Evaluation of Network Structure Using Similarity of Posts on Twitter

Yusuke Sato[1(✉)], Kohei Otake[2], and Takashi Namatame[3]

[1] Graduate School of Science and Engineering, Chuo University,
1-13-27 Kasuga, Bunkyo-Ku, Tokyo 112-8551, Japan
al3.bthf@chuo-u.ac.jp
[2] School of Information and Telecommunication Engineering, Tokai University,
2-3-23, Takanawa, Minato-Ku, Tokyo 108-8619, Japan
otake@indsys.chuo-u.ac.jp
[3] Faculty of Science and Engineering, Chuo University,
1-13-27 Kasuga, Bunkyo-Ku, Tokyo 112-8551, Japan
nama@indsys.chuo-u.ac.jp

Abstract. Social networking service (SNS) is very popular in our lives, with expanding internet environments and mobile device. Through the SNS, user can submit their opinion or reputation freely, anytime and anywhere. These activities are getting great attention on a various business scenes in recently. Twitter is one of the most popular SNS, and used by numerous people in the world. In addition, since various information is posted on Twitter, it is expected to be utilized as a business strategy, and there have been many studies on the marketing using Twitter data. Moreover, we can get some information about user's network in Twitter. In this research, we attempt to evaluate the network structure using similarity of post on Twitter. We created the user network using similarity of posts mentioned about four titles of Japanese TV drama, and we grasped the post categories that is easy to get user's interest. From the result, we discussed the difference between TV drama and suggestions for promotion strategies of TV drama production company.

Keywords: Social Networking Service · Twitter · Network analysis
Graph representation · Natural Language Processing

1 Introduction

Social Networking Service (SNS) is very popular in our lives, with the development of information technology and mobile device such as smart phone. By using SNS, it is possible that user can share various information through their friends freely anywhere and anytime. From this reason, the information transmission between consumer on SNS is actively performed, and it sometimes affects the real world. Therefore, SNS have gotten a lot of attention in the business scene as a promotion and marketing tool in recent years [1, 2]. Furthermore, SNS is regarded as an important tool that make it possible to transmit information to many people efficiently in various industries such as retailers, EC sites, political activities and so on.

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10914, pp. 315–329, 2018.
https://doi.org/10.1007/978-3-319-91485-5_24

Twitter is one of the most popular SNS in the world. By using Twitter, users can perform various actions such as “Tweet” and “Retweet”. Moreover, there are a variety of information such as user’s opinion and reputation on Twitter. Using Twitter or other SNS data including those information, we can elucidate various phenomenon occurring on Twitter (e.g. information diffusion and network of friendship between users). Therefore, there have been numerous research related to marketing activities using SNS data. To understand the user behavior on SNS, various researchers have applied studies [3–5]. In these studies, they targeted posts data about specific products or the structure of SNS itself and analyzed the SNS data. On the other hand, regarding the TV drama targeted in this research, it is inferred that there are various phenomenon caused by audience (such as post activity in real time of broadcasting time or the period from the episode to the next episode). Therefore, it can be said that elucidation of post activities on Twitter by audience is important analysis for the promotion strategies of TV drama.

2 Related Studies and Our Purpose

In this section, first, we introduce some related studies about SNS analysis. Next, we show the objective of this study.

Yang et al. [3] analyzed the information diffusion phenomenon on Twitter. Especially, they proposed model that able to capture the three main specifics of information diffusion (speed, scale and range) using survival analysis. As the result, they found that some specific of the tweets can predict the diffusion phenomenon. Matsumura et al. [4] proposed an influence diffusion model that express how articles and words were diffused. As the result, using the above model, they identified influencer who post information that gets interest of others and words reflecting consumer insights. Matsuo et al. [5] investigated the network structure of the user networks created on the largest SNS site in Japan. Moreover, they confirmed the structure of the community formed by the relationship of users on the network.

