Information diffusion model using continuous time Markov chain on social media

Firdaniza*, B N Ruchjana, D Chaerani
1Department of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

*firdaniza@unpad.ac.id

Abstract. On social media, information spreads quickly and can affect other users. In order for information to spread to many users, it is important to know how the information diffusion model or the flow of information is and who is the influencer on social media. In this paper, the Information Diffusion Model in social media was developed using the Continuous Time Markov Chain (CTMC). The tweet-retweet activity of social media users such as on Twitter can be seen as the CTMC model, because it fulfills the nature of Markov, i.e. retweets from subsequent users depend only on the current user's retweet and do not depend on the retweet history of previous users. Users engaged in retweet tweets are state of CTMC. By using the tweet-retweet network simulation data including 10 users and 4 topics, the influencer rankings were obtained. An influencer rating is determined not only by the number of retweets, but also by the time it takes to get those retweets.

1. Introduction

The use of social media as a medium for delivering information is familiar to all of us. Every social media user is free to express opinions or information in real time. The information that has been uploaded can spread to many other users and can affect many people. A user is called an influencer if his activities on social media cause the action of many other users to spread the information that the influencer has uploaded [1]. How the pattern of disseminating information on social media and how to determine influencers is better known as influence analysis has become a popular topic in social media analysis [2]. Research on influence analysis includes viral marketing [3,4] which aims to determine which users are influencers to maximize marketing. In Kempe’s work [4], the Linear Threshold model and the independent cascade model have been used. Other studies related to influence analysis on social media are recommendation of tweets based on Diffusion rate [5], dynamic influencer predictions [2], finding groups that disseminate information on social media using various methods such as Discrete Particle Swarm Optimization [6], with A Bayesian Networks Based Approach [7], with SI model [8], and with Gray Wolf Optimization Algorithm [9].

Some of these studies are based on more descriptive rather than predictive models. In the dissemination of information on social media, it is necessary to predict who is the influencer for a period of time to come. One of the prediction methods that can be used is the Markov chain. Applying Markov's nature, future circumstances can be predicted based on current conditions. For tabular data, the use prediction model was Discrete Time Markov Chain (DTMC) such as [10] predicting the properties of industrial pipe structures using the Markov chain model, but for data that increases continuously in real time, it is more suitable to use the prediction model with CTMC. Diffusion of information on social media such as Twitter fulfills the character of Markov, namely information that reaches a person only depends on
information from current users, and does not depend on previous users. Like tweet-retweeting activity on Twitter, after the first user uploads information in his tweet, then within a certain time the second user will retweet the first person and within a certain interval the third user will retweet the second user and so on, thus forming chain dissemination of information.

In this paper, an information diffusion model using CTMC was built, a model that not only involves transitions between states, but also involves the time duration of each state. In this model, the state was social media users and the duration of time was the time between tweet and retweet between one user and another. To rank influencer, several measures can be used, such as closeness (Cc), betweenness (Cb), metric on Twitter, PageRank and by topic content [11]. However this measurement is based on a static network, meaning it ranks users who have occurred at certain intervals. This ranking is subject to change based on time. Furthermore, we determine the influencer ranking dynamically for a certain period of time.

2. Method

2.1. Continuous time Markov chain (CTMC)

Let \( X(t) \) represents social media users who convey a particular information at time \( t \), then \( \{X(t), t \geq 0\} \) is a form of CTMC, the next user who will talk about the topic only depend on the current user, not depending on previous users. Formally, Ross [12] defined CTMC as a continuous-time stochastic process with discrete state space \( S \), which satisfies

\[
P\{X(t + s) = j|X(s) = i, X(\alpha) = x(\alpha), 0 \leq \alpha < s\} = P\{X(t + s) = j|X(s) = i\} \quad \forall s, t \geq 0, i, j \in S.
\]

(1)

In other words, a CTMC is a stochastic process having the Markovian that the conditional distribution of future \( X(t + s) \) given the present \( X(s) \) and past \( X(\alpha), 0 \leq \alpha < s \), depend only on the present and is independent of the past. If \( P\{X(t + s) = j|X(s) = i\} = P\{X(t) = j|X(0) = i\} \) does not depend on \( s \), then CTMC is said to be stationary or has a homogeneous transition probabilities. Let

\[
p_{ij}(t) = P\{X(t + s) = j|X(s) = i\}
\]

denote the probability that a process presently in state \( i \) will be in state \( j \) a time \( t \) later. These quantities are called the transition probabilities of the CTMC.

