International Journal of Engineering & Technology, 7 (4.38) (2018) 950-954

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Detecting Influencers in Social Media Using Social Network Analysis (SNA)

Siti Nurulain Mohd Rum, Razali Yaakob, Lilly Suriani Affendey

Faculty of Computer Science & Information Technology
Universiti Putra Malaysia, 43400 Serdang, Selangor
**Corresponding author E-mail :snurulain@upm.edu.my*

Abstract

Social media has now become a key part of life in modern society; it is a place where people share their ideas, view, emotions, and sentiments. The explosion in the popularity of social media has led to an immense increase in data over the past few years. Users engage with this platform to share their experiences, feelings, and opinions on a broad range of topics, such as politics, personalities, news, products or events. Social media has also become a phenomenal platform that provides a powerful way for businesses to enhance their prospects and reach customers. Extracting and interpreting information from user-generated content is a trending topic in the scientific community as well as in the business world, and has attracted the interest of many commercial organizations. With the wise use of social media, the marketing process for promoting products and brands can be accelerated to reach the target audience. The beauty and health industry is one of the industries that make use of this platform as their digital marketing solution to integrate communications. Recently, many leading companies and brands have used digital influencers as their strategy for marketing campaigns in management and development. Therefore, the analysis of information extracted from social media is of great importance offering valuable insights and where the importance of each actor or individual in social media can be identified. This can be achieved through the use of Social Network Analysis (SNA). This research work aims at probing the effectiveness of SNA in social media in detecting the influencers in the area of beauty and health.

Keywords: *Social Network Analysis (SNA), Social Media, Centrality Measurements, Digital Influencer*

1. Introduction

Today, the beauty and health industry in Malaysia is very different to earlier. It is also expected that this industry will grow steadily in future. The greater exposure of information about beauty and health from the Internet has resulted in a growth in demand for personal care products from Malaysian consumers. This is also due to their self-awareness about health. They have become more sophisticated about their purchases. For instance, more consumers are now becoming experts in purchasing consumer health products for self-medication if they are infected with common infections or minor illnesses. In addition, it is possible that the growth of social media has led to consumers becoming more influenced by the reviews of beauty and health products by the users of Facebook, Instagram, and Twitter, as well as a variety of bloggers. Today social media is one of the platforms used by many entrepreneurs as word-of-mouth marketing to promote their products and services. Social media mining is a hot topic across disciplines from sociology to computer science. In general, the analysis of social networks is a branch of sociology that can be seen as a set of entities connected in a network through mathematics. Social Network Analysis (SNA) has been used by many researchers to measure the relationship and flows between groups, organizations, people and other connected knowledge entities. This is generally denoted by a collection of edges and vertices. The edges in the network refer to the relations between vertices and vertex is the term used to represent an actor in the network. In mathematics, this is called a graph (Scott, 2017). Precisely, a graph usually represented as G(V,E), where V denotes a set of vertices and E stands for a set of connected vertices through edges in V. The common notation, m, represents the number of edges and the n notation denotes the number of vertices. A subgraph of G(V,E) can be represented as a second graph S(V',E') provided it satisfies the rules of $V' \subseteq V$ and E' (all edges exist in) $\subseteq$ E. There are two types of graph; namely, unweighted and undirected graphs. There are always two-way relations in an undirected graph, and all edges in the network are equally strong in an unweighted graph. The directed and weighted graph is an example of a general type of graph. The relations exhibited in a directed graph can possibly be derived from an existing network in the form of a two-way relation or in single-way relation. In the weighted graph, it is possible for the relations to have varying importance. This is denoted as w to represent the weight; all weights are equal to 1 in an unweighted graph (Scott, 2017). A graph is usually represented by the use of adjacency matrices and adjacency lists (i.e., computer network). The n-by-n of an adjacency matrix has A(i,j) = w with the condition that there is an association between i and j of weight w with A(i,j) = 0 otherwise. The array of each vertex in the adjacency list represents the number corresponding to each neighbouring vertex as well as the relations' weight. Recently, graphs have placed a lot of attention on the scientific field (Mehra, 2005) and have become increasingly common with the advancement of the Internet that allows people to connect around the world. Its applications range from marketing, economics to biology (Wildemuth, 2016) and are often used to describe groups of people. These services are be-