In this study, we focus on Twitter data and attempt to evaluate the network structure among users using similarity of posts. For the analysis, we used tweet data posted about four Japanese TV drama. The information about TV drama are frequently posted on Twitter by audience, and its contents are various (e.g. contents about story, actor or actress, etc.). Focusing on its situation, we also try to evaluate the users’ interest to post categories by dividing the user network into several communities and comparing network indexes. From these results, it is possible to identify the post category that is easy to get interest among users, and it is expected to obtain a useful suggestion for the promotion strategy performed by the TV drama production company.

3 Data Summary

We targeted four titles of Japanese TV drama and collected tweets data posted about these titles. In this study, we selected these four titles based on broadcasting period and evaluation by ranking site. We used hashtags and keywords (drama titles) and collected the data by using the application programing interface (API) of Twitter. Consequently,

we collected about 577,000 tweets in total. These tweet data were posted during broadcasting period of each TV drama and include User ID, tweet date and time, tweet text, the number of favorite and Retweets and so on. Summary (e.g. broadcast period (time zone: JST) and category of each title) of targeted TV drama and collected tweets data are shown Tables 1, 2 and 3.

Table 1. Broadcast period, frequency and time of targeted Japanese TV drama

Title	Period	Broadcast frequency	Broadcast time
A	2017/04/18 ~ 2017/06/20	Every Tuesday	PM10:00 ~ PM11:00
B	2017/04/17 ~ 2017/06/26	Every Monday	PM09:00 ~ PM10:00
C	2017/04/14 ~ 2017/06/16	Every Friday	PM10:00 ~ PM11:00
D	2017/04/16 ~ 2017/06/18	Every Sunday	PM09:00 ~ PM10:00

Table 2. Summary of targeted Japanese TV drama

Title	Content	Category
A	TV drama based on Japanese comic	Love romance
B	TV drama based on Japanese novel	Mystery
C	TV drama based on Japanese novel	Mystery
D	TV drama created by Japanese TV station originally	Drama

Table 3. Summary of collected tweet data

Title	The number of tweets	The number of unique posted users
A	65,153	13,417
B	288,004	27,425
C	104,557	25,007
D	92,147	17,430

4 Evaluation of Network Structure Using Posts Similarity

In this study, we performed analysis in 3 steps. In the 1st step, we extracted representative 50 keywords by each drama and classified these keywords into 13 post categories by Natural Language Processing. In the 2nd step, we visualized the network that express the posting relationship between users and post categories. Especially, we created incidence matrix and bipartite graph by using the weight which means user's posting importance for each category. In the final step, we divide the above network into several communities. Targeting these communities, we grasped post categories which is mainly posted by users of each community and compared the network indicators such as network density between communities. From above results, we discuss the user's interest for each post category.

4.1 Identify and Classify Keywords

In the 1st step, we identify the keywords of each title and classify these keywords into post categories. Firstly, we performed morphemes analysis to divide all tweet texts of each title into columns of morphemes (minimum elements constituting sentences). Morphological analysis is a commonly used method for dividing the natural language (text data) into morphemes and discriminating parts of speech and the like of each morpheme. It is need became all of letters are connected Japanese sentence. In the morphological analysis, information such as parts of speech words defined in grammar and the dictionary is used for dividing process. In this study, we used the R language to perform morphological analysis. Moreover, the dictionary used for analysis was Mecab [6], a Japanese morpheme dictionary.

Targeting terms extracted by morphological analysis, we selected three parses (nouns, verbs, adjectives), and identified the keywords of each title using the *tfidf* method [7]. The *tfidf* method is a type of index of word weighting and is calculated by the product of *tf* (term frequency) and *idf* (inverse document frequency). The *tfidf* values of word i in the document j is calculated by the following equations.

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{1}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_S n_{S,j}} \tag{2}$$

$$idf_i = \log \frac{|D|}{|\{d : d \in t_i\}|} \tag{3}$$

where $n_{i,j}$ is the occurrence frequency of word i in document j, $\sum_S n_{S,j}$ is the summation of count of all the words in document j, $|D|$ is the total number of documents, $|\{d : d \in t_i\}|$ is the number of documents that contain word i.

We defined the top 50 words which have high *tfidf* values as the keywords for each title and classified these keywords into 13 post categories based on the its meaning. The names of those post categories and its description are shown in Table 4. The number of keywords of each post category by each title are shown in Table 5.

