Survey of Influential User Identification Techniques in Online Social Networks

Roshan Rabade, Nishchol Mishra, and Sanjeev Sharma

Abstract. Online social networks became a remarkable development with wonderful social as well as economic impact within the last decade. Currently the most famous online social network, Facebook, counts more than one billion monthly active users across the globe. Therefore, online social networks attract a great deal of attention among practitioners as well as research communities. Taken together with the huge value of information that online social networks hold, numerous online social networks have been consequently valued at billions of dollars. Hence, a combination of this technical and social phenomenon has evolved worldwide with increasing socioeconomic impact. Online social networks can play important role in viral marketing techniques, due to their power in increasing the functioning of web search, recommendations in various filtering systems, scattering a technology (product) very quickly in the market. In online social networks, among all nodes, it is interesting and important to identify a node which can affect the behaviour of their neighbours; we call such node as Influential node. The main objective of this paper is to provide an overview of various techniques for Influential User identification. The paper also includes some techniques that are based on structural properties of online social networks and those techniques based on content published by the users of social network.

Keywords: Online social networks, Influential user, Content, Active user, Viral marketing.

1 Introduction

Social Network [1] is a structure of individuals or organizations tied to any of interdependencies such as friendship, fellowship, Co authorship, Collaboration

Roshan Rabade · Nishchol Mishra · Sanjeev Sharma
School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya,
Bhopal, M.P., India
e-mail: roshan.it2010@gmail.com, {nishchol,sanjeev}@rgtu.net

S.M. Thampi et al. (eds.), *Recent Advances in Intelligent Informatics*,
Advances in Intelligent Systems and Computing 235,
DOI: 10.1007/978-3-319-01778-5_37, © Springer International Publishing Switzerland 2014

etc. Online social networking provides a platform to build the communities of special interests virtually. It gives a new dimension to business models, inspires viral marketing [2], provides trend analysis and sales prediction in market, assists counterterrorism efforts [3] and acts as a foundation for information sources. In a physical world, according to [4] 83% of people prefer consulting family, friends or an expert over traditional advertising before trying a new restaurant, 71% of people do the same before buying a prescription drug or visiting a place, and even 61% of people prefer consultation with friends before watching a movie. In short, before people buy or take decisions, they talk and listen to others opinion, experience and suggestion. The latter affect the former in their decision making, and are aptly termed as the influential [4].

The Internet is the most influential invention of the 20th century. Since its commercialization in the 1990s, is steadily penetrating almost all dimensions of modern human life. With the emergence of the World Wide Web (WWW) and the evolution of information technologies, however, online social networks reached a new level. Online social networks can be thought as a number of nodes (or persons) connected by a series of links (or relationships). A relationship or link is defined as a pattern of social interaction between two or more persons that involve in meaningful communication and awareness of the probable behaviour of the other person.

Among all nodes in a given social network, it is important and interesting to discover nodes, which can affect the behaviour of their neighbours and, in turn, all other nodes in a stronger way than the remaining nodes, we call such nodes Influential nodes [4]. It is a fundamental issue to find a small subset of influential nodes in the network such that they can attract the largest number of members in a social network.

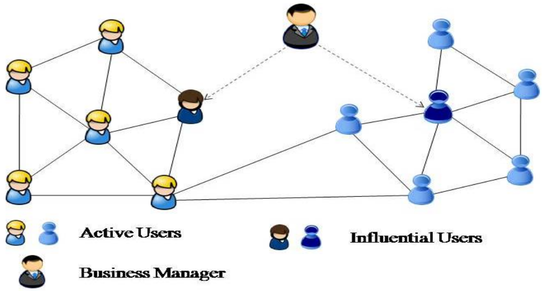

Fig. 1. Structure of Online Social Network

Figure 1 visualizes two different social groups. There is an influential node (selected by any technique) in both groups. If business manager want to advertise product via social network, then instead of interacting with each nodes, he focuses on these influential nodes only. In this way ad spread in to whole network quickly. It reduces the efforts of one to one interaction. The rest of this paper is organized as follows. Section 2 provides the basics of social network. Section 3 elaborates the related work in the area of influential user identification in the online social networks, Section 4 describes the applications of influential user identification in social network and Section 5 includes the conclusion and future work.