coming the trial-and-error for marketers to increase the word-of-mouth pervasion of information with specific influencers (Landherr, Friedl, & Heidemann, 2010). In the literature, a specific agreement about what is an influencer is not properly described (Bakshy, Hofman, Mason, & Watts, 2011). Therefore, exploring the social network influencer in a specific area can possibly become a conceptual issue that requires plenty of development criteria measurement. The word influencer refers to an individual who has the ability to influence others and has an impact that is pervasive in society (Weimann, 1991). It also refers to an individual who shows certain characteristics, such as expertise, trustworthiness or network attributes (Keller & Berry, 2003). The first category of influencer is known in the literature as the opinion leaders, key-players (Borgatti, 2006), prestigious (Gayo-Avello, 2013), spreaders (Kiss & Bichler, 2008), and innovators (Cha, Haddadi, Benevenuto, & Gummadi, 2010). The second category, as defined in the literature, are celebrities (Fraser & Brown, 2002) or experts (Keller & Berry, 2003), such as a professor at Oxford University. Quantifying and measuring the influencer is pertinent for businesses as well as for economic marketing strategies with fast delivery. These people have the potential to accelerate the process of building mutually beneficial relationships with a large scale audience by utilizing technology, such as the World Wide Web (Kozinets, De Valck, Wojnicki, & Wilner, 2010). A measure of central tendency (centrality measures (CM)) has been identified as an appropriate method for identifying influencers in social media. Although plenty of centrality measures have been developed, four basic concepts have been recognized – degree, closeness, betweenness, and eigenvector. In this research work, we applied all these methods to centrally measure the collected data, and describe their patterns in terms of accuracy, interpretability and robustness. Figure 1 is the graphical illustration of the closeness centrality to show how each node in the network is independent of the others. For example, three nodes are connected to node 2 (node 1, node 3, and node 4) as shown in the diagram. In order to reach node 5, there is a need to pass a message through node 4. To reach the whole network, there is only one intermediary that node 2 needs to depend upon. However, for node 1 to reach node 3 and node 4 there is a need to depend on node 2 for two times as an intermediary. Before node 1 can reach node 5, there is also a need to go through node 2 and node 4. Therefore, for node 1 to reach all the nodes in the graph it needs to depend on node 2 for three times and node 4 once as intermediaries. In this scenario, node 2 is a less independent node compared to node 1. Node 2 has greater centrality than node 1 (Freeman, 1978).

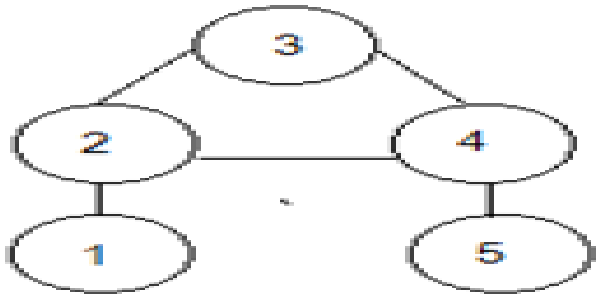

Figure 1: Closeness Centrality: Five vertices and five undirected edges (Freeman, 1978)

of a vertex that are counted, but all the other vertices are taken into account. The average distance of the vertex's source to any other vertex within the graph is measured by the metric and written formally as:

$$C_C(v) = \frac{\sum t \in V_{\setminus v} D_G(v,t)}{n-1} \quad (1)$$

where the shortest path between vertex v and t is denoted by DG(v,t), which is called the geodesic. In this situation, the duration time needed for the source to spread information will be reflected. The value of the reciprocal given in (1) is also sometimes referred to as the closeness centrality, which corresponds to the time taken for the source of the vertex to spread the information. The eigenvector centrality of the adjacency matrix refers to the largest eigenvalue. Thus, the formula can be written as:
where the shortest path between vertex v and t is denoted by DG(v,t), which is called the geodesic. In this situation, the duration time needed for the source to spread information will be reflected. The value of the reciprocal given in (1) is also sometimes referred to as the closeness centrality, which corresponds to the time taken for the source of the vertex to spread the information. The eigenvector centrality of the adjacency matrix refers to the largest eigenvalue. Thus, the formula can be written as:

