東海大学紀要情報通信学部
Vol.11,No.2,2018,pp.34-44

論文

A Study on Cluster Variation of Twitter Hashtags during a Whole Day

by

Nartlada BHAKDISUPARIT[*1] and Iwao FUJINO[*2]

(received on Sep. 28, 2018 & accepted on Jan. 10, 2019)

Abstract

Our study aims to verify the cluster variation of Twitter hashtags according to their word distributions. By gathering tweets from Twitter sample data and saving tweets in a time period of each hour separately, we extract hashtags and calculate their word distributions for each time period. Furthermore, by using Ward's hierarchical method, we cluster hashtags to a structure of groups according to the Jensen-Shannon divergence between any two hashtags and show the clustering results with dendrogram, which represents the hierarchical structure of all hashtags. Finally, we verify the differences of clustering results between each time period to analyze the details of cluster variation. The achievement of this study will contribute to develop marketing campaign or advertisement on Twitter.

Keywords: *Twitter hashtags, word distribution, Jensen-Shannon divergence, Ward's hierarchical clustering method, marketing on Twitter*

1. Introduction

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. It has one or more of the following characteristics: high volume, high velocity, or high variety. Big data comes from many sources e.g. sensors, video/audio, log files, transactional applications, web, online social media, and web browser cookie data. Analysis of big data will help us to provide more accurate prediction for to support decision-making process [1].

Nowadays the online social network is typical place that big data playing a significant role in many ways in human life. Twitter is a microblogging service that allows people to communicate with a short message called Tweet. Twitter had grown from 30 million users in 2010 to 335 million users in the first quarter of 2018 [2]. Twitter is different from other online social network. Without any relation between each pair of users beforehand, Twitter relationship model provides a simple but powerful tool to create and maintain a specific community. In the community, users can use hashtag to mention to their specific topic, where hashtag is a specific string beginning with symbol # (sharp) on Twitter.

Twitter's relationship model allows user to keep up with the latest happening of other users. By using hashtag users can share their comments and information with Twitter's online community. Twitter also helps people raise campaign for their charity events. One of a successful campaign which stared from hashtag was #nomakeupselfie. This hashtag raised 8 million Euros for cancer research in only one week. This hashtag didn't start for charity project at the beginning, but after it went on trend its social media team decided to involve and ask people to donate via this hashtag text by responding rapidly and creating a sense of community around the movement [3]. This story told us how efficient that using hashtag for advertise campaign is.

However, because hashtags are created freely from worldwide, without any specific standard and role, it is not simple to understand the content of each hashtag. As there is no any hashtag dictionary, normally Twitter's users may have to read hundreds of tweets in order to understand a new hashtag in Twitter trend. A disadvantage that reduces the chance of effectiveness in online social network marketing advertisement is an enormous data flow in online social media. Because many trends in online social media came and gone rapidly, time becomes another remarkable factor in the developing process. In addition, if we spend too much time on research and development campaign phrase, we may lose a chance to catch the relate hashtag in time. How can we respond rapidly to catch up online social network interesting trend?

Today, especially in online marketing, new approaches are needed to advertise campaign, and propaganda. Our research focuses on analyzing hashtags gathered from tweets. We aim to understand hashtags' dynamic meaning and find similar hashtags in terms of meaning, which will be significant for providing an advertisement to target customer group before they leave the topic trend on Twitter. We also concern that hashtags have their own life cycle in this paper. Based on our previous study [4], we implement the comparing procedure in addition to our last process procedures. Hashtag has a deep association with real-time event, which is the main concept of Twitter. Marketing development usually takes a time to research an associate hashtag and then launch a marketing campaign. However, many campaigns were failing to catch

*1 Graduate School of Information and Telecommunication, Course of Information and Telecommunication Engineering, Master's Program
*2 School of Information and Telecommunication Engineering, Department of Communication and Network Engineering, Professor

the relate hashtag before it fades out from Twitter's hashtag trend list. In fact, as Twitter users may use different hashtag to reach into the same topic, the related hashtag for one topic can be more than one hashtag. Problem is how can we know the similar of each hashtag's content. After finding out this problem, the objective in our study aim to know more about the variation of Twitter hashtag in a minimum time, so we can manage and develop proper campaign or advertisement on Twitter.

The rest of the paper is organized as follows. Section 2 briefly reviews current works related to Twitter influence and clustering Twitter hashtags. Section 3 shows the data process procedures and explains each procedure. Section 4 shows a summary of gathered tweets, words distribution of hashtags and clustering results of hashtags with dendrogram. Finally, section 5 concluded our work and gives a few directions for future work.

2. Related Works

Yan Mei, Weiliang Zhao and Jian Yang mentioned how to select seed user or influential users to help the propagation of the advertising information [5]. There were two ways for social media advertising as (1) advertisers can take the advantage of various users' information, including their interests, demographics (such as gender, age, race, level of education, etc.), and their behaviors in the social networks, to deliver the advertisements directly to the target audiences; (2) advertisers can identify some influencers in the social networks as the seeds and propagate the advertisements through these seed users' social circles. In this study they told us about influence maximization problem which was: (1) the influence probability model, which determines how influence probabilities between users are calculated; (2) the influence diffusion model, which reflects how influence propagates in the networks; (3) the seed nodes selection algorithm, which is used to select the influential users from the social network in order to maximize the expected influence spread. When a company wants to launch a new advertising campaign on Twitter, it is needed to select some influential users (i.e. seed users) to help the propagation of the advertising information. In this paper, a new influence probability model and the diffusion model have been proposed. Comparing with existing works, these models can better reflect the real situations of advertising information spread on Twitter. The cumulative probabilities are calculated according to users' action history.

