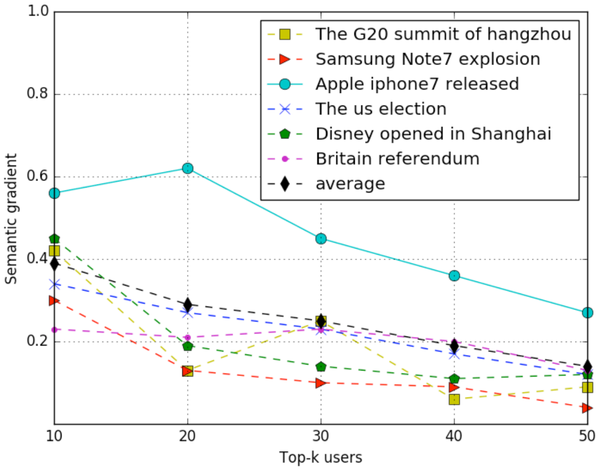


(b) Average semantic coverage of two methods excluding their overlap users

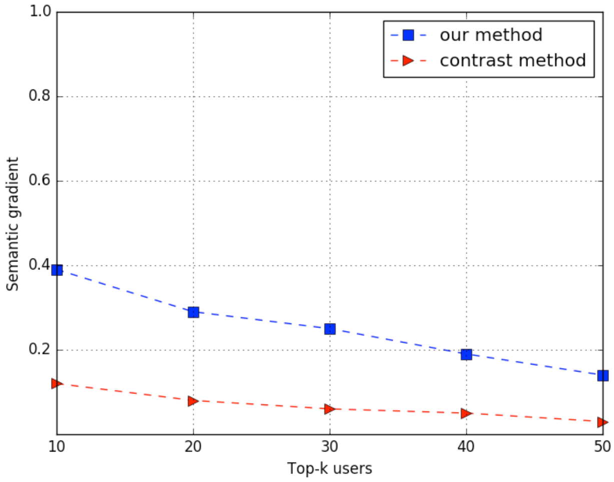
Figure 5. Semantic coverage of top-k users excluding the overlap users of two methods

Next, we remove the overlapping users in top-k ranking of two methods, which means the left influential users of each method are unique. And then we compare the semantic coverage of unique users of each method. Figure 5(a), shows the semantic coverage of six web events and average semantic coverage of our method with the overlapping users removed. Figure 5(b) shows the average semantic coverage of six web events of two method in the condition that the overlapping users have been removed. It can be seen that the semantic coverage of our method is significantly higher than the contrast method when the overlapping users are removed, which demonstrates that influential users selected by our method have richer semantic information.

**Semantic contribution**



(a) Semantic contribution of unique users of six events and their average in overall semantic



(b) Average semantic contribution of unique users of two methods in overall semantic

Figure 6. Semantic contribution of top-k users unique users of each method in overall semantic

Then we compute the semantic contribution of unique users of each method to the overall semantic of top k influential users. Figure 6(a) shows the semantic contribution of six web events and their average semantic contribution of unique users of our method. Figure 6(b) shows average semantic contribution of six web events of two methods. That shows the unique users selected by our method have higher semantic contribution for over semantic, which also demonstrate that influential users selected by our method have richer semantic information.

## 5 APPLICATION

The work of finding influential users can be valuable in many applications. First, discovering the influential users will help for viral marketing. “ZEALER China”, who has million followers in Weibo, is a well-known Weibo user. He often post technology-related microblogs, especially about mobile phone test, experience introduction about the new released phone, etc. “ZEALER China” is also an influential user in the event new Iphone released. Therefore, if mobile phone manufacturers is releasing a new phone and want to advertise through social media, influential users of tech events like “ZEALER China” is their best choice. But in this paper, we just find the global influential users of web event rather than influential users in specific populations. That is to say, we could not do targeted advertising in teenager community or adult community. Besides, influential user discovery also can be applied in intelligent friend recommendation. For example, if a Weibo user interested in sports, then influential users in sports events will be recommended to this Weibo user. Besides Moreover, identifying influential users can also help carding the information spread of web event. Due to the activity and participation in web event of influential users, influential users play a promoting role in the information spread course of web event. Therefore, identifying influential users could help grasp the information spread of web event.