$$C_E(v) = \frac{1}{\lambda} \sum_j R_{ij} (C_E(v))_j \quad (2)$$

The eigenvector centrality given in (2) can be applied to both graphs – the directed and undirected graphs. It can also be used by both weighted as well as unweighted graphs where neighbours of each vertex will receive information through the iterative process. The strength of the vertex's relation and the significance of the vertex itself are determined by the amount of information to be sent. The size of the adjacency matrix is n×n, where n typically has a very large value that makes the calculation of the eigenvectors computationally costly. Therefore, the iterative method is the solution through which the graph of the adjacency matrix is repeatedly multiplied until reaching equilibrium. The representation of the kth iteration is as follows:

$$b_{k+1} = \frac{Ab_k}{\|Ab_k\|} \quad (3)$$

The common initial value given to b0 is 1 for the equal starting value of each vertex. The complexity of this algorithm (time) is represented by O(n+m). To analyse the graphs, the betweenness centrality concept is commonly used to measure the number of paths (geodesic) that need to be reached between pair from all geodesics. Formally, the sum of the overall vertices' pair can be expressed as:

$$C_B(v) = \sum_{s \neq v \neq t \epsilon V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (4)$$

The shortest path from s to t is represented by σst(v) (4) executed through v, where t refers to the total number of nodes. The shortest paths in the graphs of all the vertices need to be counted, which will always cause computational heaviness. This can be extremely burdensome, especially for large graphs in terms of time consumption. Therefore, (Brandes, Borgatti, & Freeman, 2016) proposed the betweenness centrality algorithm, which incorporates the breadth first search (BFS) to calculate the sparse networks in order to find all the shortest paths that, later, can be used to calculate the closeness centrality of the source of the vertex. The order O(nm) is the algorithm's computational cost, and the way it is calculated is similar to the regular betweenness centrality except that consideration is given to all the possible paths of vertices (between any pair), not just the geodesics.

The Aim of Research

This paper conducted research to develop a new method for data

collection through the friendship graphs from social media. The collected data were analysed using the centrality measurement to determine influencers within the health and beauty industry.

2. Method of Research

Twitter is a popular social media platform that offers an application program interface (API) that allows public users to access and extract huge amounts of data on the fly. This platform also provides an API to access the real-time data of users' tweets as well as users' personal information in the form of a sample. Facebook is another social media platform that provides features that have a similarity to Twitter. However, extracting the information of personal users on Facebook is more difficult due to the protection set up by the platform itself. In addition, the API introduced by Facebook only enables the authorized Facebook application to fetch information. This makes it impossible to map relevant networks due to the poor data extraction process. Moreover, there are huge amounts of documentation available for reference with practical examples given on the Twitter's API. The library provided on Twitter is standardized so that it can be used across languages in a simple way to connect the OAuth through the API. In Twitter, a follower refers to a user that can automatically see all the messages sent or posted by the user account(s) that is being followed. However, in this research work, we define followers as the in-degree relationship or in simple words, as a number of connections (users) pointing to a node (a member). Whereby friends refer to the relationships; namely, out-degree (from a node to other links). All tweets, including a friend or personal information, can be seen by followers. In this study, we made a condition for data extraction from the beginning; only reciprocal communication will be considered in the network structure (talk about the same topic) rather than one-way communication. In order to identify influential users on Twitter, two steps of data extraction are carried out. The first step involves the identification of prominent founders of beauty and health products in Malaysia. The account selection is based on the number of followers on Twitter; more than 1 million followers. The second step involves the process of extraction of the 20 most common users using the map network. The first extraction of 20 lists accounts from the first iteration is considered as our initial experimental data, and this process is iterated a few times until the number of users reaches a list of 100 Twitter accounts. The assumption of the data extraction process is based on a specific topic on users' Twitter that talks about the same topic from the large influencer Twitter from the beauty and health domain. The whole process of data extraction is summarized as follows: -