From above result, it is found that whether post categories were posted or not differ depending on title of TV drama. For example, "Location" and "Other_TVshow" are posted on only title D, "Broadcasting_station" were posted on title A and B.

4.2 Bipartite Graph of Users and Categories

In the 2nd step, we created the network among users by the bipartite graph to grasp posting relationship between users and post categories. In this research, we targeted the top 500 users who posted frequently during each title of TV drama. The network graph was constructed by using the relationship which is whether user posted the post category or not and their weight for post category based on post ratio. In particular, we created the incidence matrix and the bipartite graph as the following procedure.

Table 4. The names of categories and its description

Category name	Description
L_character	Leading character
S_character	Supporting character
L_actor/actress	Leading actor or actress
S_actor/actress	Supporting actor or actress
Location	Shooting place of each title
Emotion (+)	Positive emotion about content of each episode
Emotion (−)	Negative emotion about content of each episode
Main_topic	The main topic of each title
Contents_topic	The subtopic of each title
Broadcasting_type	The word meaning the type of episode (e.g. the end of the series of each title)
Broadcasting_station	The station which broadcasts each title
Other_TVshow	TV programs other than each TV drama
Title	Title of each TV drama

Table 5. The number of keywords in each title[a]

Category	Title			
	A	B	C	D
L_character	4	7	6	5
S_character	2	1	1	0
L_actor/actress	7	5	10	11
S_actor/actress	4	1	3	3
Location	0	0	0	1
Title	4	2	2	1
Emotion (+)	10	14	9	10
Emotion (−)	4	5	4	2
Main_topic	2	3	1	4
Contents_topic	9	12	3	11
Broadcasting_type	6	8	16	7
Broadcasting_station	1	1	0	0
Other_TVshow	0	0	0	2

[a] Each keyword can belong to multiple categories.

1. We calculated the frequency of keywords for all tweet texts of targeted users. Here, we counted presence or absence of keywords, without consideration for that same keywords appears more than once in a tweet text.
2. Based on the post category of each keyword, we calculated the post ratio of each post category by each user.
3. We calculated a weight $W_{i,j}$ for post category j of user i in accordance with the following conditions. In addition, we defined the matrix constituted by these

weights as the incidence matrix X which means the user's posting importance for each post category.

$$W_{i,j} = \frac{r_{i,j} \times 100}{T_j} \tag{4}$$

where $r_{i,j}$ is post ratio of post category j of user i, T_j is the number of terms which belongs to post category j.

4. Based on the incidence matrix X, we created bipartite graphs in which nodes are users and post categories, weights of edges are $W_{i,j}$.

For the visualizing the bipartite graph, we used Fruchterman-Reingold algorism [8]. Fruchterman-Reingold algorism is a method based on dynamic model for visualizing network. This algorithm has a feature to arrange the connected nodes close to and to locate unconnected nodes far from each other. Figures 1, 2, 3 and 4 shows the bipartite graph of users and post categories of each title of TV drama. In the bipartite graph of each title, dark gray edges express high weighted edges which have weights in the top quartile points of all weights. In addition, Fig. 5 shows the result of calculating the number of edges connected to the category node that is the number of users who posted the category by weight.

