International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 18-23

Measuring User Reputation on Twitter Using Page Rank Algorithm

*N.Abirami[1], K.Manohari [2]
[1]*(Department Of Computer Science, Theivanai Ammal College for Women, India)*
[2]*(Department Of Computer Science, Theivanai Ammal College for Women, India)*

Abstract: *In the recent days, In the recent days, The enlargements made in the web knowledge's have greater impact over the social network systems. A different social media system offers different variety of online platforms to analyze dissimilar categories of formation and disorderly data. Besides, the social users are also connected with each other, and hence, instantly, the events are updated. In a time period, lots of users may take different activities such as posting and retweeting at these social networking sites which enforces a challenging issue, furthermost valuable social events. Hence, the related news is extracted from the set of online news via ranking approach. This study proposes a novel ranking framework that arranges the social user communications and activity based on the users active performance to be calculated. Intuitively, the social user communications are observed. Relied on this, the active score is predicted for each social user. To consider the social communications in the method of interchange based on the retweet communication between Twitter users. Specially, here it refers to the mutual acceptance of each other's tweets between two users in the model of a retweet. Second, the regression evaluation is done for the scored active. By performing so, we can accomplish accurate results on the current events. Experimental evaluation is approved on the real-time application, Twitter which shares the varied events instantly that further to show the efficiency of the proposed PageRank algorithm.*

Keywords: *PageRank Algorithm, Regression, Social media, Twitter, user activity*

I. Introduction

The development of web technology, social media service has been provides at many online platforms. The social networking service smooth the building of social networks or social connections among users who, for instance, share interest, activities, background and physical connections. . Through such service, users could stay connected with each other and be informed of friends' behavior such as sharing at a platform, and consequently be impact by each other. For instance, in today's Twitter (one of the most popular social networking sites in China), a user can get the live updates about his connected friends' sharing pages and could another retweet or comment the sharing post. Within a time period, millions of users may take different behavior such as posting and retweeting at these social networking sites. One important problem is how to rank users based on their active with lot of data. An almost active ranking of users will add great insight for many applications in most online social media sites. For instance, online ads provides may make gets track for convey their ads via considered the ranked active of users; site operators may design good practices for online survey via averaging the ranking list. While it is very encouraging for many parties to provide an active ranking of users. We consider the social interaction in the notion of cooperation's based on the retweet communications between Twitter users. We can achieve specific results on the current events. Experimental analysis is drifting on the real-time application, Twitter which shares the different events instantly that assist to show the efficiency of the proposed algorithm. For instance, suppose one user has had many time communications with most of his friends in a time period, we may conclude different active of this user when most of his friends also have had many time communication in the same time period versus when most of his friends do not have had many communications. Second, as the scale of social networks increases, it becomes more challenging to rank the active of users because a large number of user may impact the active of an individual user. Thirdly, as the social medias in many online sites evolve over time, the active of users may also manage over time. Thus efficient methods are needed to dynamics obtain the effective of users at different times of period. Social user also connect with twitter website to share the information, knowledge and experience. We can achieve the user communications based to provide a active ranking of users. In the literature, researchers have made some achievement on ranking users in social media sites. For instance, a Twitter user ranking algorithm was proposed to identify authenticate users who often give useful information.The proposed algorithm mainly works based on the user-tweet graph, rather than the user-user social graph. an extension of PageRank algorithm named TwitterRank was developed to rank Twitter users based on their impact. They first build topic-specific relationship network among users, then apply the TwitterRank algorithm for ranking. In a modified K-shell decomposition algorithm is developed to measure the user impact in Twitter. Furthermore, some explicit measurements such as retweets and mentions are developed to measure and rank user impact in Twitter. However, most of these measurements quantify the impact in an isolated way, rather than in a collective way. Furthermore, the focus of these methods is on impact, which is still different from the active that we address in this paper. To this end, in this paper, we propose two types of node active ranking algorithms that analyze the active of all nodes in a collective way. First, for a node A that has many communications with his friends in a time period, if most of his friends do not have many communications with their friends, it is very likely that the node A has high active. Based on this intuition, we define two measurements to quantify the active level of each node and propose the first algorithm. Second, by exploiting the mutual dependency of active among all users within a social network, we propose the second algorithm that infers the active level of users in a Pagerank way. Complete the iteration, all nodes' measurements propagate through the network and affect each other.

