Journal of Theoretical and Applied Information Technology
30th November 2015. Vol.81. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IDENTIFYING INFLUENTIAL USER IN TWITTER : ANALYSIS OF TWEET CONTENT SIMILARITY IN WEIGHTED NETWORK

WARIH MAHARANI
Telkom University, Telkom School of Computing Bandung Indonesia
E-mail: wmaharani@telkomuniversity.ac.id

ABSTRACT

In recent years, Social Network Analysis (SNA) is still growing rapidly. The mapping and measurement of the interaction in SNA can be used in many areas, for example to find the most influential users to improve the marketing strategy in Small and Medium Enterprise (SME). In order to find the most influential users in a network, we can apply the centrality measurement such as degree centrality, betweeness centrality, closeness centrality and eigenvector centrality. In this manner, degree centrality is conceptually the simplest one, which is defined as the number of links incident upon a node. While recent works has focused on number of nodes with the weighting between nodes according to its interaction such as following, followed, mention, retweet and reply. In this study, we investigate the combination of tweet content similarity and the interactions between users in twitter using Opsahl method. In this paper, we compare the proposed method with the baseline system from previous research. The experimental result show that the tweet content similarity affect the result of the most influential user in comparison with existing method.

Keywords: *Social Network Analysis (SNA), Centrality Measurement, Degree Centrality, Betweeness Centrality, Closeness Centrality, Eigenvector Centrality, Opsahl Method*

1. INTRODUCTION

Nowadays, people can connect easily through social media. Social media is an online media, where users can interact and share information. Social networking can bring people to talk, share ideas and interests, or make new friends. Twitter is an online social networking service that enables users to interact each other through a short messages consist of maximum 140 character called "tweets". The graf on Twitter consists of nodes and edges. Node represents a Twitter user and the edge represents the relationship between users. The graf used on twitter is a directed-graph unweighted, which represents the direction of the flow of information. However, how can we explore or exploit the flow of information that occurs on the network?

Social Network Analysis (SNA) is a mathematical method that allows to map and measure the flow of interactions formed between nodes on a network. The purpose of SNA is to understand how groups of individuals relate and how consequently, as well as understand how they behave. Scott [1] defines the SNA as a set of methods used to investigate aspects of the relationships in the data structure. Thus, SNA can analyze interactions within a group by analyzing nodes with the relationship between them. Otte and Rousseau [2] shows that SNA can be used to capture implicit information, including the relationship between user and their interaction. There are various ways that are often used to define centrality in a network to identify the influence of each node : degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality [3], [7]. Degree centrality is conceptually the simplest one, which is defined as the number of links incident upon a node. Meanwhile betweeness centrality and closeness centrality are global metrics which more complex but it can better to identify the influential nodes.

The most influential nodes are the nodes that have the greatest weight value, both in terms of the number of interactions or the number of connections between users. Centrality measurement used to identify the most influential nodes in a network. Generally, an indicator of a network is represented in the number of connected nodes, known as degree centrality [8], [11]. The node degree has been developed further by the amount of weight, called nodes strength [4], [8], [12]. However, in previous research, the node strength

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

did not consider the content of information that flows between nodes [6], [9], [10]. The degree centrality value will be less representative if only considering the number of nodes that are connected, regardless of the content of information that occurs between nodes. While the tweet content and interactions that occur on twitter namely the relations following, followed, retweet, mention and reply also determine the effect of a node. Previous studies have focused on the number of nodes with the node weighting and focused on the interaction relationship of a node with other nodes, but did not consider the content of the information conveyed in the network, [5], [8–10].

Opsahl method is a method which can be used to combine the tweet content and its relationship [8]. Opsahl method is able to analyze the effect of the node degree, the node weight and the content relationships between nodes [8]. Therefore, in this paper, we combine the content similarity and its relationship, to find the most influential users on microblogging twitter by using Opsahl method. The most influential user can be used as an alternative in marketing strategy through twitter. This paper will focus on those combination, which will be compared with the baseline system. The baseline system implemented Opsahl method without considering the tweet content similarity. The purpose of this paper is to get a comprehensive analysis of the content similarity and its relationships between nodes. The difference of this paper with the previous paper [6], [8–10] is the combination of tweet content similarity in the calculation of node strength.

The paper is organized as follows : Section 2 presents the centrality measurement by using degree centrality ; Section 3 describes in detail the proposed method ; Section 4 discusses the result and analysis ; finally, Section 5 concludes the paper and challenges as well as opportunities for further research.