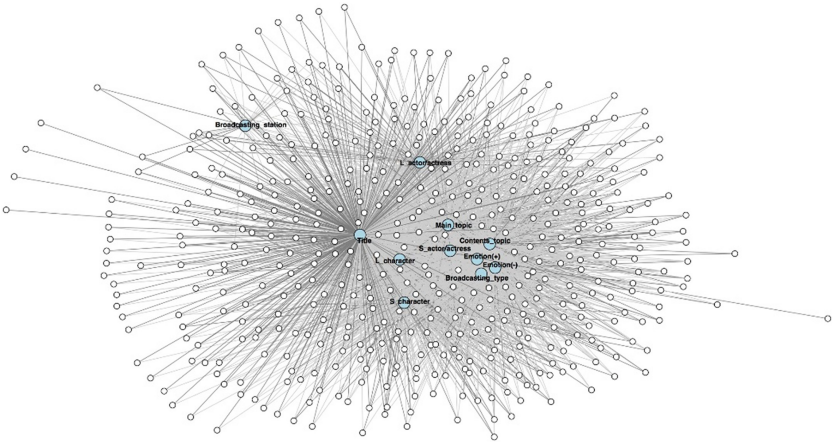

Fig. 1. Bipartite graph of users and post categories of title A

From above result, it turned out that the category most posted from users on title A and B is the "Title", users of title C and D frequently post about "L_character". In addition, about those categories, we can see that the number of edges with high weight (dark gray edge on the bipartite graph) is more than the number of edges with low weight (light gray edge on the bipartite graph). On the other hand, we can see that other posting categories show reverse trends, in other words, most edges are composed of edges with low weight.

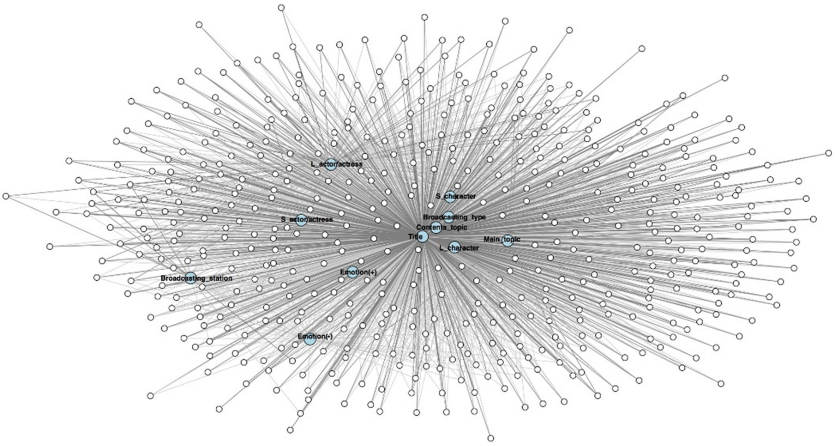

Fig. 2. Bipartite graph of users and post categories of title B

Fig. 3. Bipartite graph of users and post categories of title C

4.3 User Network Using Similarity of Post Categories

In the final step, we created user network based on their similarity of the post categories. Moreover, dividing the network into several communities, we evaluated the relationship between network indicator and post categories of each community. In order to detect some communities from the network, we need adjacency matrix rather than incidence matrix. Firstly, by using incidence matrix X of previous step, we created new incidence matrix X'. Especially, we created the new matrix based on following condition to more strictly define whether the user posted each posting category or not.