- First iteration: Selection of top 20 active friends from the Twitter accounts of top five influencers and entrepreneurs in beauty and health industry in Malaysia
- Second iteration: Selection of top 20 active friends from the first iteration list
- Third Iteration: Selection of top 20 active friends from the second iteration list
- Fourth Iteration: Selection of top 20 active friends from the third iteration list
- Fifth Iteration: Selection of top 20 active friends from the fourth iteration list
- Combining all the list of account IDs obtained from the first iteration until the fifth iteration to build an adjacent matrix of 100 x 100

The whole process of data extraction is presented in pseudocode, as shown in Figure 2, Figure 3, and Figure 4. In the pseudocode algorithm (Figure 2), the UA variable stored the list of the top five influencers in the beauty and health industry in Malaysia. The selection was based on the number of their followers, all of which had more than one million followers. These top five selected influencers have produced a number of products from hair care to skincare solutions. Their product brands have been established in the Malaysian market for over 10 years and have gained popularity among Malaysian consumers to fulfil their personal needs as well as to make them look good and feel good. The R programming language is the preferred programming language to realize this research work for several reasons; firstly, the language is open source; secondly, this language supports the data structure, such as list, vector, arrays, and matrices. I also supports the object-oriented programming and is very suitable for machine learning, statistics, and data analysis work; and, thirdly, it is easy to learn the language and provides good documentation for reference.

```
For Each FL
  CNT_FL = Get total number for Each FL
End For
SORT = Get Top 20 from CNTL_FL
L2 = SORT

SORT = NULL;
FL= NULL;
Inputting L2
L2= Lookup users L2
For Each L2
        FL= Get FriendID
        FL = FL Not in RM
End For
For Each FL
  CNT_FL = Get total number for Each FL
End For
SORT = Get Top 20 from CNTL_FL
L3 = SORT

SORT = NULL;
FL= NULL;
Inputting L3
L3= Lookup users L3
For Each L3
        FL= Get FriendID
        FL = FL Not in RM
End For
For Each FL
  CNT_FL = Get total number for Each FL
End For
```

Figure 3. Algorithm for data extraction and adjacency matrix building

```
SORT = Get Top 20 from CNTL_FL
L4 = SORT

SORT = NULL;
FL= NULL;
Inputting L4
L4= Lookup users L4
For Each L5
        FL= Get FriendID
        FL = FL Not in RM
End For
For Each FL
  CNT_FL = Get total number for Each FL
End For
SORT = Get Top 20 from CNTL_FL
L5 = SORT

ADJ = Build Matrix Adjacent Combining (L1, L2, L3, L4, L5)
GRP = Build graph for ADJ
BTWN= Build Betweenness Matrix from GRP
CLN = Build Closeness Matrix from GRP
IDGR= Build In-Degree Matrix from GRP
ODGR= Build Out-Degree Matrix from GRP
EIGN= Build Eigenvector Matrixfrom GRP
End
```

Figure 4. Algorithm for data extraction and adjacency matrix building

3. Analysis and Discussion

The graph's network structure obtained is depicted in Figure 5. The Igraph package in R is a network visualization and analysis package to analyse the graphs. This package also provides the utility to compute the vertex centrality directly. The graph is plotted using our adjacency matrix through the use of the built-in package and library of igraph. Visually, the generated graph's structure shows that the nodes are strongly connected in the centre

of the graph. Literally, the graph illustrates that the distances between the nodes in the graph are reduced with the degree of connectivity between each node in the network. The highest the in-degree score owned by the group of nodes, the more central it is located in the network; and the extremity node is the one with the lowest in-degree score. The graph structure also indicates that there is no presence of any sub-group as the diagram seems to be cyclic.