2 Definitions of Online Social Network

In this section of paper, definition of social network and its component discussed. The general idea of construction of society is the basis for the social network definitions. A society is not just a simple collection of individuals; it is rather the sum of the relationships that connect these individuals to one another. So the social network can be defined as the finite set of nodes (actors) and connecting edges (relationships). Every researcher defines social network in different forms, stated as follows:

In 1994, Wasserman and Faust [5] proposed a very sociological approach which defines: Actor**-** An actor is a discrete individual, corporate or collective social unit. Relation**-** A set of ties of a specific type; a tie is a linkage between a pair of actor. And Social network- The finite set or sets of actors and one or more relations defined on them.

In 2006, Yang, Dia, Cheng, and Lin [6] proposed a very formal way which defines as: Actor- A node in a graph; each node represents a customer. Relation- The undirected, unweighted edges in the graph; each edge represents the connectedness between two nodes. And Social network- An unweighted and undirected graph.

3 Related Work

Social and economic networks have been studied for decades in order to mine useful information not only to organize these networks in a way that maximum efficiency is accomplished but also to understand the role of nodes or groups in a network. By the emergence of social networks in recent years, finding in Influential nodes has absorbed a considerable amount of attention from researchers in this area. In this section, we are going to discuss the earlier methods to identify influential users in a social network, the methods to measure influential power of a user:

3.1 Techniques Based on Structural Measures of Online Social Networks

Social network research community describes a variety of structural measures for the identification of existing Influence in a network. This paper will briefly summarize the well-known centrality measures and number of link topological ranking measures.

Centrality Measures. Structural location of the node is advantageous to find the relative significance in the graph. There are various types of centrality measures of a node used to find the importance in the social network structure.

Degree Centrality. The degree centrality of any node in a graph or network is defined by the count of edges that are incident with it, or the count of nodes adjacent to it. Degree centrality of node in case of directional networks is given in two ways as in-degree and out-degree. The out-degree centrality is defined as [7]:

$$C_{DO}\ (i)\ =\sum_{j=1}^{n} a_{ij} \tag{1}$$

Where a_{ij} is 1 in the binary adjacency matrix A if a link from node *i* to *j* exists, or it is 0. Similarly, the in-degree centrality is defined as [7]:

$$C_{DI}\ (i)\ =\sum_{j=1}^{n} a_{ji} \tag{2}$$

Where *i* describe the node *i* and a_{ji} is 1 if a link from node *j* to *i* exists, or it is 0.

Closeness Centrality. Closeness centrality is a measure that identifies how fast it will take to flow information from node to all other nodes sequentially**.** Beauchamp [8] explains, nodes occupying a central position with respect to closeness are very productive in distributing information to the other nodes. Hakimi [9] and Sabidussi [10] developed a measure of closeness centrality as:

$$C_C\ (i)\ = \frac{1}{\sum_{j=1}^{n} d(i,j)} \tag{3}$$

Where $d(i,j)$ represent the distance between node *i* and *j,* which is the measured minimum length of any path connecting *i* and *j.*

Betweenness Centrality. Betweenness centrality counts the number of times a node founds (acts) as a bridge along the shortest route between two other nodes. It was introduced by Linton Freeman [11] as:

$$C_B\ (i)\ = \frac{\sum_{i\neq j\neq l} g_{jl}(i)}{g_{jl}} \tag{4}$$

Where $g_{jl}(i)$ is the number of shortest route linking the two nodes *j* and *l* containing node *i.*

Eigenvector Centrality. Eigenvector centrality measures the influence of a node within a network. It is based on the concept that links to high-scoring nodes attract more in comparison to the low-scoring nodes. A variant of Eigenvector centrality measure is Google's PageRank [12].