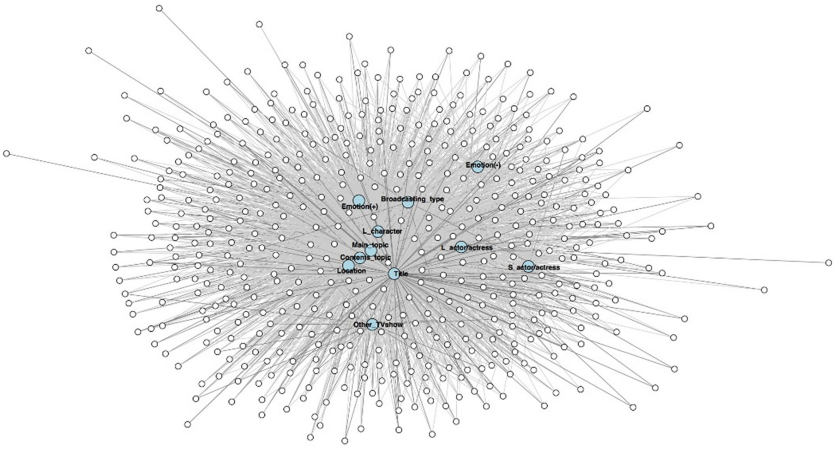

Fig. 4. Bipartite graph of users and post categories of title D

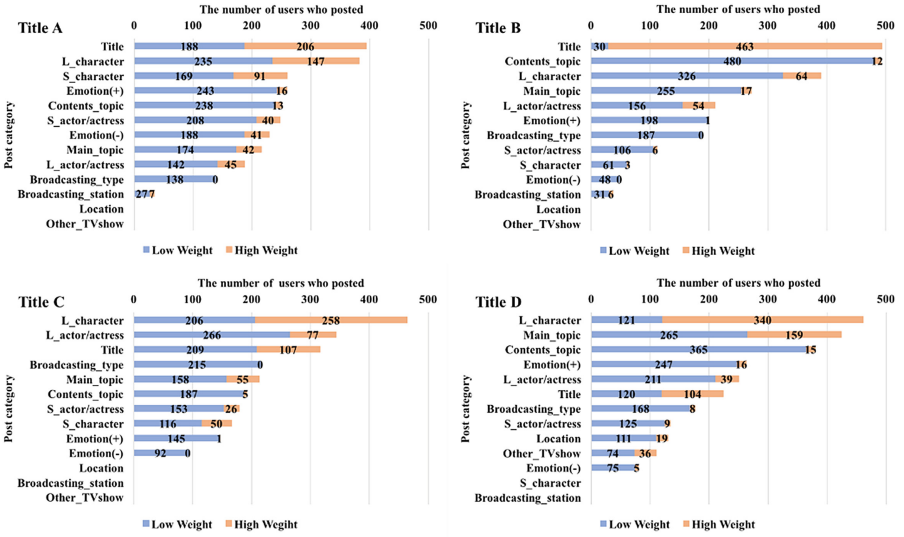

Fig. 5. The number of users who posted about each category

$$W'_{i,j} = \begin{cases} 1 & W_{i,j} \geq V_{3rd,j} \\ 0 & W_{i,j} < V_{3rd,j} \end{cases} \tag{5}$$

where $W_{i,j}$ is the weight for post category j of user i which is calculated on the previous step, $V_{3rd,j}$ is the weight in the top quartile points of all weights of post category j.

So, incidence matrix X' means that only users with high weights for each category are redefined as "Users who posted the category".

Next, we created adjacency matrix Y which describes the similarity of post categories between users. Especially, we converted the incidence matrix X' into adjacency matrix Y as follows.

$$Y = X'X'^{t} \tag{6}$$

where X'^{t} is the transposed matrix of incidence matrix X', the all diagonal elements of the adjacency matrix Y are 0.

Table 6. The network indexes using user similarity of post category of each title

Title	The number of nodes	The number of edges	Density
A	499	57,745	0.462
B	487	51,223	0.410
C	491	56,260	0.450
D	498	60,743	0.486

From the above transformation, the adjacency matrix Y means the similarity of the post category between users. Using adjacency matrix Y, we created and visualized the user network. For the visualization, we deleted isolated nodes not connected to any other node out of 500 user nodes of each title. The network indexes of user network of each title are shown Table 6.

From Table 6, regarding the user network of all titles, we can see that the nodes are connected with moderate density compared to general social networks of friendship between users.

Finally, we divide these user network into several communities by spin glass method [9]. Spin glass method is one of the most popular method to detect communities from the network. This method assigns each node to the community so as to minimize the Hamiltonian function expressed by the following equation.

$$\begin{aligned}\mathcal{H}(\{\sigma\}) = &-\sum\nolimits_{i\neq j} a_{ij}Y_{ij}\delta(\sigma_i,\sigma_j) + \sum\nolimits_{i\neq j} b_{ij}(1-Y_{ij})\delta(\sigma_i,\sigma_j) \\ &+ \sum\nolimits_{i\neq j} c_{ij}Y_{ij}\left[1-\delta(\sigma_i,\sigma_j)\right] - \sum\nolimits_{i\neq j} d_{ij}(1-Y_{ij})\left[1-\delta(\sigma_i,\sigma_j)\right]\end{aligned} \tag{7}$$

where Y_{ij} denotes the adjacency matrix of the graph, σ_i denotes the group index of node i in the graph, and $a_{ij}, b_{ij}, c_{ij}, d_{ij}$ denote the weights of the individual contributions.