Thus the second algorithm is able to collectively analyze the active score of all nodes by considering the whole network. Furthermore, upon our in-depth understanding about user active, we propose an improved model to predict the active of users. The successful calculate results will another benefit many applications on social networking sites. Finally, we conduct intensive experiments on both user active ranking and prediction with two large-scale real world data sets. The experimental results demonstrate the effectiveness and efficiency of our methods. For instance, suppose one user has had many communications with most of his friends in a time period, we may conclude different active of this user when most of his friends also have had many communications in the same time period versus when most of his friends do not have had many communications. Second, as the scale of social networks increases, it becomes more challenging to rank the active of users because a large number of nodes (users) may impact the active of an individual node (user). Third, as the social networks in many online sites evolve over time, the active of users may also change over time. Thus effective methods are needed to dynamic obtain the active of users at several times.

II. Related Work

As the online social network such as Twitter grows fast these years, there are several works to study the authoritative or valuable users in the online social media for purpose such as maximize the spreading of impact [13] and growing marketing [14]. PageRank [15] and HITS [16], which are basically used to rank the web pages in the network which is made up of web pages, are commonly used in this new circumstance to rank the users in the network which is prepared of users. TwitterRank [17] extends PageRank by proposing a new aspect of the field of tweets. However, both PageRank and HITS are derivational based on their own convention. PageRank expects there is a surfer randomly visiting the web pages. HITS consider an educational scenario in which there are two aspects: authorization and hub. Due to the limitations of the convention, PageRank and HITS doesn't take into account the communications of users in social network, which may be the key point to discover the valuable of users in social network. Under this consideration, we proposed a quite different model, in which Twitter users' retweet activity are treated as reciprocal social activity. In this circumstance, the constitutional values of users are completed by the continuous communication between them. Further, other works which acknowledge the social features includes [18] and [19]. [18] is an application of HITS in the Twitter setting. It identifies authoritative users who are able to extended information immediately and influent others effectively. It introduces the "passive users" who are cautious to be determined. In this model, higher values in the ranking indicate that that particular user cans even influent most passive users. [19] Models active and sensibility in Twitter. In this model aggressive information circulation is due to viral users, aggressive items and exposed users. These models provide the other directions to calculate the users in the social network. In this section, related work can be grouped into two categories. The first category is most relevant that includes the work on measuring and ranking user in social network system. The second category is about the work on measuring user in network system. First, the Pagerank ranking algorithm in social network system has drawled a lot of attention in the research literature. The best known node ranking algorithms are Pagerank and HITS. Sergey Brin and Lawrence [1]proposed the pagerank to rank websites on the Internet. Pagerank is a link analysis algorithm which based on the directed graph (web graph).

The rank value indicates an importance of a particular node that represents the like-hood that users randomly clicking will arrive at any particular node. And, in[2], the authors presented two sampling algorithms for PageRank efficient approximation: Direct sampling and Adaptive sampling. Both methods sample the transition matrix and use the sample in PageRank computation. The hyper-link-induced topic search (HITS) was developed by Jon Kleinberg [3]. This algorithm is link analysis algorithms which rank the webpages. The authors presented a set of algorithms tools for rating and ranking the webpages from the directed graph of Internet environments. Furthermore, this work proposed a formulation of the notion of authority. PageRank/HITS is to find important websites that are linked to more different important websites and they do not consider the difference of nodes contribution to links at all, but in this paper we want to find those nodes that relatively contribute more to the interactions linked to them. However, Meeyoung Cha et al. [4] proposed a method to measure the user influence in Twitter used the directed links information, and present the comparison of three static measures of influence. However, they investigate the dynamics of user influence across topics and time which give a guide to the following research. Meanwhile, Yuanfeng Song and Wilfred Ng et al. [5] proposed a theoretical analysis on which frequent patterns are potentially effective for improving the performance of LTR and then propose an efficient method that selects frequent patterns for LTR. Also, Weng et al. [6] developed a Twittered rank algorithm based on PageRank to measure the influence of Twitterers. With a focus on both the topical similarity and the link structure into account, they proposed to measure the influence of users in Twitter with a topic-sensitive which means the influence of users vary in different topics. Besides, the user ranking based on the influence of user, in [7], [8], the expertise is considered as the ranking factor, both of them propose to measure the expertise level for user with the iteration information. There are other ranking factors for user ranking like [9] that rank the user with the authority score. In those user ranking algorithms, the Pagerank ideas are widely used in [6], [7] which pay more attention to the link analysis than content analysis. The algorithms based on link analysis were used for measuring the ranking factor that carried out as a research project which rank the transferred emails. In [10], [11] their work found that the ranking algorithms used link analyses have better results than the content methods. Nonetheless, the user rank is still under-explored with the influence and expertise score. Instead, in this paper, we focus on the ranking of user active level in social networks rather than focusing on measuring the influence or other factors. Measuring factor is not limited by the value of user, in [4], the work expand the value to influence which can better reflect the characteristics of user in social network system. Romero et al. [12] have developed the influence of user based on the information forwarding activity of user, the influence model is based on the concept of passivity and used the similar method to HITS to quantify the influence of users. In addition, in, the author computing the influence on Twitter by tracking the diffusion of URL from one user to another with three assignments. Furthermore, those predicting the individual user or URL influence by the regression tree model.