## 6 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a two-layer network-based method to find the influential user in the evolution process of web event. The proposed method mainly consists of two procedures: the construction of two-layer network and iteratively find the influential users based on the constructed network. For the construction of two-layer network, the user behavior network has been constructed according to the user interaction behaviors and the user content network has been constructed according to the contents of the user's microblogging and comment respectively, and then two networks has been associated to form a two-layer network. The two networks are homogeneous networks and the associated network are heterogeneous network. The social influence of user is not only reflected by the interaction behaviors, but also affected by the semantic information of his/her microblogs. Therefore, for the procedure of iteratively calculation, the PageRank algorithm has been applied to homogeneous network and HITS algorithm has been applied to heterogeneous network. We have combined these two algorithms into our framework, and finally get an influential user ranking. Our model can be summed up as an important entity discovery model based on entity network and background knowledge network. At last we select six events to validate the effectiveness of our method. The experimental results show the good performance of the proposed method.

In the future work, we will study the evolution of user influence. The influence of a user will change in different stage of web event and in different web events. We hope we can get a deeper insight into mechanism of social influence. The study of user influence evolution will make social influence analysis appropriate for a widely application.

## ACKNOWLEDGEMENT

The research work reported in this paper was supported in part by the National Science Foundation of China under grant NO. 91746203, 61471232 and 61640212. This work was jointly supported by the Shanghai Science International Cooperation Project under grant NO.16550720400. Xiangfeng Luo and Hai Zhuge are the correspondence authors. All affiliations are equal in this work.

### REFERENCES

1. Kempe D, Kleinberg J M, Tardos É. Maximizing the Spread of Influence through a Social Network[J]. Theory of Computing, 2015, 11(4): 105-147.
2. Bhagat S, Goyal A, Lakshmanan L V S. Maximizing product adoption in social networks[C]//Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012: 603-612.
3. Shang S, Hui P, Kulkarni S R, et al. Wisdom of the crowd: Incorporating social influence in recommendation models[C]//Parallel and Distributed Systems (ICPADS), 2011 IEEE 17th International Conference on. IEEE, 2011: 835-840.
4. Song X, Tseng B L, Lin C Y, et al. Personalized recommendation driven by information flow[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 509-516.
5. Katz E, Lazarsfeld P F. Personal Influence, The part played by people in the flow of mass communications[M]. Transaction Publishers, 1966.
6. Deutsch M, Gerard H B. A study of normative and informational social influences upon individual judgment[J]. The journal of abnormal and social psychology, 1955, 51(3): 629.