Lotfi A. Zadeh, Ali M. Abbasov and Shahnaz N. Shahbazova interested in an analysis of hashtag from the point of view of their dynamics [6]. They identified groups of hashtags that exhibit similar temporal patterns, look at their linguistic descriptions, and recognize hashtags that are the most representative of these groups, as well as hashtags that do not fit the groups very well. They used a fuzzy clustering algorithm to gain insight related to temporal trends and patterns of hashtags popularity concluded that the changes of hashtags popularities are an attempt to look at the dynamic nature of the user-generated data. The application of fuzzy clustering shown here provides several interesting benefits related to the fact that categorization of hashtags is not obvious.

Also, Orianna Demasi, Douglas Mason and Jeff Ma studied about understanding communities via hashtag engagement [7]. In this study, they told us that hashtag had the semantic meaning and this meaning can also be ambiguous due to temporal dynamic, e.g. the city of a sports team is often used to refer to the team during the match and not the geographic location. They separated hashtag into 4 dynamic classes by used K-means clustering algorithm. Also, they considered about bot and spammer in their experiment which are entitled that send out an abnormally high number of Tweets with title value and often not receptive audience. In result stage was shown in 4 global clusters. The cluster was validated by extensive manual inspection of a randomly selected subset of hashtags. 4 classes of clustering results were stable (daily chatter), periodically recurring, single event and irregular stochastic patterns. Tweet with hashtag received a Retweet or a favorite, human engagement. Higher engagement means that a large portion of Tweets with that hashtag received some human engagement and implies that the hashtag more successfully engages users on a broad scale.

3. Process Procedures

3.1 *Software environment*

The R environment is a convenient software facility for data manipulation, calculation and graphical display. R developer prefers to present R as an environment within which statistical techniques are implemented. R can be extended with the package [8]. Package streamR and stringR are the main packages we used in this study. Package streamR provide a function to access Twitter API via R and gather tweets sample [9]. This package was not only gathering function for our study. Because streamR also provides an import function which can import gathered data into R data frames. Package stringR provides a function to operate string format data [10]. This package plays the main role in the data cleaning procedure, extracting string data with assign format e.g. URL, stop word, encoded characters which can be processed in the program and replace this matched format with a blank data. Also, this package has been using in the hashtag extracting procedure. By assigning a hashtag format which starts with # (sharp) following with a string of word and number. In addition to these two packages, we also used some other R packages, which will be mentioned in the following subsections.

3.2 *Summary of process procedures*

Our study purpose was achieved by the following process

procedure.

1. Collect tweets sample for 24 time periods and clean these tweets
2. Extract hashtag and create top hashtag frequency list
3. Extract tweet samples for each top hashtag
4. Calculate word distribution for each top hashtag
5. Calculate Jensen-Shannon divergence between each hashtag
6. Cluster hashtags by means of Ward's hierarchical clustering method and generate dendrogram from clustered results
7. Compare the result for each time period and make conclusion.

3.3 *Collect tweet samples and clean data process*

The sampleStream is a function in streamR package. This function use for collected tweets via Twitter's API and save data in to json file. Because of the limited in hardware resource, we collected tweet data in small json files. Each json file is contained tweet during the one hour collecting process. After gathering process, we import and extract only tweets which posted in English language. Next, we implemented the data clean process simplify data. This process will prevent unexpect character encoding problem during the analyzing procedure. After creating a data frame, we also create a Rdata file as known as .Rda file to backup gathered data. Rdata file is a binary file in a compact size and can be stored any time of R structure and fast to import again by R language.

3.4 *Extract hashtags and create top frequent hashtag list*

We used the data frame obtained from the previous procedure to extract all hashtags from tweet samples. Hashtag has an assigned format. All hashtag will begin with character # (sharp) and then followed by a string of characters or numbers. According to this format, we can extract all hashtags from tweets sample. The stringR is a function use to extract string data which has a matching format. In our study we use stringR to extract all hashtag from gathered tweets. After extracting all hashtag string data, we convert these data into term document matrix format. Then we calculate the frequency of each hashtag string data by summary same hashtag in the term-document matrix. We used function TermDocumentMatrix [11] from package tm [12] to implement the calculation frequency of each hashtag. Then we sorted all hashtags by frequency and make a hashtags list data frame for each json file of time periods for preparing for the next procedure.

3.5 *Extract tweet samples for each top hashtag*

According to the hashtag list, we created each hashtag subset data frame from tweet data frame. We used function grep from package base for searching matching argument pattern within the element character vector in data source [13]. A tweet which has a matching hashtag in the list will be pasted into a new data frame. Each hashtag data frame will contain tweet with itself hashtag preparing for the next process.

3.6 *Calculate word distribution for each top hashtag*

From each hashtag's data frame, we calculated its word probability distribution for every word in each hashtag's data frame and save the result into new column in its own data frame. By counting the appearance number of each word in all tweets and calculate its probability as following Eq. (1).

$$P(A) = \frac{the\ number\ of\ words}{the\ number\ of\ all\ words\ in\ a\ document} \qquad (1)$$

3.7 *Calculate Jensen-Shannon divergence and generate dendrogram from clustered results*

This procedure will calculate divergence between each pair of hashtag's word probability and then cluster them according to the results of divergence.

