IOP Conf. Series: Journal of Physics: Conf. Series **803** (2017) 012089 doi:10.1088/1742-6596/803/1/012089

Estimation of social network user's influence in a given area of expertise

E E Luneva[1], **V S Zamyatina**[1], **P I Banokin**[1], **A A Yefremov** [1]

[1] Tomsk Polytechnic University, 30, Lenina ave., Tomsk, 634050, Russia

E-mail: lee@tpu.ru

Abstract. Nowadays social networks are frequently used to express personal opinion on a topic of interest. Some users' opinion has more informational influence than others do. These users are called influential users. There are services that allow evaluating how popular and influential users are; however, any information on evaluation methods is proprietary and represents know-how of such software services. Furthermore, most services could not provide extensive data on the influential users within the specified area of knowledge. This article proposes the method of evaluating a user influence index within a social network in a given area of expertise.

1. Introduction

Modern social networks are frequently used to express users' opinion on topics of their interest, events, products or services. However, some users' opinion could have an informational influence on opinion of the others. Data on the most influential users in a certain area could be used in various fields of expertise, ranging from marketing researches to political forecasting. There is a variety of services that allow evaluating user's popularity and influence. However, these services could not identify expert users within a specified area of expertise [1-4] and, what is more, they provide user's popularity index evaluation, performed by proprietary algorithms.

The purpose of this work is to develop the influence index evaluation method for users of a social networking service in a given area of expertise that would be able to provide stable results on incomplete (stream) data as well as on historical data.

2. Theoretical analysis

The review of literature on obtaining user's mutual information influence in social networks shows that this kind of problem could be described as a problem of identifying the set of "key players". The social graph is used as a mathematical model representing users and the structure of connections between them. According to Borgatti [5], there are two types of key players: KPP-NEG (Key Player Problem/Negative); KPP-POS (Key Player Problem/ Positive).

The problem of identifying influential users in a social network belongs to KPP-POS problem [5]. There are following methods and approaches to solution of "key players" problem [5-8]:

- An approach based on calculating the centrality indices of the social graph. The instability of the centrality indices is a significant drawback of this approach in solving a problem, stated in this paper.

- Method based on the usage of combinatorial optimization and a greedy algorithm. These algorithms require considerable computational power of a web-server.
- Approach based on the measuring the information entropy.
- Approach based on measuring the communication efficiency.

To identify influential users of a social network, the last approach was chosen due to its strong performance, relatively easy implementation and low demand for computing power. Efficiency E of certain social graph G is defined by the following equation [9]:

$$E(G) = \frac{\sum_{i \neq j \in G} \varepsilon_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}},$$

where N is the number of nodes in graph G, ε_{ij} is the communication efficiency, d_{ij} – the shortest path length between nodes i and j .

In order to calculate efficiency E, the nodes of the social graph should be successively excluded and the shortest distances between remaining nodes should be found. The decline of the efficiency indicator shows the importance of the excluded node, which means that it represents the influential user.

There is an unresolved issue of defining a rule of social graph construction and evaluating weights of its edges in order to calculate the efficiency indicator. In modern social networks, the information is transmitted via a subscription mechanism. If user *A* is a follower of user *B*, then user *A* receives all of the open publications of user *B*. However, the fact of subscription does not mean that user *A* is actively interested in user *B* publications on the specific topic; it also does not confirm the influence of the author's opinion on the opinion of the others.

Interest of one user in opinion of another can be identified by the following characteristics: the amount of comments to the publications, the amount of reposts or reposts with comments and the amount of mentions of another user. These characteristics could be considered as a measure of interest of followers in publications of some user. Therefore, the measure of interest of user *A* in user *B* publications can be described by function $f(x, y, z, l)$, where x – is the amount of reposts with comments, y – the amount of reposts, z – the amount of comments, l – the amount of user mentions. Obviously, x, y, z, l characteristics are not equivalent.

Concurring with the literature [8-10], we find that all of the aforementioned characteristics can be ranked as follows: x – a very high degree of interest, y – high degree of interest, the z – an average degree of interest, l – a certain degree of interest. The application of the analytic hierarchy process, developed by Saaty [11], allows obtaining numerical values of ranks in the form of weight coefficients and calculating values of $f(x, y, z, l)$. Tables 1 and 2 show the comparison matrix and the result of evaluating priority vector V that contains weight coefficients of characteristics x,y, z, l.