III. The Proposed Approach

To present the user ranking problem to rectify the problem used for Page rank ranking algorithm.

1. Page Rank Algorithm

To develop the page rank algorithm achieve for user ranking problem and spamming issues. The aim is to estimation the popularity, or the importance, of a webpage, based on the interdependence of the web. The explanation behind it is (i) a page with more incoming links is more important than a page with less incoming links, (ii) a page with a association from a page which is known to be of high usefulness is also usefulness. Page proposed a formula to calculate the Page Rank of a page A as stated below

: $PR(X) = (1-c) + c\,(PR(T1)/Y(T1) + ... + PR(Tm)/Y(Tm))$

where, PR(X) – Page Rank of page X, PR(T1) – Page Rank of pages Ti which link to page X,Y(T1) - number of outbound links on page T1, c – restraining factor which can be set between 0 and 1.

```
PageRank-Algorithm(G)
P0 ← e/n
k ← 1
repeat
(1 ) ; k -1
T
k P ← − d e + dA P
        k ← k + 1;
until ||Pk – Pk-1||1 < ε
return Pk
```

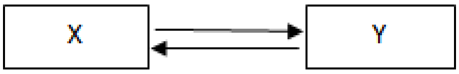

Fig.1example of page rank algorithm

Each page has one outgoing link. So that means Z (X) = 1 and Z (Y) = 1.

3.1 Page Rank Calculation

The PageRank of each page depends on the PR of the pages declaring to it. but won't know what PR those pages have till the pages disclosing to them have their PR computed and so on. Following a web graph with n nodes, where the nodes are pages and edges are hyperlinks. i) Assign each node an initial page rank.ii) Repeat until convergence. Calculate the pagerank of each node.

Table Page Rank Calculation

	Iteration 0	Iteration 1	Iteration 2	PageRank
P1	1/5	1/20	1/40	5
P2	1/5	5/20	3/40	4
P3	1/5	1/10	5/40	3
P4	1/5	5/20	15/40	2
P5	1/5	7/20	16/40	1

$R1(P_5) = 1/5 + 1/5 * 1/4 + 1/5 * ½ = 7/20$

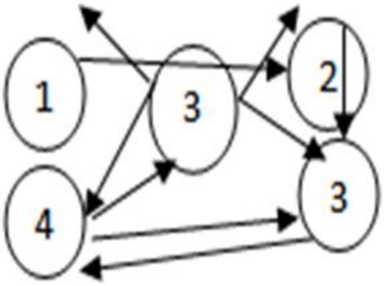

Fig.2 calculation of pagerank algorithm

3.2 Spam Issue With Pagerank

One of the main issues in PageRank is spamming. Spamdexing is a way used to affect search engine indexes. While effecting use of the PageRank algorithm, some web pages use spams in order to development the rank of certain pages. When a page receives link from a page which has better PageRank then the PageRank of current page will also developed. Hence they try to cause use of such links to develop the PageRank of their pages. This appearance the chance of sharing page rank in associated nodes. Inspired by this problem, researchers are attracting in a new area called Personalization. It is an area yet to be researches as the mathematical computation are hard. A best known approach is that of Trust Rank which only used the good pages. They rely on the extension of a set of good pages. The idea is to achievement trust from the good pages and iteratively goes to the outgoing links. But the problem that appears here is that interferer can leave bad links somewhere in the page. The solution for this is found by attaching a threshold value, if a page is below threshold then it is consulted as spam.

IV. Experiments

In this section, to evaluate the performances of the activity ranking algorithms with two real-world data sets.

4.1 The Experimental Setup

The experiments were achieved with two real-world network data sets. One of them is social media data set and the other one is academic networking data set. The social media data set was collected from a social networking system that is actually one of the biggest micro blog systems in China and has millions of active users per day. The academic networking data set was collected from the DBLP site which includes thousands of authors and articles. Compared with the DBLP data set, the microblog data is much more complex because it includes a variety of information. TABLE 2 shows the important information available in the twitter data.