For two given probability distribution P and Q, the definition of Jensen-Shannon divergence (JSD) is shown in Eq. (2).

$$D_{JS}(P \parallel Q) = \frac{1}{2}\big(D_{KL}(P \parallel R)\big) + D_{KL}\big((Q \parallel R)\big) \qquad (2)$$

where $R = \frac{1}{2}(P + Q)$

It is a measure of how much one probability distribution diverges from another probability, or in reverse words, it is a measure of how much one pair of the probability distribution is similar to each other. We calculated JSD between all pair of distribution of hashtag. However, Jensen-Shannon divergence is a symmetrized and smoothed version of the Kullback-Leibler divergence [14], defined as shown in Eq. (3). Result calculated by Kullback-Leibler show an asymmetrical result.

$$D_{KL}(P \parallel R) = \sum_i P(i) \log \frac{P(i)}{R(i)} \qquad (3)$$

First, word distribution data frame of top 15 hashtags which prepared from the previous procedure are merged together into the same data frame. The divergence will calculate from each pair of hashtag's word distribution, result from this process will be given in the 15x15 matrix with 225 results of divergence. These divergence results calculate by Jensen-Shannon divergence with CalcJSDivergence function. CalcJSDivergence function is a function used to calculate the Jensen-Shannon divergence from the row or column of the numeric matrix or for two numeric vectors [15].

This procedure of JSD and clustering hashtags according to its value and show the result with dendrogram. Dendrogram in this procedure creates by function plot from package graphics. This function can provide many types of

graph according to R object. Combining with function hclust from package statscalculate a hierarchical cluster on a set of similarities [16].

3.8 ***Compare results for each time period and conclude the results***

From the previous procedure, we can understand hashtag clustering by Jensen-Shannon divergence in hierarchical cluster dendrogram. In our study we considered in 4 points from experimental results shown in the following:

1. The volume of each time period shows us a peak and valley of tweets online traffic. These data can use to analyze when to lunch a campaign on Twitter peak usage time.
2. Hashtag top list uses to compare each period help us to understand the flow of Twitter trend via hashtag top list. In different a period shows us a different trend.
3. Jensen-Shannon divergence clustering show group of hashtags which clustered by its own word distribution. Each group clustered and show how similar of each hashtag. Hashtag is a dynamic generate content or meaning of each hashtag can be change anytime. In this point, we can determine hashtag contain or meaning in that analyze time. Considering of hashtag meaning is a significant process, we cannot use hashtag as a tool for advertisement or marketing plan if we use wrong hashtag that does not match with advertisement categories the advertisement or campaign may not reach to target customer groups.
4. The last point we considered is about the hashtag life cycle. Many hashtags always appear in gathered data. We determined these hashtags as a hashtag which has a long-life cycle. We defined hashtag life cycle as a cycle that hashtag still in part of top hashtag list.

4. Experimental Results

4.1 ***Summary of gathered tweets***

We gathered 974,967 tweets which use English language for 24 hours on 17th September 2019. Unfortunately, we found a problem in period No.15 gathering procedure, the network connection during the gathering process had been cut out after this section start 15 minutes. The number of gathered tweets in each section is shown in Table 1. From this table, we can analyze the variation of the tweets amount for each time period, except for period No.15. Also, a time chart in this table is shown in Fig.1. From this chart, we found that the lowest tweet amount started around 16:00 o'clock and the highest tweet volume started to around 19:00 o'clock. According to this result, we can choose a better time to launch a new online social media advertisement or campaigns during the peak usage time. So that we may expect to increase the percentage of advertisement to reach the target customer group.

Table 1. The number of gathered tweets in each period

Period No.	JP Time	US Time	UK Time	Tweets
1	02:56	13:56	18:56	52005
2	03:56	14:56	19:56	50171
3	04:56	15:56	20:56	49827
4	05:56	16:56	21:56	49920
5	06:56	17:56	22:56	47913
6	07:56	18:56	23:56	47420
7	08:56	19:56	24:56	45279
8	09:56	20:56	01:56	45619
9	10:56	21:56	02:56	48455
10	11:56	22:56	03:56	47630
11	12:56	23:56	04:56	44069
12	13:56	24:56	05:56	38167
13	14:56	01:56	06:56	33636
14	15:56	02:56	07:56	28296
15	16:10	03:10	08:10	26268[1]
16	17:10	04:10	09:10	25745
17	18:10	05:10	10:10	24687
18	19:10	06:10	11:10	26139
19	20:10	07:10	12:10	30437
20	21:10	08:10	13:10	36860
21	22:10	09:10	14:10	43052
22	23:10	10:10	15:10	48343
23	24:10	11:10	16:10	51220
24	01:10	12:10	17:10	52710

Generally, people read the tweet's text or check username of the tweet. But tweet's data also has many other attributes that can be very useful for data analysis. In our study, we created a data frame in our program of R language for allocating all attributes of gathered tweet's data. In this data frame, there is an attribute called "country_code" which shows the country name where the tweet is posted from. According to this country code data we identified the countries on the tweets and created the country list of each time period. The top 20 countries in period No.1, No.16, and No.23 are shown in Table 2, 3, and 4.