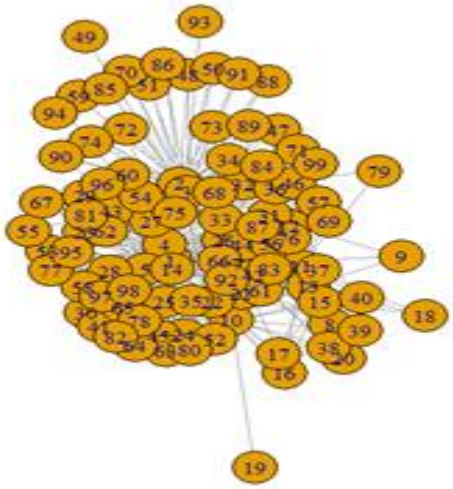

Figure 5: Graph translation of our adjacent matrix

Table 1. Statistical analysis of all centrality measurements

	Mean	Standard Deviation	Minimum	Maximum
In-degree	0.872	0.121	0	0.656
Out-degree	0.872	0.121	0	0.656
Closeness	0.329	0.04	0.425	0.56
Betweeness	0.01	0.035	0	0.245
Eigenvector	0.249	0.179	0.26	1

Table 2: Correlation study of centrality scores

	In-degree	Out-degree	Closeness	Betweennes s	Eigenvector
In-degree	1				
Out-degree	1	1			
Closeness	0.534	0.534	1		
Betweeness	0.930	0.930	0.473	1	
Eigenvector	0.951	0.951	0.637	0.849	1

Table 1 presents the statistical analysis of all the centrality measurements generated based on 100 Twitter accounts (extracted from five iterations). The breadth-first search (BFS) is used for the degree centrality matrix to find the shortest path measurement. This algorithm is the most effective way to calculate the distances between the nodes from a large network dataset (Easley & Kleinberg, 2010). It always starts from the arbitrary node (source) of the graph to explore the nearest nodes (neighbours), before traversing to the next level of the network structure. The vector score obtained for the in-degree and out-degree measurement are similar and identical; therefore, the mean, standard deviation, the minimum and the maximum are also identical. This result indicates that the level of influence has the same level of popularity and vice versa. The mean for both the in-degree and out-degree is 0.872, the standard deviation is 0.121, the minimum is 0, and the maximum is 0.656. The average of the closeness centrality score is 0.329 with a standard deviation equal to 0.04. The minimum value for the closeness centrality is 0.56 whereby the maximum value is 0.425, and there is no extreme value present. It can be concluded that the majority of scores of centrality is close to the mean. In general, the utilization of the closeness centrality measure in this research work is conceptually relevant in order to determine the minimum and maximum time spent by the central nodes to communicate with other nodes according to the point of view of (Scott, 2017). In addition, the higher the score of the closeness centrality the faster the messages spread to the other nodes in the network as there is a huge number of links pointing to the node. The mean of the eigenvector centrality is equal to 0.249, and the standard deviation is 0.179. The maximum for the eigenvector centrality value is 1, and the minimum value is 0.26.

Table 2 presents the correlation study between all the rankings and the results are consistent with the studies done by (Valente, Coronges, Lakon, & Costenbader, 2008). There is a strong correlation between the in-degree and out-degree; this result revealed that those with a huge number of friends are the ones that have the highest number of followers. This might be due to the nature of the BFS algorithm used by both measurements. The number of paths traversed along the network is measured by the betweenness centrality. According to (Cha et al., 2010) the popularity of the audience size is usually measured by the in-degree; however, it does not necessarily represent the actual degree of influence. For instance, a user with the highest number of followers may increase his popularity, but that does not mean that they are a great influencer. However, the result obtained in this research work presents that there is a strong relationship between the in-degree, out-degree centrality measurement and the eigenvector measurement that provides almost identical ranking with 0.951. This might be due to the computation of the eigenvector centrality that gives the most influential node in a network from the inward as well as from the outward neighbours. The more important the node in a network the higher the eigenvector score obtained by the node in the network, and, hence, should be the one considered as an influential node. The basic factor in determining the most influential nodes in a network with centrality measurement using eigenvector centrality is the eigenvector value among its adjacency (Bloch, Jackson, & Tebaldi, 2016; Brandes, 2001). The result also shows that there is a strong relationship between the eigenvector centrality and the betweenness centrality scores with r=0.849, as well as the eigenvector centrality with the closeness centrality that provide the ranking of 0.637. This result reveals that the more important a node in the network is, the closer the geodesic path between the node with other pairs of nodes in the network. In other words, an important node (a node is important if it is connected to other important nodes) makes more people depend on them to make a connection with others and vice versa. The result also indicates that the more important the node, the more central and closer that node is to the others.