In contrast to DTMC which focuses on the transition probability matrix, which plays an important role in CTMC is the generator infinitesimal matrix or commonly known as the transition rate matrix. In this case, the topic spread rate is assumed to be constant, so the information diffusion model used is time-homogeneous CTMC.

Transition rate matrix \( Q \), with entries \( q_{ij} \), indicates the rate of spread of the topic from user \( i \) to user \( j \) where

\[
q_{ij} = \lim_{h \to 0} \frac{p_{ij}(h)}{h} = p_{ij}(0), \quad i \neq j.
\]

(3)

Because \( q_{ii} = p_{ij}(0), i \neq j \), then \( \sum_j q_{ij} = 0 \). The rate of transition out of state \( i \), \( v_i = q_{ii} = -\sum_{j \neq i} q_{ij} \) and for any \( i,j \), the value \( q_{ij} = v_i p_{ij} \) is called the instantaneous transition rate.

Thus the transition rate matrix \( Q \) can be written in the form

\[
Q = \begin{pmatrix}
-\sum_{j \neq 0} q_{0j} & q_{01} & q_{02} & \cdots \\
q_{10} & -\sum_{j \neq 1} q_{1j} & q_{12} & \cdots \\
q_{20} & q_{21} & -\sum_{j \neq 2} q_{2j} & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}.
\]

(4)

Let \( T_i \) is the time for the spread of topics from user \( i \), namely the length of time a topic remains in user \( i \) before it reaches other users, so that \( P\{T_i > s + t|T_i > s\} = P\{T_i > t\} \) for each \( s, t \geq 0 \). Therefore, the random variable \( T_i \) is memory-less and has an exponential distribution with mean \( v_i \). In this case, it is obtained \( v_i = \frac{1}{E(T_i)} \). From the property \( q_{ij} = v_i p_{ij} \) then the value of \( q_{ij} \) is determined by equation (5).
\[ q_{ij} = \sum_m v_i e^{-v_i t_{ij(m)}} \]  
with \( m \) is the number of topics (information) disseminated by user \( i \) to user \( j \), and \( t_{ij(m)} \) is the time between user \( i \) uploading topic \( m \) and user \( j \) retweeting the same topic [2].

The transition probability matrix \( P(t) \) can be estimated based on the Chapman-Kolmogorov equation [12],

\[ p_{ij}'(t) = \sum_{k \neq i} q_{ik} p_{kj}(t) - v_i p_{ij}(t). \]  
In matrix form, equation (6) can be written in form

\[ P'(t) = QP(t). \]  
Equation (7) is a differential equation with solutions

\[ P(t) = P(0)e^{Qt}. \]  
Because \( P(0) = I \), then

\[ P(t) = e^{Qt} = \sum_{n=0}^{\infty} (Qt)^n/n!. \]  
For \( n \) large enough, the equation (9) can be approached with

\[ P(t) = \lim_{n \to \infty} \left(1 + \frac{t}{n}Q\right)^n. \]  

2.2. User Diffusion Size
From the transition probability matrix that has been obtained according to the desired time period based on equation (10) the diffusion size of user \( i \) for time period \( t \), is determined from the multiplication of the number of transitions probability from user \( i \) to other users with the number of users who retweet the topics uploaded by user \( i \). Measure the diffusion of user \( i \) over all topics, Li [2] using a formula

\[ D_{i,t} = \sum_j p_{ij}(t) n_i \]  
with \( p_{ij}(t) \) represents the probability of transition from user \( i \) to user \( j \) at time \( t \), \( n_i \) states the number of users who retweet after \( i \) upload a tweet within a certain time . The influencer rating is determined by the size of the user's diffusion.

2.3. Research Procedures
The taken steps in building an information diffusion model with time-homogeneous CTMC and determining the influencer ranking are considered as follows:
1. Tweet-retweet network
   In this case, the number of users involved indicates the number of states, \( T_i \) is user \( j \)'s retweet time after user \( i \) uploads his tweet.
2. Determines the rate of transition out of state \( i \) for each topic, \( v_i = \frac{1}{E(T_i)} \).
3. Calculate the transition rate matrix \( Q \) that satisfies equation (4), with entries \( q_{ij} \) using equation (5).
4. The transition matrix \( P \) can be calculate from the diffusion model using equation (10).
5. Determine the diffusion size using equation (11) to obtain the influencer rank.