From this result, we noticed that time is another of the significant factor to consider target customer group. The countries in the list change during the different time period. According to human habit, day-time and night-time effect online social network user's behavior. Country list focusing

[1] This is a modified data. The original data is 6567 tweets for 15 minutes. We use 6567x4=26268 tweets for one hour as in other sections.

Table 2. Top 20 countries identified from gathered tweets in period No.1

No.	Country	Frequency	Percentage
1	United states	675	35.5
2	United Kingdom	135	7.1
3	Canada	36	1.9
4	South Africa	26	1.4
5	India	21	1.7
6	Nigeria	15	0.8
7	Malaysia	8	0.4
8	Philippines	8	0.4
9	Pakistan	7	0.4
10	Ireland	6	0.3
11	Arab Emirates	5	0.3
12	Deutschland	5	0.3
13	Ghana	5	0.3
14	Indonesia	4	0.2
15	Kuwait	4	0.2
16	Germany	4	0.2
17	Brazil	3	0.2
18	Spain	3	0.2
19	Finland	3	0.2
20	Nederland	3	0.2

Table 3. Top 20 countries identified from gathered tweets in period No.16

No.	Country	Frequency	Percentage
1	United states	137	7.2
2	United Kingdom	81	4.3
3	South Africa	25	1.3
4	Nigeria	22	1.1
5	India	18	0.9
6	Australia	17	0.9
7	Philippines	11	0.6
8	Malaysia	9	0.4
9	Arab Emirates	8	0.4
10	Ireland	7	0.3
11	Pakistan	7	0.3
12	Kenya	5	0.3
13	New Zealand	5	0.3
14	Indonesia	4	0.2
15	Belgium	3	0.2
16	Spain	3	0.2
17	Canada	3	0.2
18	Thailand	3	0.2
19	Ghana	3	0.2
20	Uganda	3	0.2

Table 4. Top 20 countries identified from gathered tweets in period No.23

No.	Country	Frequency	Percentage
1	United states	574	30.1
2	United Kingdom	82	4.3
3	Philippines	32	1.7
4	South Africa	31	1.6
5	India	24	1.3
6	Canada	19	1
7	Nigeria	15	0.8
8	Malaysia	12	0.6
9	Ghana	9	0.4
10	Indonesia	6	0.3
11	Turkey	5	0.3
12	Singapore	5	0.3
13	Kenya	5	0.3
14	Pakistan	4	0.2
15	Ireland	4	0.2
16	Germany	3	0.2
17	Saudi Arabia	3	0.2
18	Brazil	3	0.2
19	Italy	3	0.2
20	Spain	3	0.2

on Unites States in period No.1 is relatively similar to that in period No.23. The difference of period No.16 is the time. In this time period, tweets were gathered while United States was an early morning time. The highest percentage from United States is in period No.1. However, if focusing result on Philippines. The highest percentage is in period No.23. According to this result, country is also important factor for Twitter's advertisement. Therefore, the time and country are the considering factors in the advertisement or marketing campaigns.

4.2 *Results of extracted hashtags*

We gathered enormous Twitter data during 24 hours. Unfortunately, we have no efficient computer hardware resource to analyze these data. By this reason, we decided to choose gathered data from 3 different time periods. The selected sections were period No.1, No.16 and No.23.

The hashtag list contains top 15 hashtags from each section and these lists are shown in Table 5,6, and 7. The first column of the table is a hashtag in lower case character and the second column is frequent of each hashtag gathered in each time period.

By comparing these tables, we can confirm the variation of hashtag usage in each time period. In period No.1 we found hashtag #job listed in our list, but it had gone from list in period No.16 and it appeared again in period No.23. This pattern can be found for #careerarc and #hiring too. A list in period No.16 has a new hashtag which not appeared in other

section e.g. #jungkook, #blackpink, #jimin, #love and #jin. This result shows us how the rapid changes of hashtag top list in the Twitter's social network flow. Some hashtag usages from user decrease in a short period but usage volume can be increased back into hashtag top list. In another way some hashtags fade out from the top list and didn't come back again like hashtag #nct127 in period No.1.

Table 5. Top 15 hashtags in period No.1

No.	Hashtag	Frequency
1	#got7	105
2	#lullaby	84
3	#mpn	82
4	#bts	78
5	#selenagomez	66
6	#anitters	47
7	#nct127	46
8	#regular	46
9	#job	37
10	#maiareficco	36
11	#btsarmy	32
12	#hiring	32
13	#larissamanoela	30
14	#careerarc	29
15	#nct	27

Table 6. Top 15 hashtags in period No.16

No.	Hashtag	Frequency
1	#bts	67
2	#got7	52
3	#mpn	52
4	#lullaby	48
5	#btsarmy	33
6	#jungkook	30
7	#anitters	24
8	#selenagomez	24
9	#blackpink	21
10	#jimin	20
11	#love	18
12	#emmys	17
13	#exo	17
14	#maiareficco	16
15	#jin	14

Table 7. Top 15 hashtags in period No.23

No.	Hashtag	Frequency
1	#got7	114
2	#bts	93
3	#mpn	87
4	#lullaby	84
5	#selenagomez	54
6	#anitters	52
7	#larissamanoela	51
8	#job	46
9	#exo	43
10	#mark	43
11	#hiring	43
12	#careerarc	35
13	#btsarmy	35
14	#marktuan	29
15	#jackonwang	29

4.3 *Results of word distribution for hashtags*

In this study, we were focusing on hashtag #got7 to check the variation of hashtag word list in Twitter trend during our monitored time. The following Tables 8, 9 and 10 are shown the word frequency for hashtag #got7. From these tables, we found many changes of word position in the hashtag word list. Differences between word list show the change of how user used this hashtag in Twitter. We assume that "lullaby" is a new album from brand name Got7. In word list of Table 8 and Table 9, we can see the word "album". But in Table 10, the word "album" has gone from the word list. We assume that user used the same hashtag #got7 but the content in tweets has been changed.