Table 1. The comparison matrix of user interest characteristics

Characteristics	*x*	*y*	*z*	*l*
x	1	9	7	5
y	1/9	1	1	2
z	1/7	1	1	1
l	1/5	1/2	1	1

Obviously, the authors with the greatest influence in social networks are those whose publications are frequently mentioned by the others, actively discussed and widely spread via a reposting mechanism. Thus, we propose to analyze the social graph with nodes connected by edges according to the following principle: there is a directed edge from node *B* to node *A,* if user *A* has commented, reposted or mentioned user *B* in his publications. The weight of the edge is equal to $1/f(x,y,z,l)$. Thus,

efficiency ε_{ij} between nodes i and j is inversely proportional to the shortest distance d_{ij} between these nodes.

Table 2. Priority vector V

Weight coefficients	Priority vector V
a	0.68
b	0.12
c	0.1
d	0.09

Therefore, the interest of user A in user B's publications can be determined by the following function:

$$f(x, y, z, l) = ax + by + cz + dl \qquad (1)$$

The previously described method can be illustrated with the following example: Figure 1(a) depicts a social graph consisting of six nodes. Weights of the edges were obtained by Equation 1. Efficiency E of the base graph is 0.39. Evaluated efficiency indices are represented in the plot Figure 1(b).

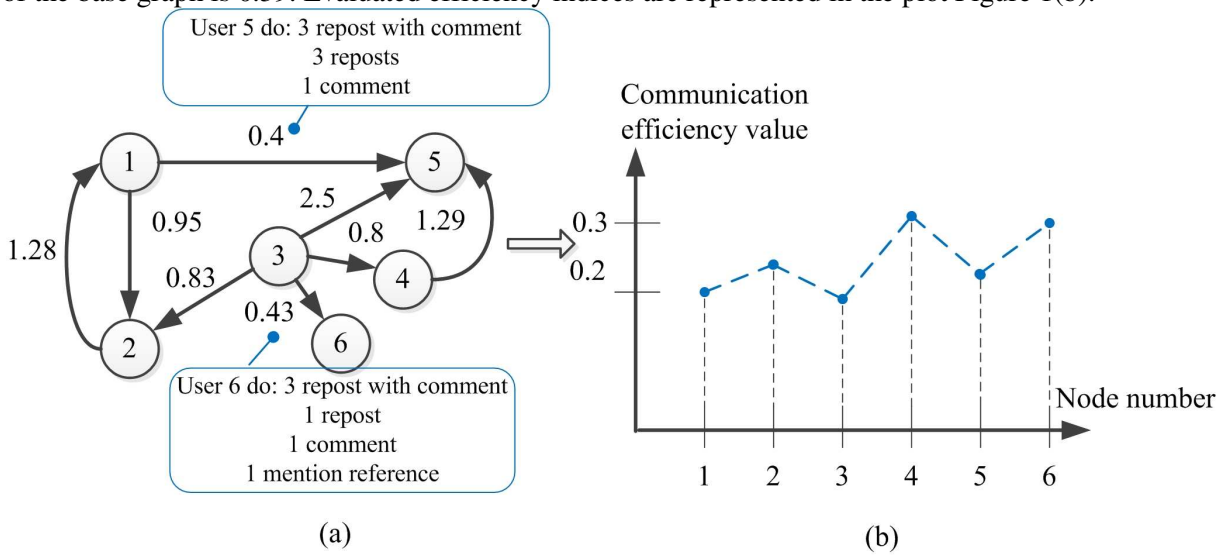

Figure 1. Steps of influential index evaluation: a – social graph, obtained from social network data; b – communication efficiency values of the social graph nodes.

Table 3 lists evaluated efficiency indices and their absolute deviation ΔE caused by successive removal of nodes and calculated by the following equation [9]:

$$\Delta E = E(G) - E(G - node_i),$$

where $E(G)$ - efficiency E of the base graph, $E(G - node_i)$ - the graph obtained by removing node i in graph G.

According to the data represented in Table 3, the minimal value of efficiency indicator emerges when node number three is excluded from the social graph. It means the user corresponding to the third node is the most influential.

Thus, the proposed method of user's influence evaluation consists in evaluating the communication efficiency indices for a weighted social graph, with edge weights defined by function of interest $f(x, y, z, l)$ according to equation 1.