2. CENTRALITY MEASUREMENT

In general, social networks can be distinguished by the type of nodes and their relationships. There is a social network that has multiple types of nodes with similar relationships. In addition, there is also a social network that has uniform nodes with different weights and has a varying relationship, or any other combination among the type of node, the type of node weighting and the types of relationships. The type of weighting on a network can vary depending on the case and the field of social networking implementation. In this study, we used microblogging twitter, which have a uniform node and having similar types of relationships such as user relationships with its weighting in each relationship. The weight value of user relationship are determined based on the the interaction occurs between nodes such as following, followed, retweet, reply, and mention [6]. Those value will increase if there is a similarity of tweet content and their interactions, which will lead to the greater weight values.

Centrality is an important measurement which most frequently used in the SNA [8], [11]. According to Freeman [11] at the end of 1977, centrality measurement is a measurement that involves three aspects : degree, closeness and betweeness. Degree centrality is defined as the number of degrees of relationship that is connected to a node, for example in a user relationship in a community [7]. The central user is one with many connections. In this research, people who are considered the most influential user is one who has the most user relationship, so that the value of the measurement for the degree centrality is calculated based on the number of relationship connections. As mentioned in previous chapter, this paper will focus on degree centrality to find most influential user in microblogging Twitter.

For example, if there is a graph with n nodes, then the degree centrality is defined in the formula (1) below [8]:

$$C_D(V_i) = \sum_{k=1}^{n} a(V_i, V_k) \quad (1)$$

Where a(u,v) = 0 if u and v not connected by a relationship, otherwise a(u,v) = 1.

According to Opsahl, the measurement of the degree centrality can be formulated with [8] :

$$k_i = C_D(i) = \sum_j^N x_{ij} \quad (2)$$

Where i = focal node, j = representation to all nodes, N = total node, x = adjacency matrix x_{ij} = a cel connectedness (x_{ij} = 1 if interconnected and x_{ij} = 0 if it is not related).

For a weighted networks, Opsahl et al. [8] proposes a calculation degree centrality by combining the number of ties as a node degree and the total weight of ties as a node strength, because both of these measures indicate the node involvement in a network. Generally, degree

Journal of Theoretical and Applied Information Technology
30th November 2015. Vol.81. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

centrality can be expanded by considering the total weight to analyze the weight and the node strength of the network [8] :

$$s_i = C_D^w(i) = \sum_j^N w_{ij} \quad (3)$$

Where w = weight of the adjacency matrix, $w_{ij} \neq 0$ if nodes i and j are connected.

To combine the node degree and node strength requires a tuning parameter (α), which its function is to determine the number of relations compared to the weight of the edge. The centrality can be determined by adjusting the weighted average by tuning parameter with the results of the number of nodes that are connected with focal node [8].

$$C_D^{w\alpha}(i) = K_i \times \left(\frac{s_i}{k_i}\right)^{\alpha} = k_i^{(1-\alpha)} \times s_i^{\alpha} \quad (4)$$

3. PROPOSED METHOD

This research applied degree centrality by combining the node degree and node strength which consist of tweet content similarity between nodes and its weighting according to relation of following, followed, mention, reply and retweet. Due to the tweet content is one of factors that influence the interactions that occur on twitter, then this study focused on the incorporation of tweet content similarity as a node strength. The value of tweet content similarity between users is computed based on term similarity. As a baseline, the results will be compared with the method without considering tweet content similarity between nodes. In this paper, the value of node degree obtained from the level of connectedness of a node in a graph, based on the relationships that occur between nodes, such as following, followed, reply, retweet and mention relationship. The value of the node degree and node strength will become the main indicator in the network.

This research consists of a preprocessing phase, the measurement of degree centrality, weighting and ranking phase. The purpose of the preprocessing stage is to eliminate the ambiguous data due to duplication of data. The preprocessing phase will produce a node table and a relationships table. The relationship table will be represented in a matrix (n x n), where n is the number of users. This process is done by using NodeXL tools[1].

The second phase is the degree centrality measurements by using Opsahl method. The measurement is done by calculating the node relationships as well as the tweet content similarity between nodes. This phase begins by searching the degree of each node that is adjacent to the central node using the equation (1) and (2).

The idea of weighting method in this study were drawn from previous studies of bibliometric and co-authorship. In the study of co-authorship, the calculation of the weighting between the author calculated based on the number of paper documents and journals that have been prepared by both simultaneously. The greater number of documents, then the weight will increase according to the number of the document. As well as on studies of bibliometric, the weighting is calculated based on the amount of interaction between web link.

In the case of twitter dataset, each tweet can be analogized as a micro documents distributed to various other users. So that, for each additional tweets can be considered as an additional parameter weights, either retweet, reply or mention. For each of these interactions always involve only 1 piece tweet as the smallest unit, in a sense, for every mention, reply and retweet can not be done in 2 pieces tweet or more at once. So, if the weight measurement process based on the number of tweets that flow, then the interaction mentions, reply and retweet will only be worth 1 weights, as shown in Figure 1.