7. Granovetter M S. The strength of weak ties[J]. American journal of sociology, 1973, 78(6): 1360-1380.
8. Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter: The million follower fallacy[J]. Icwsm, 2010, 10(10-17): 30.
9. Pal A, Counts S. Identifying topical authorities in microblogs[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 45-54.
10. Bonacich P, Lloyd P. Eigenvector-like measures of centrality for asymmetric relations[J]. Social networks, 2001, 23(3): 191-201.
11. Horowitz D, Kamvar S D. The anatomy of a large-scale social search engine[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 431-440.
12. Katz L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.
13. Jiang J, Shi P, An B, et al. Measuring the social influences of scientist groups based on multiple types of collaboration relations[J]. Information Processing & Management, 2017, 53(1): 1-20.
14. Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 807-816.
15. Yang J, Leskovec J. Modeling information diffusion in implicit networks[C]//Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 599-608.
16. Matsubara Y, Sakurai Y, Prakash B A, et al. Rise and fall patterns of information diffusion: model and implications[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 6-14.
17. Weng J, Lee B S. Event detection in Twitter[J]. ICWSM, 2011, 11: 401-408.
18. Yang Y, Cui P, Zhu W, et al. User interest and social influence based emotion prediction for individuals[C]//Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013: 785-788.
19. Wang X, Jia J, Tang J, et al. Modeling emotion influence in image social networks[J]. IEEE Transactions on Affective Computing, 2015, 6(3): 286-297.
20. Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential Twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270.
21. https://klout.com/
22. https://kred.com/
23. http://twitalyzer.com/
24. Embar V R, Bhattacharya I, Pandit V, et al. Online topic-based social influence analysis for the wimbledon championships[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1759-1768.
25. Hakimi S L. Optimum locations of switching centers and the absolute centers and medians of a graph[J]. Operations research, 1964, 12(3): 450-459.
26. Sabidussi G. The centrality index of a graph[J]. Psychometrika, 1966, 31(4): 581-603.
27. Freeman L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977: 35-41.
28. Feng P. Measuring user influence on Twitter using modified k-shell decomposition[J]. 2011.
29. Wang G, Jiang W, Wu J, et al. Fine-grained feature-based social influence evaluation in online social networks[J]. IEEE Transactions on parallel and distributed systems, 2014, 25(9): 2286-2296.
30. Zhuge H. Multi-dimensional summarization in cyber-physical society[M]. Morgan Kaufmann, 2016.
31. Tian J, Cao M, Liu J, et al. Sentence Ranking with the Semantic Link Network in Scientific Paper[C]//Semantics, Knowledge and Grids (SKG), 2015 11th International Conference on. IEEE, 2015: 73-80.
32. Xuan J, Lu J, Zhang G, et al. A Bayesian nonparametric model for multi-label learning[J]. Machine Learning, 2017, 106(11): 1787-1815.
33. Xuan J, Lu J, Zhang G, et al. Doubly Nonparametric Sparse Nonnegative Matrix Factorization Based on Dependent Indian Buffet Processes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017.
34. Wan X, Yang J, Xiao J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction[C]//ACL. 2007, 7: 552-559.
35. Ma Q, Luo X, Luo Y. Information Entropy Based the Stability Measure of User Behaviour Network in Microblog[C]//Semantics, Knowledge and Grids (SKG), 2014 10th International Conference on. IEEE, 2014: 67-74.
36. Luo X, Xu Z, Yu J, et al. Building association link network for semantic link on web resources[J]. IEEE transactions on automation science and engineering, 2011, 8(3): 482-494.

37. Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.
38. Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604-632.
39. Zhu Z, Su J, Kong L. Measuring influence in online social network based on the user-content bipartite graph[J]. Computers in Human Behavior, 2015, 52: 184-189.
40. Luo X, Zhang J, Ye F, et al. Power series representation model of text knowledge based on human concept learning[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 44(1): 86-102

**APPENDIX:**

Influential users in event British Referendum on June 25 2016

| nickname | value | rank |
|---|---|---|
| 华尔街日报中文网 | 1 | 1 |
| 环球市场播报 | 0.937376 | 2 |
| 央视财经 | 0.84001 | 3 |
| 英国报姐 | 0.738961 | 4 |
| 财经网 | 0.719898 | 5 |
| 汇通财经 FX678 | 0.694709 | 6 |
| IT 之家 | 0.649651 | 7 |
| Wind 资讯 | 0.643993 | 8 |
| 香港商報網 | 0.640904 | 9 |
| MACD 波段操作王 | 0.636259 | 10 |

Influential users in event British Referendum on June 26 2016

| Nickname | value | rank |
|---|---|---|
| 环球市场播报 | 1 | 1 |
| 头条新闻 | 0.849617 | 2 |
| 英国报姐 | 0.649485 | 3 |
| 尹国明 | 0.628847 | 4 |
| 今晚报 | 0.605365 | 5 |
| 香港商報網 | 0.601348 | 6 |
| 人民日报海外版-海外网 | 0.58993 | 7 |
| 中国经营报 | 0.568288 | 8 |
| 新华国际 | 0.559279 | 9 |
| 西部商报 | 0.547324 | 10 |

Influential users in event British Referendum on June 27 2016

| nickname | value | rank |
|---|---|---|
| 环球市场播报 | 1 | 1 |
| FT 中文网 | 0.999022 | 2 |
| 华尔街日报中文网 | 0.921435 | 3 |
| 汇通财经 FX678 | 0.747072 | 4 |

| | | |
|---|---|---|
| 证券日报之声 | 0.669715 | 5 |
| 股通大亨 | 0.664634 | 6 |
| 投资界微博 | 0.595073 | 7 |
| 早报网 | 0.593667 | 8 |
| 北美新浪 | 0.568021 | 9 |
| 香港商報網 | 0.529355 | 10 |