Table 8. Top 15 words for hashtag #got7 in period No.1

No.	Word	Frequency
1	lullaby	171
2	presentyou	55
3	7official	52
4	hard	45
5	album	41
6	love	37
7	really	37
8	finally	36
9	guys	36
10	hope	36
11	itgot7	36
12	work	36
13	check	35
14	itwe	35
15	outgo	35

Table 9. Top 15 words for hashtag #got7 in period No.16

No.	Word	Frequency
1	lullaby	73
2	presentyou	40
3	7official	34
4	album	28
5	love	16
6	guys	15
7	hard	15
8	itgot7	14
9	work	14
10	check	13
11	finally	13
12	hope	13
13	itwe	13
14	outgo	13
15	really	13

Table 10. Top 15 words for hashtag #got7 in period No.23

No.	Word	Frequency
1	lullaby	107
2	180918	65
3	preview	57
4	presentyou	56
5	7official	50
6	0904	38
7	mark	37
8	yugyeom	30
9	marktuan	28
10	bambam	26
11	review180918	25
12	dance	22
13	pratice	21
14	jacksonwang	20
15	Live93	20

4.4 ***Result of JSD matrix of Twitter hashtags***

As an example, a calculated result of Jensen-Shannon divergence with a 15x15 matrix is shown in Table 11. This result was calculated from the word probability distribution of hashtags in period No.1.We can see that JSD result between same hashtags will always provide 0. From this fact, we can conclude that as small as the JSD value between a pair of different hashtags are, they will more similar to each other, or in reverse words, as big as the JSD value between a pair of different hashtags are, they will more dissimilar to each other.

4.5 ***Results of clustering hashtags for each time period***

Finally, we clustered these hashtags by means of Ward's hierarchical clustering method according to the Jensen-Shannon divergence matrix of 15 hashtags in each section. Lastly, we represented the hierarchical structure of the associated hashtag group with dendrogram. Here, as for example, we only showed 3 dendrograms, which were the clustering results of period No.1, No.16 and No.23. In Fig.2, we had two group hashtag which disappeared in Fig.3. The first group was a group of hashtags #nct, #nct127 and #regular. The second group was a group of hashtags #job, #hiring and #careerarc. The difference between these two hashtag groups was their hashtag life cycle. From Fig.4 we noticed the second group in the result again with the same grouping form. But in the first hashtag group, all the hashtags from the first group faded out from the top hashtag list in the same time. This means each hashtag life cycle associated with other hashtags in the same group. Short hashtag life cycle was represented by the first hashtag group. And long hashtag life cycle was represented by the second hashtag group. Hashtag which had a long hashtag life cycle will keep coming back in the hashtag top list event it fades out from the list sometime. In the sense, hashtag which had a long-life cycle could be defined as stable hashtag too. The content or meaning of use in these hashtag types is rarely to change.

5. Conclusions and Future Works

In this study, we used the word probability distribution for understanding the meaning contained in hashtag and then used Jensen-Shannon divergence for clustering hashtags. We represented the hierarchical clustering results with dendrogram. We implemented the comparing process procedure to study more deeply about the hashtag life cycle. With the results of our study, we can understand more about each customer behavior. By including the related hashtags into advertisement campaign, we can improve the percentage of tweets reached to the target customers. So that Twitter users can access to the advertise campaign easier. Comparing with Ref.5, our study presented a simpler algorithm for clustering hashtags on Twitter. Comparing with Ref.6, we used Jensen-Shannon Divergence as a measure to cluster hashtag, which reflects precisely the dissimilarity between two probability distributions.

For future work, we will continue our study on hashtag life cycle in a relatively long time period. For this objective, we must continually gather long-term, e.g. many days ar many months, tweets from Twitter. From gathered data we expect to discover the factor that effect change in hashtag life cycle and the patterns of hashtags.

References

[1] IBM website, *Big data analytics*. (2018) [Online]. Available: https://www.ibm.com/analytics/hadoop/big-data-analytics

[2] Statista website, *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in milions)*. (2018) [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users

[3] The Guardian website, *Five social media charity campaigns you need to know about*. (2018) [Online]. Available: https://www.theguardian.com/voluntary-sector-network/2014/apr/03/five-social-media-charity-campaigns.

[4] Nartlada Bhakdisuparit and Iwao Fujino, "Understanding and clustering hashtags according to their word distributions," in *IEEE 5th International Conference on Business and Industrial Research*, 2018.

[5] Yan Mei, Weiliang Zhao and Jian Yang, "Maximizing the Effectiveness of Advertising Campaigns on Twitter," in *IEEE 6th International Congress on Bigdata*, 2017.