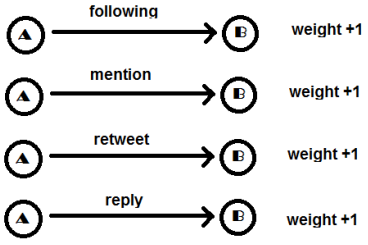

Figure 1:The interaction weight

3.1 Degree centrality

For example, there are 4 user (A, B, C, D) in a twitter network, with connectivity to one another in the form of a matrix (n x n) - weighted undirected graph that has involved a following, followed, retweet, mention and reply relationship. Preprocessing phase will produce a matrix (n x n) as follows :

Table 1: Initial matrix

User	A	B	C	D
A	0	2	1	3
B	2	0	2	3
C	1	2	0	3
D	3	3	3	0

Journal of Theoretical and Applied Information Technology
30th November 2015. Vol.81. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The above example shows that every user connected to another user with adjacency relation = 3 and each interaction of following, followed, mentions, reply and retweet has the same weight (weight = 1 unit). Based on the equation (4), the value of the node degree and node strength with the tuning parameter α of each node, as shown in the table 2.

Table 2: Degree Centrality

U	C_D	C_D^W	α			
			0	**0.5**	**1**	**1.5**
A	3	6	3	4.2426	6	8.4852
B	3	7	3	4.5825	7	10.692
C	3	6	3	4.2426	6	8.4852
D	3	9	3	5.1961	9	15.588

α is a tuning parameter with positive values contained in a research conducted by Opsahl et. al. [8]. If the tuning parameter ranges between 0 and 1, the node that has a high value for the degree and strength, will increase the value of degree centrality of the node. Whereas if the tuning parameter-value above 1, then the node that has a low value for the degree, but has a high value for the strength, will increase the value of degree centrality of the node.

Ranking phase derived from user rankings that have the greatest weight was measured using Opsahl method, which consider the node strength and the node degree, as well as combinations of tuning parameter value (α). Based on the table above, the ranking of the most influential user = D, B, A, C.

3.2 Content-based degree centrality

Content-based approach was calculated based on tweet content similarity. This study examined the various scenarios by simply applying tweet content-similarity and the incorporation of content and following, followed, retweet, mention and reply relationships. Here is an example of tweet content similarity based on term similarity between users A, B, C and D :

Table 3: Initial matrix based on tweet content similarity

U	A	B	C	D
A	1	0.6	0.2	0.15
B	0.6	1	0.3	0.2
C	0.2	0.3	1	0.4
D	0.15	0.2	0.4	1

The example above shows that each user connected to another user with adjacency relation = 3. The calculation of degree centrality will involve node degree and node strength regulated by tuning parameters, as shown in Table 4.

Table 4: Degree centrality based on tweet content similarity

U	C_D	C_D^W	α			
			0	**0.5**	**1**	**1.5**
A	3	0.95	3	1.6881	0.95	0.5345
B	3	1.1	3	1.8165	1.1	0.6660
C	3	0.9	3	1.6431	0.9	0.4929
D	3	0.75	3	1.5	0.75	0.375

The result of user ranking based on tweet content similarity is B, A, C and D, which are relatively different from the results in section 3.1. In order to analyze the influence of the tweet content that used to find the most influential user, this study will combine the tweet content and the following, followed, reply, retweet and mention relationships as a node strength in degree centrality calculation. Based on the those examples, it can be seen that the node relationship combined with the tweet content similarity as a node strength might form a perfect network, that represent its connection and its content between nodes.

4. RESULT AND ANALYSIS

Experiments conducted to analyze the effect of the relationships, weight and tweet content similarity, which is contained in the relation between nodes. Specifically, experiments conducted to examine the effect parameter α, which is used as a tuning parameter in the calculation of degree centrality. The system will generate the user's ranking of the most influential in the dissemination of information content on social networking twitter.