In addition, we determined the number of communities by using the modularity Q [10]. This value is an index for evaluating the accuracy of community detection and calculated by following equation. We can define the dividing the community has high Q value as appropriate detection.

$$Q = \frac{1}{2M}\sum_{i \neq j}\left[A_{ij} - \frac{k_i k_j}{2M}\right]\delta(\sigma_i, \sigma_j) \tag{8}$$

where k_i is the number of edges of node i, M is summation of the number of edges which exist in the network.

Table 7. The result of community detection of each title

Title	The number of communities	Modularity
A	3	0.22
B	3	0.28
C	3	0.23
D	3	0.22

We can conduct appropriate community detection by adopting the division result of high Q value. As the result of community detection based on modularity Q, the user network was divided three community by each title. The modularity Q of community detection of each title are shown Table 7.

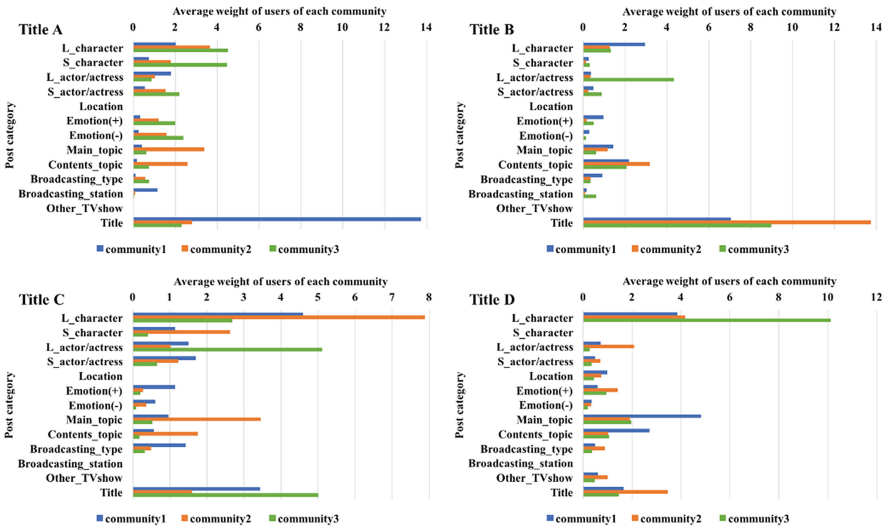

Fig. 6. The average weight for each category of users belonging to each community

Targeting these communities, we attempt to grasp the post categories which is mainly posted by users who belong to each community. Figure 6 shows the user's average weight to each post category by each community of each title. By using result of Fig. 6, we defined the topic of each community shown in Table 8.

Table 8. The topic of each community of each title

Title	Community	Topic
A	1	TV Drama Headline
	2	Content
	3	Character and Emotion (±)
B	1	Content and Emotion (+)
	2	Content
	3	TV Drama Headline
C	1	Content and Emotion (+)
	2	Content
	3	TV Drama Headline
D	1	Content and Emotion (+)
	2	Content
	3	TV Drama Headline

From Table 8, it turned out that the topics posted for each community differed. Considering all the communities, there are six types of topics. "TV Drama Headline" is user group posted tweet including words that have potential of becoming headline of TV drama such as title names of TV drama and actor or actress names. "Content" is the topic which is mainly posted both content of episode and character of TV drama. "Character and Emotion" or "Content and Emotion" is the topic that have been posted the characters of TV drama or content of the episode with user's emotion. However, there are not only positive emotions but also negative emotions. Regarding community 2 of title D, this community posted about actor or actress names, title names and other TV show names. Therefore, we defined this community as the user group which mainly posted about promotion of TV drama, named "Promotion". Moreover, about community 3 of title D, we named "Actor or Actress" which is user group posted only leading actor or actress. As the total tendency, it is found that user's emotion, regardless positive or negative, are posted with content of episode or character of TV drama.