[6] Lotfi A. Zadeh, Ali M. Abbasov and Shahnaz N. Shahbazova, "Analysis of Twitter Hashtags: Fuzzy Clustering Approach," in *North America Fuzzy Information Processing Society,* 2015.

[7] Orianna Demasi, Douglas Mason and Jeff Ma, "Understanding Communities via Hashtag Engagement: A clustering Based Approach" in *Proc*. *ICESM*,2016.

[8] R Project website, *What is R?*, (2018) [Online]. Available: https://www.r-project.org/about.html

[9] Pablo Barbera. *streamR: Access to Twitter Streaming API via R.* [Online]. Available: https://cran.r-project.org/web/packages/streamR/index.html. Accessed on: Sep. 24, 2018.

[10] RDocumentation website, *stringr*. (2018) [Online]. Available: https://www.rdocumentation.org/packages/stringr/versions/1.1.0

[11] RDocument website, TermDocumentMatrix, (2018) [Online]. Available: https://www.rdocumentation.org/packages/tm/versions/0.7-5/topics/TermDocumentMatrix

[12] RDocument website, *tm*, (2018) [Online]. Available: https://www.rdocumentation.org/packages/tm/versions/0.7-5

[13] RDocument website, *base*, (2018) [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.5.1/topics/grep

[14] Hiroya Takamura, *Introduction to Machine Learning for Natural Language Processing* (In Japanese), Corona Publishing Co.,LTD, Tokyo, Jpn, 2015.

[15] RDocument website, *textmineR*, (2018) [Online], Available: https://www.rdocumentation.org/packages/textmineR/versions/2.1.3/topics/CalcJSDivergence

[16] RDocument website, *hclust*, (2018) [Online], Available: https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/hclust

Table 11. Result of calculated JSD matrix of hashtags in period No.1

	#got7	#lullaby	#mpn	#bts	#selenagomez	#anitters	#nct127	#regular	#job	#maiareficco	#btsarmy	#hiring	#larissamanoela	#careerarc	#nct
#got7	0.000	0.136	0.535	0.475	0.556	0.560	0.553	0.563	0.504	0.560	0.538	0.516	0.562	0.526	0.542
#lullaby	0.136	0.000	0.546	0.475	0.560	0.564	0.557	0.566	0.506	0.564	0.540	0.518	0.567	0.528	0.548
#mpn	0.535	0.546	0.000	0.445	0.201	0.246	0.575	0.580	0.536	0.242	0.429	0.547	0.338	0.552	0.565
#bts	0.475	0.475	0.445	0.000	0.521	0.528	0.529	0.539	0.493	0.528	0.284	0.509	0.529	0.514	0.515
#selenagomez	0.556	0.560	0.201	0.521	0.000	0.154	0.584	0.590	0.557	0.152	0.531	0.568	0.269	0.573	0.577
#anitters	0.560	0.564	0.246	0.528	0.154	0.000	0.590	0.594	0.561	0.211	0.539	0.571	0.224	0.577	0.582
#nct127	0.553	0.557	0.575	0.529	0.584	0.590	0.000	0.093	0.561	0.590	0.578	0.571	0.593	0.577	0.132
#regular	0.563	0.566	0.580	0.539	0.590	0.594	0.093	0.000	0.564	0.593	0.583	0.575	0.596	0.580	0.155
#job	0.504	0.506	0.536	0.493	0.557	0.561	0.561	0.564	0.000	0.560	0.529	0.141	0.563	0.152	0.553
#maiareficco	0.560	0.564	0.242	0.528	0.152	0.211	0.590	0.593	0.560	0.000	0.541	0.571	0.315	0.576	0.582
#btsarmy	0.538	0.540	0.429	0.284	0.531	0.539	0.578	0.583	0.529	0.541	0.000	0.541	0.538	0.545	0.564
#hiring	0.516	0.518	0.547	0.509	0.568	0.571	0.571	0.575	0.141	0.571	0.541	0.000	0.574	0.099	0.564
#larissamanoela	0.562	0.567	0.338	0.529	0.269	0.224	0.593	0.596	0.563	0.315	0.538	0.574	0.000	0.579	0.585
#careerarc	0.526	0.528	0.552	0.514	0.573	0.577	0.577	0.580	0.152	0.576	0.545	0.099	0.579	0.000	0.569
#nct	0.542	0.548	0.565	0.515	0.577	0.582	0.132	0.155	0.553	0.582	0.564	0.564	0.585	0.569	0.000