Let A again be the binary adjacency matrix of the network and $\vec{x}$ be the principal eigenvector corresponding to the maximum eigenvalue θ. The eigenvector centrality for a node i can be defined as a single element of the eigenvector, calculated as [7]:

$$C_E\ (i)\ =\ x_i\ =\frac{1}{\theta}\sum_{j=1}^{n} a_{ji}\, x_j \tag{5}$$

Edgevector Degree Centrality. Weighted digraph describes how often a user (node) called another one, or how many text messages sent by him/her. In this case, the element of an adjacency matrix a_{ij} describe the numeric weights of a connection from node i to j. Each weighted graph can be converted into a multigraph, where the same pair of nodes can be connected by multiple edges [13]. In paper [7], edge-weighted degree centrality is defined as:

$$C_{ED}\ (i)\ =\ \sum_{j=1}^{n}(a_{ij}\ +\ a_{ji}) \tag{6}$$

Link Topological Ranking Measures. Centrality measures (except eigenvector centrality) did not consider the type of node in the network. There exist some very influential nodes to which connection has more value than to others. With regard to social networks, a connection with a high centrality node might be more valuable than with only one neighbour node [7]. Web search engines leverage link topological ranking by means of HITS (Hyperlink-Induced Topic Search) and PageRank algorithms.

*HITS***.** Kleinberg [14] proposed a Web search algorithm called HITS which identifies authoritative pages and a set of hub pages. An iterative algorithm is used to find the equilibrium values for the authority and hub weights of a web page or node in a network, respectively. Equilibrium is reached if the difference of the weights between two iterations is less than a threshold value. For each page i a nonnegative authority weight $C_A(i)$ and a nonnegative hub weight $C_H(i)$ are exist. The weights of each type are normalized so their squares sum to 1 and are defined as:

$$C_A\ (i)\ =\ \sum_{j=1}^{n} a_{ji}\ C_H\ (j) \tag{7}$$

$$C_H\ (i)\ =\ \sum_{j=1}^{n} a_{ij}\ C_A\ (j) \tag{8}$$

Where a_{ji} is 1 if an edge from node j to i exists otherwise 0.

*PageRank Algorithm***.** The PageRank algorithm, which was initially developed by Brin and Page [12], the founders of the Google search engine; they maintain only a single metric for each web page. The so called PageRank is transmitted from the source page to the link target, and the value depends on the PageRank of the source page. The PageRank of the page or node i is the sum of contributions from its incoming links or edges. A constant damping factor f is the probability at each page that the "random surfer" will get bored and requests another random page. Additionally *(1–f)* is added to each node. This is done because if a node has an out-degree of zero then his PageRank would be zero. This zero-value would be passed down to the real node. To avoid this, a constant is added to the PageRank. The PageRank can be defined as:

$$C_{PR}\ (i) = (1 - f) + f \sum_{z \in Mi} \frac{C_{PR}\ (j)}{C_{Do}\ (j)} \tag{9}$$

Where Mi is the set of source pages that link to i and $C_{Do}\ (j)$ is the out-degree of page j, as described in the previous subsection. The damping factor f is often set to a value of 0.85 [12].

3.2 Techniques Based on the Diffusion Model of the Online Social Network

The diffusion model of the social networks also helps in the identifying the influential user in networks. The diffusion models [15] are classified into three categories: linear threshold model, independent cascade model and a model that combines both the features of the linear threshold model and independent cascade model.

Linear Threshold Model. The model proposed by Granovetter and Schelling [16] is based on the use of thresholds. Linear Threshold Model (LTM) is a variant of the original model. In this model a node u is influenced by its neighbour v with weight$b(u, v)$, subject to the constraint $\sum_v b(u,v) < 1$.