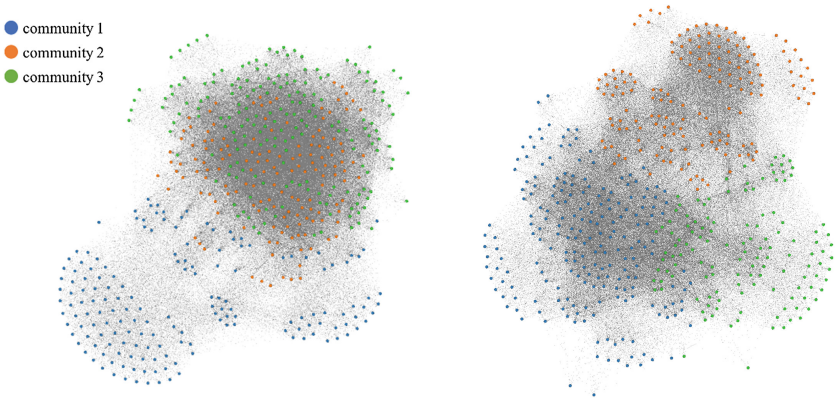

Fig. 7. User network using similarity of post category of title A (left) and B (right)

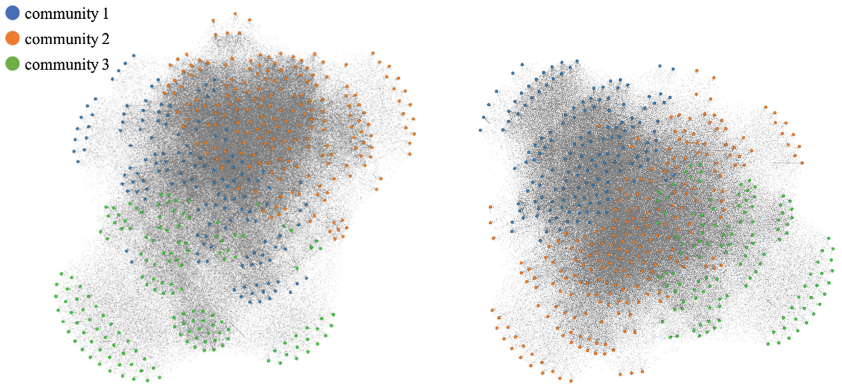

Fig. 8. User network using similarity of post category of title C (left) and D (right)

Table 9. Network indexes of each community of title A

	Community 1	Community 2	Community 3
Topic	TV Drama Headline	Content	Character and Emotion (±)
The number of nodes	165	148	186
Density	0.662	0.912	0.777
Average degree	108.727	134.202	143.817

Table 10. Network indexes of each community of title B

	Community 1	Community 2	Community 3
Topic	Content and Emotion (+)	Content	TV Drama Headline
The number of nodes	206	153	128
Density	0.676	0.894	0.907
Average degree	138.660	135.921	115.296

Table 11. Network indexes of each community of title C

	Community 1	Community 2	Community 3
Topic	Character and Emotion (±)	Content	TV Drama Headline
The number of nodes	142	198	151
Density	0.771	0.777	0.804
Average degree	108.845	153.111	120.649

Table 12. Network indexes of each community of title D

	Community 1	Community 2	Community 3
Topic	Content and Emotion (−)	Promotion	Actor or Actress
The number of nodes	153	223	122
Density	0.916	0.645	1.000
Average degree	139.268	143.210	121.000

Furthermore, by using network indexes of each community, we evaluate the user's interest to the topic of each community. We used the density and average degree of network as network indexes. Figures 7 and 8 show the user network using similarity of post category and Tables 9, 10, 11 and 12 show network indexes calculated by each community by each title.

From the above result, even if it is a similar topic, it turned out that there is a difference in the density of the user for each title, that is, the degree of user's interest for the topic. In terms of the result of each title, in the title A, there are not many users who pay attention to "TV Drama Headline" and "Character and Emotion (±)". On the other hand, topic of "Content" gets a lot of interest of users. About title B, it turned out that "Content" and "TV Drama Headline" topic is posted by relatively large number of users.

Regarding title C, there aren't so many users who pay attention to "Character and Emotion (±)" and "Content" topic. In the title D, many users are interested in "Content and Emotion (−)" and "Actor or Actress". Moreover, a few users posted about topic "Promotion".