Table 3. Communication efficiency values of social graph G

Node number	*1*	*2*	*3*	*4*	*5*	*6*
$E(G\text{-}node_i)$	0.20	0.24	0. 19	0.31	0.23	0.30
ΔE	0.19	0.15	0.20	0.08	0.07	0.09

3. Experimental analysis

A software component based on MVC (Model Viewer Controller) architectural pattern was developed for experimental analysis. The Microsoft Visual Studio development environment and programming language C# were used for the component implementation. Figure 2 demonstrates the software architecture.

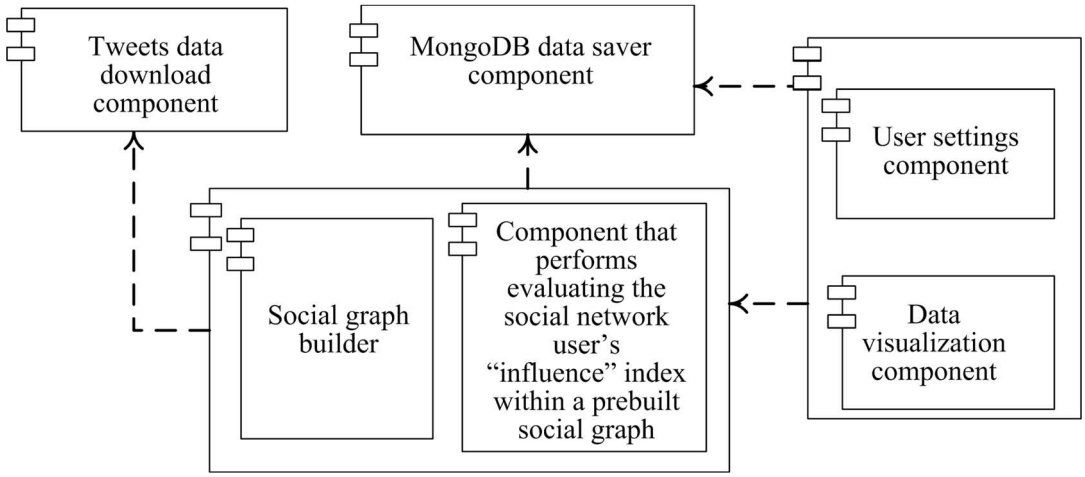

Figure 2. A component diagram of the developed experimental software component

AN operating scenario of the software is as follows. A user of the software application enters a key word, which is used for initial data search. The software application then sends a request to the Twitter social network service and retrieves the data related to the domain, selected by the keyword. This data is used for social graph creation. Based on the built graph, user social media influence analysis is performed. The analysis result is displayed to the user and stored in the database.

The first phase of an experimental analysis was performed on the samples prepared in advance, the second on the real data from the social network Twitter.

In order to validate the proposed method and the model, data samples were prepared with influential users defined in advance, and the greater interest regarding a chosen influential user was simulated. This experiment was conducted 50 times on networks of 50 nodes. The proposed method correctly determined the influential users in 84 % of cases. In 92 % of cases the influential users turned out to be in the top three of the most influential users (Figure 3).

Table 4 shows the accuracy of the method based on conducted experiments, resulting in the average relative error of 0.084%. Data in Table 4 is represented in an ascending order of the relative error values. For 42 experiments, values of E_{min} are equal to the values of E_{true} causing a zero-valued relative error, therefore they are not listed in the table.

In the second stage, preloaded samples from social network Twitter were used for the testing of the software component. The testing was conducted on social graphs, containing from 6 to 500 nodes. The experimental results proved operability of the social influence calculation method. The main drawback of the software component is its performance. This issue is noticeable when larger datasets are used. It could be eliminated by modification of the shortest distance calculation algorithm and dataset sorting algorithms [12, 13]. Additionally, a possibility of domain search by a set of keywords and hashtags is worth studying. This type of search would allow selecting a dataset more precisely in comparison to search by a single keyword.