Each node has a predefined threshold value $\theta u \in [0,1]$. This show how difficult is to influence the node u when their neighbours are active nodes. Normally the threshold value is chosen randomly with uniform distribution, but in some cases entire network has a uniform threshold $(\theta u = 1/2)$. the process starts with a set of active nodes. The diffusion process continues in discrete steps, where the nodes that become active in step t, remain active until the end. At each step, every node u whose neighbour's total weight is at least θu is activated $(\sum_v b(u,v) \geq \theta u)$.

Independent Cascade Model. The Independent Cascade Model (ICM), proposed by Goldenberg, Libai and Muller [17] starts with a set of active nodes Ao, then the process triggers a cascade of activations in discrete steps. When a node u becomes active in step t, then it has only one chance to activate each inactive neighbour v with a success probability $P(u, v)$. the order of attempts to activate inactive neighbours is arbitrary. Independently of the success or not of the activation, u cannot make further attempts to activate v until the end of the process. The process runs until no more activation is possible. This process is called independent because the activation of any node v is independent of the history of activations.

3.3 A Technique Based on Community Mining in Online Social Networks

Kempel [18] state that optimization problem of influence maximization is NP-Hard. Accordingly to solve this type of problem, Greedy algorithms with provable approximation may give better outcome. But Greedy methods are expensive in computation, so as a result it is not practicable to social network. Yu Wang, Gao Cong, Guojie Song, Kunqing Xie [19] proposed novel method called "Community based Greedy algorithm for mining top-K influential nodes" which divides the network into a number of communities, and then selects individual community to find top-K influential nodes. The community structure is a main property of social network features: Individuals within a community have frequent contact; in contrast, individuals across communities has much less contact with each other and thus is less likely to influence each other. This property suggests that it might be a good approximation to identify influential nodes within communities instead of the whole network. This work gives several directions to expand research in construction of location based social network to find influential over time.

3.4 Techniques Based on Content Mining in Online Social Networks

In early studies, influential users identified based on the structural properties of a social network. In some social networking platforms (for e.g. blog network) structural properties may not exhibit the actual influence between users. In online social networks, user's content can bring on many activities that address its dynamic nature. This section includes content based influential user identification techniques.

Topic Level Influence. Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang proposed a method [20] that brings out the idea of mining the strength of direct and indirect influence. This proposed a generative graphical model that uses heterogeneous link information as well as the textual content related to each node in the online social networks to identify topic-level direct influence. Based on this learned direct influence, proposed topic-level influence propagation and aggregation algorithms apply to derive the indirect influence between various nodes. This paper focuses on mining topic level influence and propagation method to propagate whole network. In future by basis of this method behaviour prediction model can be employed in a social network.

ExpertRank Algorithm. G. Alan Wang, Jian Jiao, Alan S. Abrahams, Weiguo Fan, Zhongju Zhang presents a novel algorithm “ExpertRank” [21], to identify a topic-aware expert for online knowledge communities. The ExpertRank algorithm evaluates expertise based on both documents-based relevance and one's authority in his or her knowledge community. For this they modify the PageRank [12] algorithm to evaluate one's authority, so that it reduces the effect of certain biasing communication behaviour in online communities. ExpertRank algorithm explores three different expert ranking strategies for combining document-based relevance and authority: Linear Combination, Cascade Ranking and Multiplicative Scaling. This method is proficient to develop expert databases or organizational memory systems for providing key knowledge exchange among employees.

Content Power Users on Blog Network. Seung-Hwan Lim, Sang-Wook Kim, Sunju Park, and Joon Ho Lee [22] proposed a new method to identify content power user in Blog network. There are special users who bring on other users to actively utilize blogs, these users called influential users. Identifying such influential users in the blogosphere is important when establishing new business policy for the blog network. This paper defines the user, whose content exhibit significant influence over other users as content power users (CPUs) and proposed a method of identifying them. They analyse the performance of the proposed method by applying to an actual blog network and compare results with pre-existing methods for identification of power users. The experimental results of their analysis, demonstrate that the definition of content power user is adequate to address the dynamic nature of the blogosphere and the main concerns of the blog industry.