3. Result and Discussion
3.1. Research Data
At the time of this paper, the real tweet-retweet data from Twitter that formed the network could not be obtained simply with the Twitter API. Therefore, in this paper the information diffusion model with CTMC is applied to the simulation data. The simulation data collection is carried out in the following steps:
1. Determine the amount of information (topics) to be disseminated (tweet) via social media; for example 4 topics.
2. Estimate the number of social media users involved in the dissemination of the four topics; for example there are 10 users (A,B,C,D,E,F,G,H,I,J).
3. Specify the user who will upload the first tweet for each topic.
4. Determine the time of distribution of each topic for the first time; in this case, suppose topic 1 is distributed at 7:00 a.m, topic 2 at 8:00 a.m, topic 3 at 9:00 a.m, and topic 4 at 10:00 a.m.
5. Specify which users retweet each topic, after the topic was first shared and the time each user retweeted.

The tweet-retweet simulation data between Twitter users, namely A,B,C,D,E,F,G,H,I,and J with the 4 uploaded topics are presented in Table 1.

| Time  | User | Tweet   | Retweet          |
|-------|------|---------|------------------|
| 07.00 | B    | Tweet 1 | Retweet B topic 1|
| 07.10 | A    |         | Retweet B topic 1|
| 08.00 | G    | Tweet 2 | Retweet A topic 1|
| 08.10 | E    |         | Retweet E topic 1|
| 08.30 | D    |         |                  |
| 09.00 | B    | Tweet 3 | Retweet E topic 3|
| 09.20 | E    |         | Retweet B topic 3|
| 09.40 | C    |         | Retweet G topic 2|
| 10.00 | D    |         | Retweet E topic 3|
| 10.05 | C    | Tweet 4 | Retweet D topic 3|
| 10.10 | H    |         | Retweet D topic 1|
| 10.10 | A    |         | Retweet C topic 2|
| 10.15 | F    |         | Retweet H topic 1|
| 10.20 | G    |         | Retweet F topic 3|
| 10.25 | J    |         | Retweet F topic 1|
| 10.25 | B    |         | Retweet A topic 2|
| 10.55 | I    |         | Retweet G topic 3|
| 11.05 | E    |         | Retweet B topic 2|
| 11.05 | H    |         | Retweet I topic 3|
| 11.20 | E    |         | Retweet C topic 4|
| 11.25 | I    |         | Retweet E topic 2|
| 11.40 | J    |         | Retweet E topic 4|
| 11.50 | B    |         | Retweet J topic 4|
| 13.50 | A    |         | Retweet B topic 4|

3.2. Tweet-retweet Network
Based on Table 1, tweet-retweet activities network can be created from 10 users with 4 uploaded topics as shown in Figure 1.
Figure 1. Tweet-retweet network with 4 topics and 10 users from simulation data

The node represents the user and the arrow indicates the retweet activity with the retweet time. In this case, B $\rightarrow$ A means that A retweetes B so that the information conveyed by B can be received by A within the time stated above the arrow (in minutes).

3.3. Transition Rate Matrix for Homogeneous CTMC Diffusion Models

The resulting transition rate matrix

$$Q = \begin{pmatrix}
-0.000644 & 0.000314 & 0.000000 & 0.00014 & 0.000014 & 0.000091 & 0.000012 \\
0.000026 & -0.000983 & 0.000000 & 0.000107 & 0.000014 & 0.000047 & 0.000013 \\
0.000155 & 0.000141 & -0.000465 & 0.000000 & 0.000000 & 0.000000 & 0.000072 \\
0.000000 & 0.000000 & 0.000000 & -0.001492 & 0.000064 & 0.000048 & 0.000012 \\
0.000000 & 0.000143 & 0.000000 & 0.000032 & 0.000000 & 0.000000 & 0.000073 \\
0.000000 & 0.000000 & 0.000000 & 0.000000 & -0.004391 & 0.000000 & 0.000035 \\
0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & -0.000544 & 0.000020 \\
0.000032 & 0.000177 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\
0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\
0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000
\end{pmatrix}$$

3.4. Diffusion Size Results

The size of the diffusion of information was calculated based on equation (11) for 3 specific time periods (i.e. $t = 50, 500, 2000$), and the results are shown in Table 2.