The xperiments were conducted by using a dataset that crawl on twitter, as used in [9], [10]. The dataset was obtained from one SME user consist of 298 nodes. The following are the results of experiments with some variations in the α value :

Table 5: The experimental result

Rank	Degree centrality			+ Content-based		
	α = 0	**0 < α < 1**	**α = 1.5**	**α = 0**	**0 < α < 1**	**α = 1.5**
1	4431.0	5983.5	6083	398	577.20	601.80
2	3870.8	5274.1	5971	201	408.01	600.10
3	3782.3	4603.1	5601	198	302.02	532.56
4	3646.5	4480.2	5482	176	280.61	501.23
5	3509.4	4406.2	5402	134	256.83	500.98
6	3300.7	4341.8	5348	121	210.77	476.54
7	3093.0	4040.8	5348	120	178.94	465.55
8	301.29	402.72	5022	99	154.32	435.98
9	274.83	378.00	4780	89	133.20	390.89
10	231.20	360.28	4608	87	122.87	375.65

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5 shows the difference between the degree centrality with the addition of tweet content similarity between nodes. The experiment with tuning parameter (α) = 0 will have the same value of the node degree centrality as performed by Freeman [11]. With applying the standard of tuning parameter (α=0), it will consider only one indicator of degree and ignore node strength indicator. With just considering the value of a degree, then the node that has the greatest value would be the most popular node in the dataset community. In other words, the value of node degree is proportional to the value of its centrality. Strength value has no effect at all when using the tuning parameter = 0, so this means that a weight relation has no effect on the results of the degree centrality calculation. Whereas if the tuning parameter ranges between 0 and 1, the value of degree centrality of a node can be affected by two indicators: the degree and strength. The greater the degree and strength of a node, it will have a positive impact which enlarge the degree centrality of a node. However, if the value of the tuning parameter = 1, the indicator considered by Opsahl method for determining the value of degree centrality is only influenced by the value of strength. It is similar to degree centrality measurements performed by Barrat et al. [12].

In the second experiment, this study combined the tweet content similarity and the relationships of following, followed, mention, retweet and reply by taking the average value of the tweet content and relationship interactions. The experimental result shown in table 6.

Table 6: Combination degree and content-based

Rank	α = 0	0 < α < 1	α = 1.5
1	241774.5	3280.35	3342.4
2	2035.92	2841.06	3285.5
3	1990.15	2452.56	3066.78
4	1911.25	2380.41	2991.62
5	1821.7	2331.51	2951.49
6	1710.85	2276.29	2912.27
7	1606.5	2109.87	2906.77
8	200.15	278.52	2728.99
9	181.91	255.6	2585.44
10	159.1	241.58	2491.83

Based on the table 6, it can be seen that the value of degree centrality generated from the above experiment are relatively different, which will lead to differences in the results of user ranking. The value definitely affect the ranking of the most influential users. This difference is due to the addition of tweet content similarity in the node strength calculation.

5. CONCLUSION AND FUTURE WORKS

The purpose of this study was to analyze the effect of tweet content similarity to the node strength on twitter. The tweet content similarity can be used as one of the parameters that influence the determination of user ranking. In the tuning parameter values between 0 and 1 (0 <α <1), the value of a node degree and strength affect the ranking of a node. In the tuning parameter values above 1 (α> 1), the value of strength is a major influence on the ranking of the user node, compared to node degree. Based on relationship analysis in dataset, a mention and reply relationship have more influence on the node strength compared with other relationship. In the further research, this study will consider the context of the tweet as a parameter in the node strength calculation.

ACKNOWLEDMENT :

The research was supported by Hibah Bersaing DIKTI 2015 Research Grant, awarded to Telkom University Bandung Indonesia.

REFERENCES :

[1] J. Scott, *Social network analysis*. Sage, 2012.
[2] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of information Science*, vol. 28, pp. 441–453, 2002.
[3] P. Bonacich, "Some unique properties of eigenvector centrality," *Social Networks*, vol. 29, pp. 555–564, 2007.
[4] M. E. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, p. 056131, 2004.
[5] W. Maharani and A. A. Gozali, "Collaborative Social Network Analysis and Content-based Approach to Improve the Marketing Strategy of SMEs in Indonesia," *Procedia Computer Science*, vol. 59, pp. 373–381, 2015.
[6] Z. A. Rachman, W. Maharani, and others, "The analysis and implementation of degree centrality in weighted graph in Social Network Analysis," in *Information and Communication Technology (ICoICT), 2013 International Conference of*, 2013, pp. 72–76.
[7] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, pp. 215–239, 1979.
[8] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Generalizing degree and shortest paths," *Social Networks*, vol. 32, pp. 245–251, 2010.

[9] Y. Yustiawan, W. Maharani, and A. A. Gozali, "Degree Centrality for Social Network with Opsahl Method," *Procedia Computer Science*, vol. 59, pp. 419–426, 2015.

[10] W. Maharani, A. A. Gozali, and others, "Degree centrality and eigenvector centrality in twitter," in *Telecommunication Systems Services and Applications (TSSA), 2014 8th International Conference on*, 2014, pp. 1–5.

[11] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[12] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 3747–3752, 2004.