4. Conclusion

Today, social media has become a key part of our daily life and is a contributor to the radical changes in people's communication behaviour worldwide. This platform has become the marketing land for many companies to market their products and services. Finding the right ambassador with a well-integrated network in social media is the major issue for many companies. Recently, many measurements of centrality have been developed and analysed to address these issues. This research work applies the five centrality measurements; namely, in-degree, out-degree, closeness, betweenness, and eigenvector on the beauty and health topic area to identify the most important vertices within a network. In conclusion, for someone who is in-charge of the communication in the beauty and health industry, the closeness centrality is the best measurement used to determine the minimum time for a user to spread a message and information to others. In order to find well-integrated users (significant node) in a network that is highly followed by many other influencers; the eigenvector is the most appropriate measurement. The result of this research work also indicates that the more important the users in the network, the more central and closer the node is to the other users. The result also reveals that the most important actors in a network structure are

also those that highly depended on the other actors in order to make a relationship. Depending on the cases for the application of the centrality measurement, each must be interpreted according to its context. It is recommended that future research should be carried out to compare it with our result.

Acknowledgment

This work is fully supported by Universiti Putra Malaysia research grant (9595800).

References

[1] Bakshy, Eytan, Hofman, Jake M, Mason, Winter A, & Watts, Duncan J. 2011. Everyone's an influencer: quantifying influence on twitter. Proceedings of the fourth ACM international conference on Web search and data mining.
[2] Bloch, Francis, Jackson, Matthew O, & Tebaldi, Pietro. 2016. Centrality measures in networks.
[3] Borgatti, Stephen P. 2006. Identifying sets of key players in a social network. Computational & Mathematical Organization Theory, 12(1), 21-34.
[4] Brandes, Ulrik. 2001. A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[5] Brandes, Ulrik, Borgatti, Stephen P, & Freeman, Linton C. 2016. Maintaining the duality of closeness and betweenness centrality. Social Networks, 44, 153-159.
[6] Cha, Meeyoung, Haddadi, Hamed, Benevenuto, Fabricio, & Gummadi, P Krishna. 2010. Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17), 30.
[7] Easley, David, & Kleinberg, Jon. 2010. Networks, crowds, and markets. Cambridge Univ Press, 6(1), 1-6.
[8] Fraser, Benson P, & Brown, William J. 2002. Media, celebrities, and social influence: Identification with Elvis Presley. Mass Communication & Society, 5(2), 183-206.
[9] Freeman, Linton C. 1978. Centrality in social networks conceptual clarification. Social networks, 1(3), 215-239.
[10] Gayo-Avello, Daniel. 2013. Nepotistic relationships in twitter and their impact on rank prestige algorithms. Information Processing & Management, 49(6), 1250-1280.
[11] Keller, Edward, & Berry, Jonathan. 2003. The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy: Simon and Schuster.
[12] Kiss, Christine, & Bichler, Martin. 2008. Identification of influencers—measuring influence in customer networks. Decision Support Systems, 46(1), 233-253.
[13] Kozinets, Robert V, De Valck, Kristine, Wojnicki, Andrea C, & Wilner, Sarah JS. 2010. Networked narratives: Understanding word-of-mouth marketing in online communities. Journal of marketing, 74(2), 71-89.
[14] Landherr, Andrea, Friedl, Bettina, & Heidemann, Julia. 2010. A critical review of centrality measures in social networks. Business & Information Systems Engineering, 2(6), 371-385.
[15] Mehra, Ajay. 2005. The Development of Social Network Analysis: A Study in the Sociology of Science: SAGE Publications Sage CA: Los Angeles, CA.
[16] Scott, John. 2017. Social network analysis: Sage.
[17] Valente, Thomas W, Coronges, Kathryn, Lakon, Cynthia, & Costenbader, Elizabeth. 2008. How correlated are network centrality measures? Connections (Toronto, Ont.), 28(1), 16.
[18] Weimann, Gabriel. 1991. The influentials: back to the concept of opinion leaders? Public Opinion Quarterly, 55(2), 267-279.
[19] Wildemuth, Barbara M. 2016. Applications of social research methods to questions in information and library science: ABC-CLIO.