3.5 A Technique for Identifying Influential Users in Micro-blog Marketing

Micro-blog marketing has become an important business model for online social networks now a day. Micro-blog marketing make possible to publish advertisements directly to customers for attracting to buy products. Fei Hao, Min Chen, Chunsheng Zhu, Mohsen Guizani proposed a method [23] to discover the influential user in micro-blogging site (for e.g. Twitter, Sina Weibo). They try to analyse the influence of nodes in a micro-blog network and proposed the “Community Scale-Sensitive Max Degree (CSSM)”, an algorithm for maximizing the influence when placing advertisements. This work contributes mainly in the First, the influence analysis of the various nodes in micro-blog social networks. Second, a Community Scale-Sensitive Max Degree (CSSM) based influence maximization algorithm and third, an evaluation of the CSSM algorithm, using very popular micro-blog service Twitter dataset.

They observed that the influence of a node depends on following three matrices: first, centrality based on node degree. Second, sum of neighbour’s degree and third, attributes of nodes. The attributes of a node in the micro-blog network, includes activity degree, interaction degree and social prestige.

3.6 A Technique Based on Link Polarity in Online Social Networks (Blogs)

The main objective of above discussed techniques to understand the spread patterns of influence in a social network. All these approaches used to identify the influence between individual users, but do not take into account the question of what kind of influence inclined among them. So Keke Cai, Shengha Bao, Zi Yang, Jie Tang, Rui Ma, Li Zhang, Zhong Su introduced novel approach, an "Opinion Oriented Link Analysis Model (OOLAM)", [24] to characterize user's influence personae. In particular, three kinds of influence personae which take place widely in social network includes: Positive Persona, Negative Persona and Controversy Persona. Within the OOLAM model two factors, i.e., opinion consistency and opinion creditability are defined to capture the persona information from public opinion perspective. Extensive experimental studies have been performed to make obvious the effectiveness of the proposed approach to influence personae analysis using real web datasets.

4 Applications

The significant role of influence is studied extensively in sociology, communication, marketing and political science and in understanding peer pressure, trend analysis, obedience and leadership.

4.1 Opinion Leader Finding

Finding dominating people in societies has been a key question for marketing policy makers, political managers, social researchers, security analysts, engineers and computer scientists. Since any society can be considered as a network, network analysis has provided significant insight in this area. To find out the Leader, community based influential user identification approach applied.

4.2 Viral Marketing

Influence maximization has the obvious application in viral marketing [15] through social networks; in this company try to promote their products and services through the word-of-mouth propagations among friends in the social networks. The ultimate objective of the marketers to create viral messages that strongly convey to the influential users to spread in a short period of time.

Table 1 Key features of various Influential identification techniques

Name of Technique	Year launched	Key Features
Centrality Measures	1966	Identify the relative importance of Nodes in Network.
HITS Algorithms	1998	Identify Authority Pages and Hub Pages in Network.
PageRank Algorithms	1998	Maintain a single metric for information of all Web Pages.
Linear Threshold Model	1978	Focuses on Threshold (whole) behaviour of Nodes.
Independent Cascade Model		Focuses on Individual's Interaction in Network.
Community Modelling	2010	Efficient over Greedy method and orthogonal to existing algorithms of Influential detection.
Topic Level Algorithm	2010	Consider the presence of Indirect Influence with Direct Influence in Online Social Network.
ExpertRank Algorithm	2013	Document based relevance and Authority of Individuals.
Content Power User	2011	Illustrate the dynamic nature of Online Social Networks.
CSSM Algorithm	2012	Includes Activity degree, Interaction degree and Social prestige of the user.
OOLAM Algorithm	2011	Opinion consistency and Opinion creditability are used to capture the persona of user.