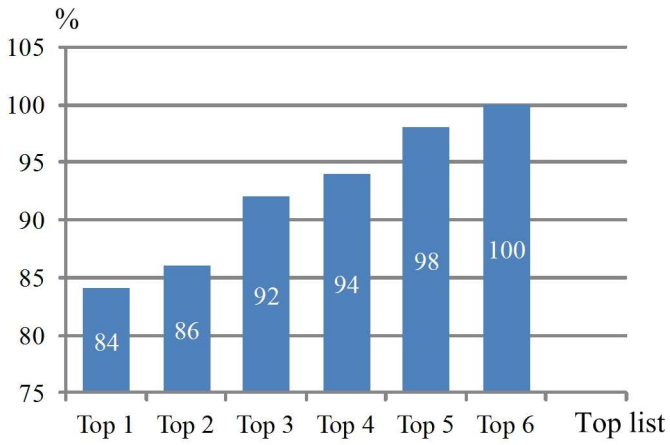

Figure 3. The result of influential user determination in conducted experiments

Table 4. Accuracy of the method in the conducted experiments

№	*1-42*	*43*	*44*	*45*	*46*	*47*	*48*	*49*	*50*
E_{min} [a]		0.198	0.560	0.173	0.252	0.245	0.170	0.780	0.277
E_{true} [b]		0.200	0.572	0.177	0.260	0.257	0.160	0.700	0.320
Relative error, *%*	0	1.000	2.098	2.260	3.077	4.669	6.250	11.429	13.438

[a] E_{min} – is the minimum value of the communication efficiency obtained during the experiment.
[b] E_{true} is – the value of the communication efficiency for the most influential pre-selected user.

The proposed method handles a social graph that is based on crisp quantitative data (i.e. number of reposts, comments, etc.). However, additional data on user's influence can be obtained by the processing messages in a natural language. For example, the sentiment response of a user's followers or readers can be processed, including their agreement or disagreement with the expressed opinion. Further work will be focused on the combined application of both crisp and fuzzy data, based on our research [14].

4. Conclusion

The evaluation method of social network user's influence was presented in the paper. The ability of source data selection by a domain area is a significant feature of the proposed method. The software component implementing the process of Twitter users' influence calculation was developed. The component testing carried out on the samples, prepared in advance, and preloaded datasets from the Twitter social network service, confirmed applicability of the proposed method and operability of the software component. Experiments on data, prepared in advance, show that in 84% of cases, the proposed method correctly determines the most influential users. Further research will be focused on increasing the accuracy of influential index evaluation by combining both crisp and fuzzy data from social network services. Also, in order to improve software component performance, it is necessary to implement more efficient graph algorithms (i.e. shortest path algorithms) together with parallel computations.

References

[1] Del Campo-AVila J, Moreno-Vergara N, Trella-López M 2013 *Proc. of 4th Int. Conf. on Ambient Systems, Networks and Technologies (Halifax)* **19** 437-444
[2] Cossu J-V, Dugue N, Labatut V 2015 *Proc. 2nd European Network Intelligence Conf (ENIC 2015)* (Karlskrona) **1** 89-30
[3] Savenko I I, Sukhodoev M S, Tsapko S G, Simonov P K 2015 *Proc. of Int. Siberian Conf. on*

Control and Communications (SIBCON 2015) (Omsk) 7147228
[4] Rao A, Spasojevic N, Li Z, Dsouza T 2015 *Proc. 2015 IEEE Int. Conf. on Big Data (Santa Clara USA)* 7364017 2282-2289
[5] Borgatti S P 2006 Identifying sets of key players in a network *Comput Math Organ Theory* **12(1)** 21–34
[6] Ortiz-Arroyo D 2010 Discovering Sets of Key Players in Social Networks *Computational Social Networks Analysis* **1** 24-47
[7] Borgatti S P, Carley K M, Krackhardt D 2006 On the robustness of centrality measures under conditions of imperfect data *Social Networks* **28(2)** 124-136
[8] Shetty J, Adibi J 2005 *Proc. 3rd Int. Workshop on Link Discovery, LinkKDD 2005* (Chicago) 117835
[9] Latora V, Marchiori M 2004 How the science of complex networks can help developing strategies against terrorism *Chaos, Solitons and Fractals* **20(1)** 69-75
[10] Burke M, Marlow C, Lento T 2010 *Proc. of Conf. on Human Factors in Computing Systems* **3** 1909-1912
[11] Saaty T L 1987 How to handle dependence with the analytic hierarchy process *Mathematical Modelling* **9(3-5)** 369-376
[12] Tarakanov D, Tsapko I, Tsapko S. Buldygin R 2015 *Proc. of Int. Conf. on Mechanical Engineering, Automation and Control Systems 2015 (MEACS 2015)* 012105
[13] Skirnevskiy I, Korovin A 2016 *Key Engineering Materials* **685** 857-862
[14] Luneva E E, Banokin P I, Yefremov A A, Tiropanis T 2016 Method of evaluation of social network user sentiments based on fuzzy logic *Key Engineering Materials* **685** 847-851