Table 2 Attribute of Twitter data

Field name	Field Type	Description
Content	Char[140]	Content of this message
Msg ID	Unit32	A given number to mark this message
Author ID	Unit 32	The Id of the author user
Time	Unit32	The time that the message published
Reposted ID	Unit32	A list of user ID who have reposted this message
Retweet ID	Unit32	A list of user ID who have commented this message

Table 3 Statistics of education network data

Period	Users	Links
2010	282526	671076
2011	296196	727594
2012	318110	807677
2013	330893	869798

Table 4 Statistics of Twitter Analysis data

Period	Users	Links
20/06/2013 10:00	363415	611713
20/06/2013 11:00	503865	105263
20/06/2013 12:00	489517	894233
20/06/2013 13:00	398745	636564

In TABLE 3 and 4, shows some statistics of those networks. Note that, the users in TABLE 3 and 4 mean online users who have posted information to the social network system. Experimental Platform. All algorithms were implemented with Java and all experiments were conducted on a Windows 10 machine with i7-4700MQ CPU and 16.00GB Ram.2. Performance on Activity RankingValidation Metrics*:* Leverage it to evaluate the performance of different ranking algorithm..To evaluate the performance via using benchmark way.

Table 5 Auxiliary information on user data

Acronyms	Implications
Nof	Number of followers
Nop	Number of Posted
Nor	Number of Repost

.

Information available in our data , which are shown in TABLE 5 , to evaluate the ranking results of different methods.

V. Conclusion

A summary of user active ranking and estimate in social networking services such as twitter application. A different social media system offers different variety of online platforms to analyze dissimilar categories of formation and disorderly data. Besides, the social users are also connected with each other, and hence, instantly, the events are updated. The social users are also connected with each other, and thus, instantly, the events are updated. This study proposes a novel ranking framework that arranges the social user communications and activity based on the users active performance to be calculated. Intuitively, the social user communications are observed. Relied on this, the active score is predicted for each social user. To consider the social communications in the method of interchange based on the retweet communication between Twitter users. Specially, here it refers to the mutual acceptance of each other's tweets between two users in the model of a retweet. Second, the regression evaluation is done for the scored active. By performing so, we can accomplish accurate results on the current events. Experimental evaluation is approved on the real-time application, Twitter which shares the varied events instantly that further to show the efficiency of the introduced PageRank algorithm. To improve the page rank algorithm achieves for user ranking problem and spamming issues. The accurate results of both user active ranking and prediction could benefit many parties in different social networking services, e.g., a user active ranking list could help ads providers to better display their ads to active users and reach more audiences using precision, accuracy, root mean squared error and mean absolute error. This metric shows generally perfect results.

References

[1]. Alex Cozzi, and Byron Dom. Expertise identification using email communications. Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks, 56(18):3825– 3833, 2012.

[2]. Wenting Liu, Guangxia Li, and James Cheng. Fast pagerank approx-imation by adaptive sampling. Knowledge and Information Systems, 42(1):127–146, 2015.

[3]. Jon M Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632, 1999.

[4]. Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. ICWSM, 10(10-17):30, 2010.

[5]. Yuanfeng Song, Wilfred Ng, Kenneth Wai-Ting Leung, and Qiong Fang. Sfp-rank: significant frequent pattern analysis for effective ranking. Knowledge and Information Systems, 43(3):529–553, 2015.

[6]. Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining, pages 261– 270. ACM, 2010.

[7]. Jian Jiao, Jun Yan, Haibei Zhao, and Weiguo Fan. Expertrank: An expert user ranking algorithm in online communities. In New Trends in Infor-mation and Service Science, 2009. NISS'09. International Conference on, pages 674–679. IEEE, 2009.

[8]. Amin Omidvar, Mehdi Garakani, and Hamid R Safarpour. Context based user ranking in forums for expert finding using wordnet dictionary and social network analysis. Information Technology and Management, 15(1):51–63, 2014.

[9]. Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In Web Information Systems Engineering–WISE 2010, pages 240–253. Springer, 2010.

[10]. Christopher S Campbell, Paul P Maglio, Alex Cozzi, and Byron Dom. Expertise identification using email communications. In Proceedings of the twelfth international conference on Information and knowledge management, pages 528–531. ACM, 2003.

[11]. Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, page 20. Association for Computational Linguistics, 2004.

[12]. Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In Machine learning and knowledge discovery in databases, pages 18–33. Springer, 2011.

[13]. Kempe, D., Kleinberg, J., Tardos, _E.: Maximizing the spread of infuence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2003) 137-146.

[14]. Shakarian, P., Paulo, D.: Large social networks can be targeted for viral marketing with small seed sets. In: Advances in Social Networks Analysis and Mining(ASONAM), 2012 IEEE/ACM International Conference on, IEEE (2012) 1-8.

[15]. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. (1999).

[16]. Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., Kleinberg, J.: Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks and ISDN Systems 30(1) (1998) 65-74.

[17]. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: _Finding topic-sensitive infuential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, ACM (2010) 261-270.

[18]. Romero, D., Galuba, W., Asur, S., Huberman, B.: Infuence and passivity in social media. Machine Learning and Knowledge Discovery in Databases (2011) 18-33.

[19]. Hoang, T.A., Lim, E.P.: Virality and susceptibility in information difusions. In: Sixth International AAAI Conference on Weblogs and Social Media. (2012).