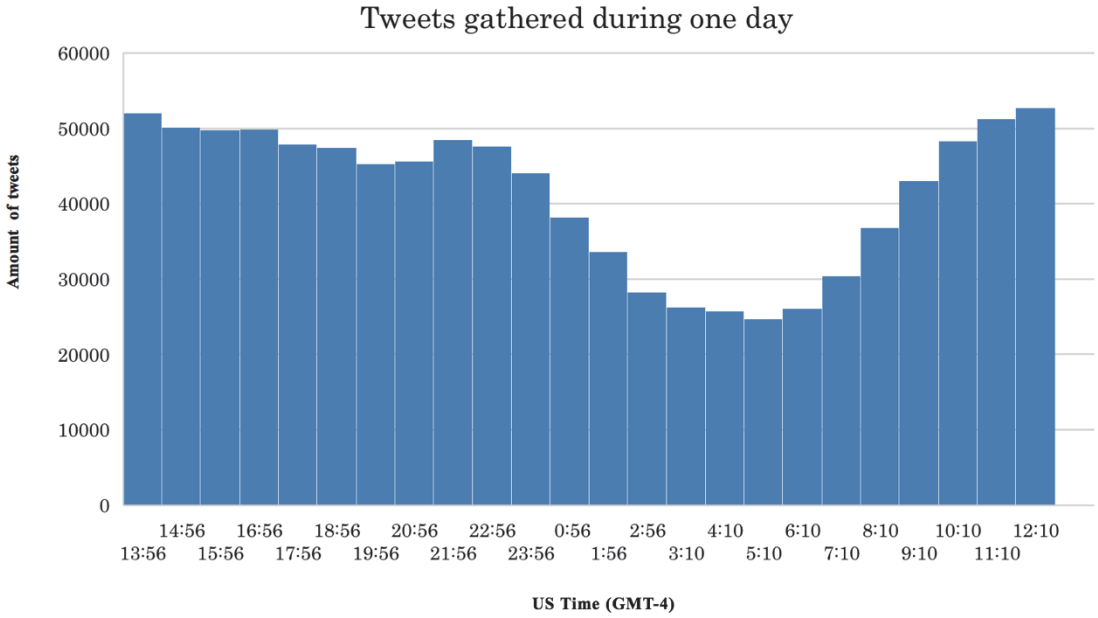

Fig.1 Time chart of Tweets gathered during one day

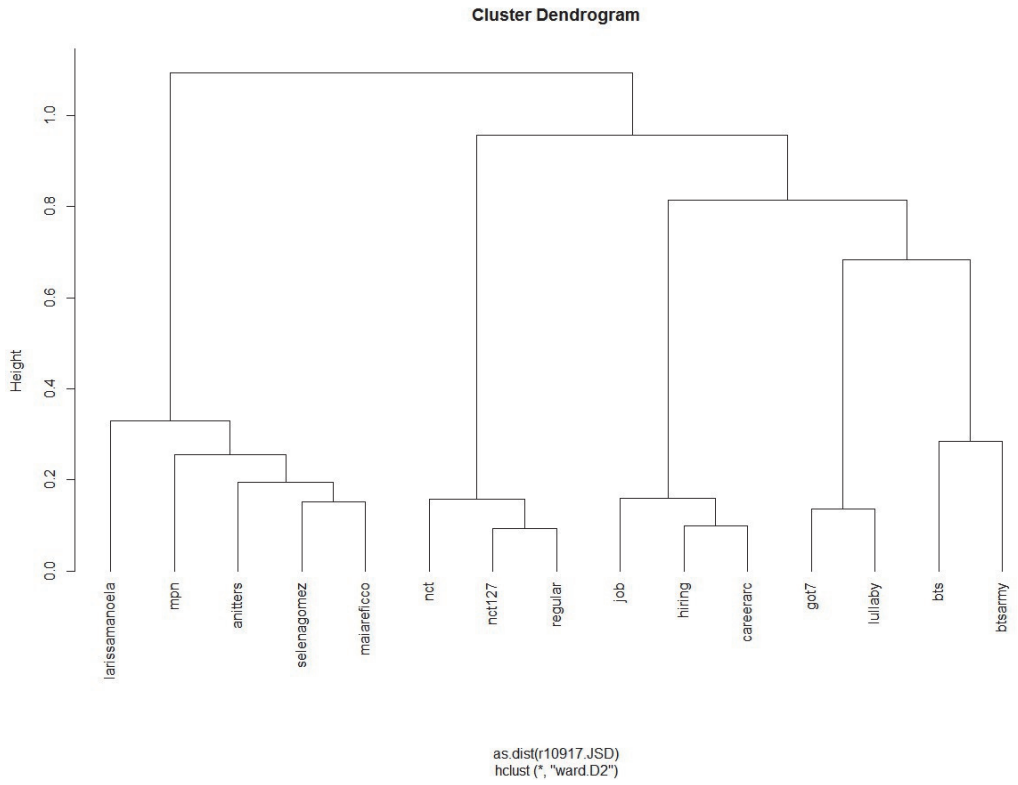

Fig.2 Cluster dendrogram of top 15 hashtags in period No.1