5 Discussion

First of all, we discuss posting categories that attract users' interest in each title. In the title A, since the users of community 2 are most closely connected, the topic on the contents of the TV drama attracts audience's interest. It is inferred that the reason for this result is that title A is a love romance TV drama including elements such as affair, and audience actively posts the contents of the episode because its story is unpredictable.

From the results of title B, the community 3 has the highest density among users. Therefore, it can be said that topics (such as title names and actor names) that have potential of becoming headlines of TV drama are actively posted. It is assumed that users are pay attention to the actor or actress and its title name rather than the contents of the TV drama because the leading actor of title B is a member of a popular idol group in Japan.

About title C, the densities of all the communities are moderate, there is not much different between them. In other words, it can be said that users are post the same importance degree for any topic and it is a TV drama that has been posted about various kinds of topics in well balance.

Regarding title D, community 1 and 3 has the high user's density. Since this title was broadcasted in the TV slot that is broadcast station and broadcast time zone in which many masterpiece TV dramas were broadcasted in the past, it is inferred that the audience pay much attention to the topic related to contents of episodes and actor or actress. In particular, community 3 is the user group that emphasizes on only actor or actress, it is assumed that there are a certain number of fans of actor or actress in the community. Furthermore, in title D, there is only one user community that mainly post the topic related to the promotion of TV drama. However, it turns out that the density of that community is not so high. This is presumably because not only whether the user is a fan of the actor (actress) but also whether user watch other TV programs also influences the importance of "Promotion" topic.

In addition, as the overall knowledge, it was found that the emotion (regardless of positive or negative) of the users was posted with the contents of episodes and characters of the TV drama. It can be said that this is a natural result as a topic on which audience express opinions.

6 Conclusion

In this research, targeting four titles of Japanese TV drama, we evaluated the network structure using similarity of posts on Twitter. Even if users posted about same title of TV drama, it turned out that there are differences of user's importance to post categories among communities by dividing the user network into several communities. Moreover, it also found that there are differences among four titles of TV drama as well. It is expected to utilize these results for the strategies for promotion or marketing on Twitter of companies related to each TV drama.

As the future work of our research, we need to evaluate user's interest for topic from various point of view such as "who are the main users in the same community?" and "what topics are easy to post simultaneously with other topics?". In addition, it is possible to obtain more useful suggestions as a promotion strategy by using follow or follower relations data among users on Twitter in combination.

Acknowledgment. We thank Rooter Inc. for providing valuable datasets and for their useful comments.

References

1. Elisabeta, I., Ivona, S.: Social media and its impact on consumers behavior. Int. J. Econ. Pract. Theor. **4**(2), 295–303 (2013)
2. Sitaram, A., Bernardo, A.H.: Predicting the future with social media. Computing **25**(1), 492–499 (2010)
3. Jiang, Y., Scott, C.: Predicting the speed, scale, and range of information diffusion in Twitter. In: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, ICWSM, vol. 10 (2010)

4. Matsumura, N., Yamamoto, H., Tomozawa, D.: Finding influencers and consumer insights in the blogosphere. In: International Conference on Weblogs and Social Media, Seattle, Washington (2008). (in Japanese)
5. Matso, Y., Yasuda, Y.: How relations are built within a SNS World: Social network analysis on Mixi. Trans. Jpn. Soc. Artif. Intell. **22**(5), 531–541 (2007). (in Japanese)
6. MeCab. http://taku910.github.io/mecab/. 23 Feb 2018
7. Ricardo, A.B., Berthier, A.R.: Modern Information Retrieval: The Concepts and Technology Behind Search, 2nd edn. Addison-Wesley Professional, Harlow (2011)
8. Thomas, M.J.F., Edward, M.R.: Graph drawing by force-directed placement. Softw. Pract. Experience **21**(11), 1129–1164 (1991)
9. Joerg, R., Stefan, B.: Statistical mechanics of community detection. Phys. Rev. E **74**(1), 016110 (2006)
10. Mark, E.J.N., Michelle, G.: Finding and Evaluating Community Structure in Networks. Phys. Rev. **69**(2), 026113 (2004)