5 Conclusion and Future Work

This paper, presents the brief knowledge about the previous work that has been done in the field of identifying the influential users in the online social networks. This paper includes the techniques based on the topological structure of the social network as well as content of users. The topological analysis of the social networking makes use of graph theory but, sometimes it becomes complex and typical when the size of the social networks increases, particularly in present scenario. A methodology based on usage of the diffusion history of the activities of the user is also useful for identifying the Influential. In future we are going to propose a novel approach which includes topological measures, content power and link polarity values in account to identifying influential users in online social network. Another major area for extending works by basis of temporal and location based methodologies.

References

1. Aggarwal, C.C.: Social Network Data Analytics, pp. 1–14. Springer Science Business Media, LLC (2011)
2. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 61–70. ACM Press, New York (2002)
3. Coffman, T., Marcus, S.: Dynamic Classification of groups through social network analysis and HMMs. In: Proceedings of IEEE Aerospace Conference (2004)
4. Keller, E., Berry, J.: One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influential. The Free Press (2003)
5. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)
6. Yang, W.S., Dia, J.B., Cheng, H., Lin, H.T.: Mining Social Networks for Targeted Advertising. In: Proceedings of the 39th Hawaii International Conference on Systems Science, vol. 6, p. 137a. IEEE Computer Society (2006)
7. Kiss, C., Bichler, M.: Decision Support Systems 46, 233–253 (2008)
8. Beauchamp, M.: An improved index of centrality. Behavioural Science 10, 161–163 (1965)
9. Hakimi, S.: Optimum locations of switching centers and the absolute centers and medians of a graph. Operations Research 12 (1965)
10. Sabidussi, G.: The centrality index of a graph. Psychometrika 31, 581–603 (1966)
11. Freemann, L.C.: A set of measures of centrality based on betweenness. Sociometry 40, 35–41 (1977)
12. Brin, S., Page, L.: The anatomy of a large scale hyper textual web search engine. In: WWW Conference, Australia (1998)
13. Newman, M.E.J.: Analysis of weighted networks. Physical Review E 70
14. Kleinberg, J.: Auth. sources in hyperlinked environment. In: CM-SIAM Symposium on Discrete Algorithms (1998)
15. Singh, S., Mishra, N., Sharma, S.: Survey of Various Techniques for Determining Influential Users in Social Networks. In: IEEE International Conference on ETCCN, India, pp. 398–403 (2013)
16. Granovetter, M.: Threshold models of collective behaviour. American Journal of Sociology 83(6), 1420–1443 (1978)
17. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 211–223 (August 2001)
18. Kempel, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: ACM SIGKDD, pp. 137–146 (2003)
19. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks. In: KDD 2010, Washington, July 25-28 (2010)
20. Liu, L., Tang, J., Han, J., Jiang, M., Yang, S.: Mining Topic-level Influence in Heterogeneous Networks. In: CIKM 2010, Toronto, Canada (October 2010)
21. Alan Wang, G., Jiao, J., Abrahams, A.S., Fan, W., Zhang, Z.: ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. Decision Support Systems 54, 1442–1451 (2013)

22. Lim, S.-H., Kim, S.-W., Park, S., Lee, J.H.: Determining Content Power Users in a Blog Network: An Approach and Its Applications. IEEE Transactions on Systems, Man, and Cybernetics – part-A: System and Human 41(5), 853–862 (2011)
23. Hao, F., Chen, M., Zhu, C., Guizani, M.: Discovering Influential Users in Micro-blog Marketing with Influence Maximization Mechanism. In: Globecom 2012 - Ad Hoc and Sensor Networking Symposium (2012)
24. Cai, K., Bao, S., Yang, Z., Tang, J., Ma, R., Zhang, L., Su, Z.: OOLAM: an Opinion Oriented Link Analysis Model for Influence Persona Discovery. In: WSDM 2011, Hong Kong, China, February 9-12 (2011)