Table 2. User diffusion measure

| User | Number of retweets | Diffusion measure ($t = 50$) | Diffusion measure ($t = 500$) | Diffusion measure ($t = 2000$) |
|------|--------------------|-------------------------------|-------------------------------|-------------------------------|
| A    | 8                  | 0.253444                      | 2.193219                      | 5.698915                      |
| B    | 15                 | 0.717702                      | 5.716307                      | 11.959944                     |
| C    | 8                  | 0.183897                      | 1.659882                      | 4.841387                      |
| D    | 7                  | 0.50319                       | 3.679313                      | 6.612732                      |
From Table 2, it can be seen that user E gets the highest diffusion size value, meaning that user E is the top ranked influencer in the three different time periods, although user E is not the most retweeted. The order of influencer ranking based on the diffusion size value is shown in Table 3.

Table 3. Influencer rating

| Rating | User (for $t=50$) | User (for $t=500$) | User (for $t=2000$) |
|--------|------------------|--------------------|---------------------|
| 1      | E                | E                  | E                   |
| 2      | F                | B                  | B                   |
| 3      | B                | D                  | D                   |
| 4      | H                | F                  | A                   |
| 5      | D                | A                  | C                   |
| 6      | A                | H                  | G                   |
| 7      | G                | C                  | F                   |
| 8      | C                | G                  | H                   |
| 9      | I                | I                  | I                   |
| 10     | J                | J                  | J                   |

Table 3 shows the ranking of influencers for three different time periods, namely $t = 50$, $t = 500$, $t = 2000$. It can be seen that for rank 2 to rank 8, the order is different for the three different time periods. This shows that the level of influence of an influencer can change depending on the time.

4. Conclusion
The information diffusion model in social media can be modeled with CTMC if there is a tweet-retweet time network. From the example of the retweet tweet time network data case, for three different time periods, the first influencer rank is user E, while the next rank changes according to the time period.

Acknowledgments
This research is funded by Doctoral Dissertation Research Grant from Universitas Padjadjaran Indonesia for the year of 2019 contract number 5879/UN6.D/LT/2019.

References
[1] Peng S, Zhou Y, Cao L, Yu S, Niu J and Jia W 2018 Influence analysis in social networks: A survey Journal of Network and Computer Applications 106 17-32
[2] Li J, Peng W, Li T, Li Q, and Xu J 2014 Social network user influence sense-making and dynamics prediction Expert Systems with Applications 41 5115–5124
[3] Domingos P and Richardson M 2001 Mining the network value of customers Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 57–66
[4] Kempe D, Kleinberg J and Tardos É 2015 Maximizing the spread of influence through a social network Theory of Computing 11 105–147
[5] Song X, Chi Y, Hino K and Tseng B L 2007 Information flow modeling based on diffusion rate for prediction and ranking 16th International World Wide Web Conference WWW2007 191–200

[6] Gong M, Yan J, Shen B, Ma L and Chai Q 2016 Influence maximization in social networks based on discrete particle swarm optimization Information Sciences 367–368 600–614

[7] Varshney D, Kumar S and Gupta V 2017 Predicting information diffusion probabilities in social networks: A Bayesian networks based approach Knowledge-Based Systems 133 66–76

[8] Ghani N A, Hamid S, Hashem I A T, and Ahmed E 2019 Social media big data analytics: A survey Computers in Human Behavior 101 417–428

[9] Zareie A, Sheikhamadi A and Jalili M 2020 Identification of influential users in social network using gray wolf optimization algorithm Expert Systems with Applications 142 112971

[10] AL-Alawi M, Bouferguene A and Mohamed Y 2018 Random generation of industrial pipelines data using Markov chain model Advanced Engineering Informatics 38 725–745

[11] Riquelme F and González-Cantergiani P 2016 Measuring user influence on Twitter: A survey Information processing & Management 000 1-27

[12] Ross S 2010 Introduction to Probability Models 10th editions (USA: Elsevier Inc UK)