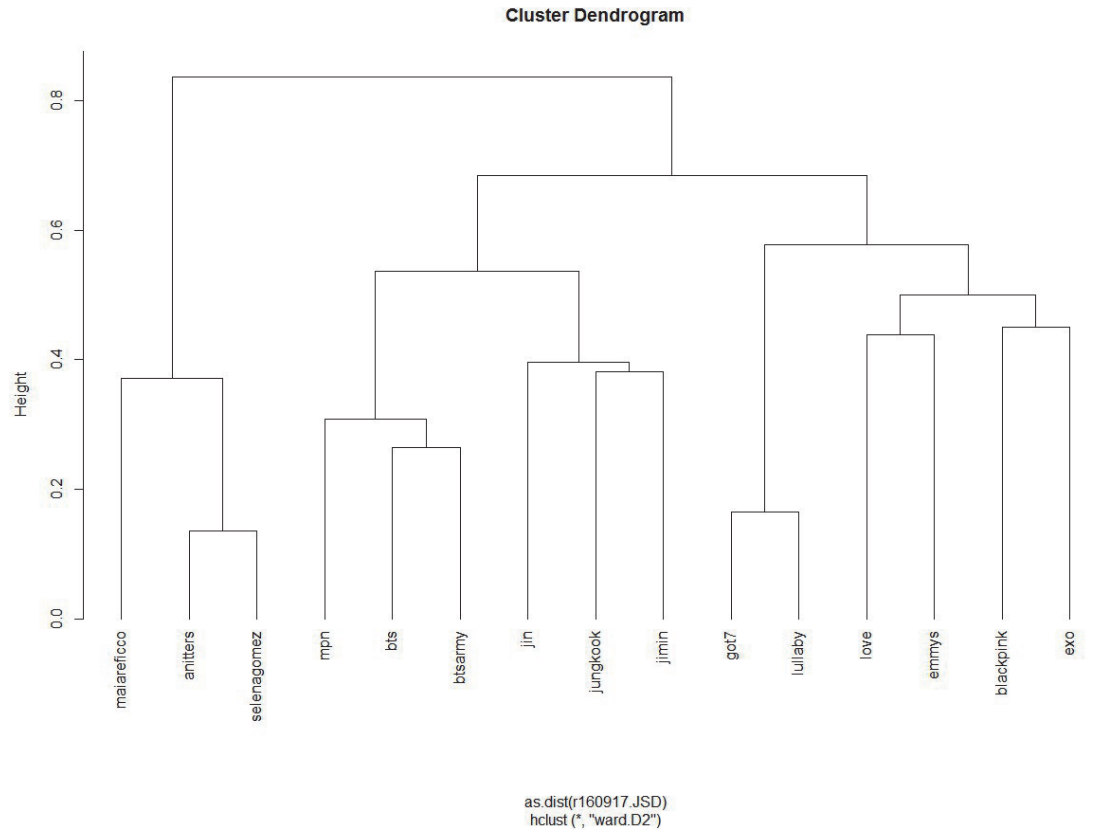

Fig.3 Cluster dendrogram of top 15 hashtags in period No.16

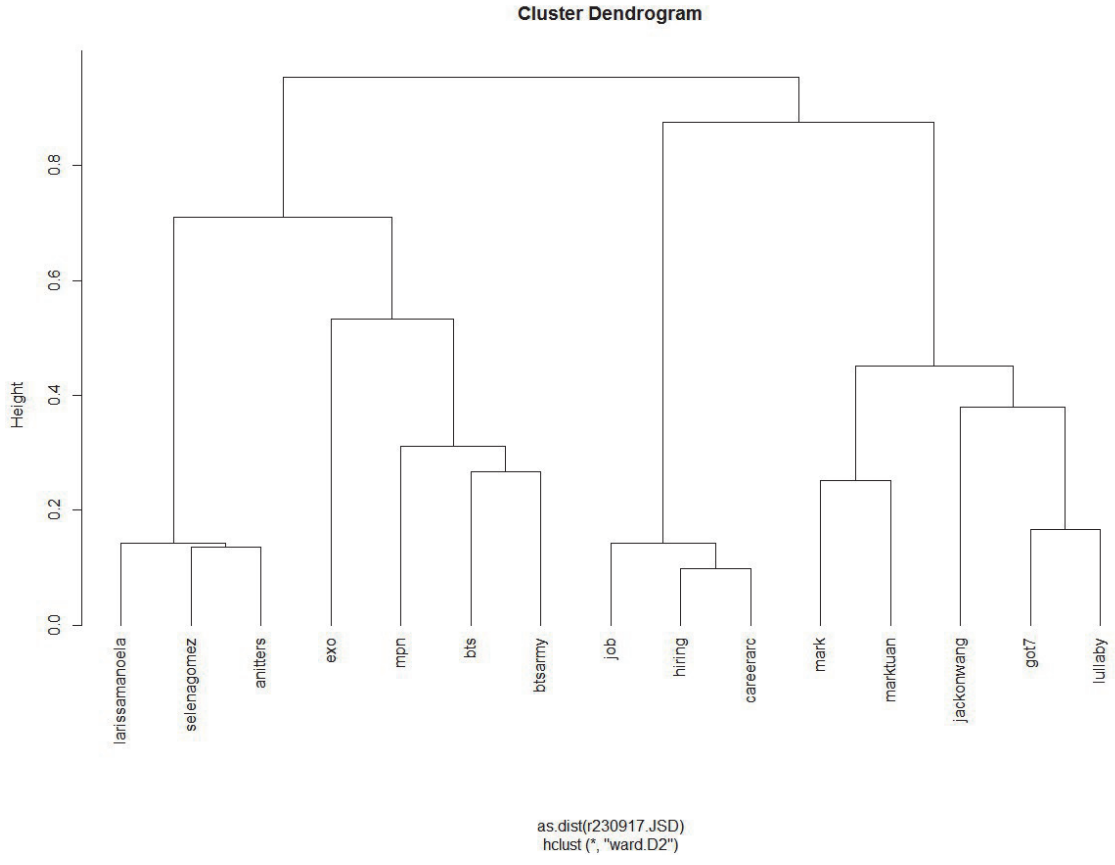

Fig.4 Cluster dendrogram of top 15 hashtags in period No.23