Time Series Topic Transition based on Micro-Clustering

1st Takako Hashimoto
Commerce and Economics
Chiba University of Commerce
Chiba, Japan
takako@cuc.ac.jp

2nd Takeaki Uno
Principles of Informatics Research Division
National Institute of Information
Tokyo, Japan
uno@nii.jp

3rd Tetsuji Kuboyama
Computer Centre
Gakushuin University
Tokyo, Japan
kuboyama@tk.cc.gakushuin.ac.jp

4th Kilho Shin
Graduate School of Applied Informatics
University of Hyogo
Hyogo, Japan
kilhoshin314@gmail.com

5th Dave Shepard
HumTech
University of California, Los Angels
California, USA
shepard.david@gmail.com

Abstract—This paper proposes a method for analyzing time series topic transition based on micro-clusters to present different situations that show people's reactions to topical problems on the Web. To form micro-clusters, we leverage our original data polishing algorithm developed by one of the authors. Our method shows that micro-clusters efficiently represent the dynamics of topic transitions: for example, events cause sudden changes in the number of clusters. This implies that there were increases or decrease of diversity of cluster contents that correspond to people's feelings and opinions to the topic. To show the method's effectiveness, we conducted an experiment on tweets targeting rumors of a petrochemical complex explosion just after the Great East Japan Earthquake in 2011. Our method easily identifies the following phases in topic transitions. First, people post the real story. Second, rumors circulate about the explosion. Finally, the rumors were corrected by the government and gradually disappeared.

Index Terms—micro-clustering, topic extraction, time-series, data polishing, social media analysis, big data analysis

I. Introduction

After the East Japan Great Earthquake on 11 March, 2011, rumors about an explosion at a petrochemical complex owned by Cosmo Oil spread rapidly on twitter. Stories of oil tanks exploding and releasing harmful substances into the air caused widespread panic until official government news releases corrected the misinformation the following day. This story demonstrates the importance of fake news detection: while an enormous real disaster (the earthquake) was unfolding, rumors of imaginary disasters spread misinformation and diffused attention on social media to imaginary dangers. After the East Japan Great Earthquake on 11 March, 2011, a wide variety of fake news rapidly spread on Twitter. One example was a rumor about a petrochemical complex explosion at the Cosmo Oil. The rumor spread stating that oil tanks were exploding and emitting harmful substances into the air.

Automatic topic extraction has been a major area of research for the last ten years. Conventional methods based on word co-occurrence and latent topic extraction from high dimensional vector space such as LDA [1] could extract major topics. However, these conventional methods fail to capture the emergence of topics over time, and how individual users react to topics. Approaches focusing on tweets with specific keywords and monitoring the number of keywords and co-occurrence words could find the trends of topics such as burst and disappearance, but it is also difficult to understand how people reacted to the topic.

This paper proposes a method for analyzing topic transition dynamics based on micro-clustering, an approach that creates smaller clusters, at least compared to conventional clustering methods. Each cluster is composed of entities similar to each other, as in community detection. In our method, micro-clusters are extracted by a data polishing algorithm [2] [3] from millions of tweets. Each micro-cluster shows independent opinions and/or aspects of the topic. Next, we analyze clusters over time using visualization methods to understand the topic transition. Further, we observe that the diversity of clusters such as the number of clusters and the number of words in one cluster should show topic transition dynamics. When people's opinions on a topic suddenly change, this leads to a drastic increase of diversity of opinion on the topic, even if the number tweets does not increase. On the other hand, when people may not consider a topic in depth, such as just repeating a rumor, we can observe no change in the number of clusters even though the number of tweets increases.

The contributions of this paper are as follows:

- Proposing a novel method for showing topic transition dynamics by quantifying cluster diversity using micro-clustering.
- Visualizing the dynamics of topic transitions by applying the method to a large amount of tweets.

This paper is organized as follows. Section II introduces related work on social big data analysis. Section III describes our proposed method. Section IV demonstrates experimental results of our method: it shows that our method can extract quality clusters by micro-clustering, and that these micro-clusters illustrate topic transition dynamics from cluster diversity. Finally, Section V offers directions for future research.

II. Related Work

Research on time series topic analysis targeting social media (e.g. Twitter) has become an active area of research. One type of approach uses word co-occurrence to track topic transition over time [4] [5]. These methods are useful for extracting dominant topics, but they do not have enough resolution to distinguish subtle but important differences between a real and a fake subtopic in a topic, or differences among incompatible opinions in a topic. It is also difficult to extract a small amount of representative keywords showing the topic content.

A number of methods have been proposed to track topic transitions over time based on conventional topic extraction methods such as LDA and LSA. In these methods, topics are extracted on each window in a time series, and connected according to their similarities to track their transitions [6]–[8]. These methods characterize clusters with a set of keywords based on word occurrences, and high-frequency words tend to be extracted as keywords.

We can extract keywords using conventional methods, but it is not easy to make sense of them. These conventional methods extract a few big clusters and many small clusters. Although these methods are good for major topic extraction, they tend to put all diverse subtopics into one big topic. For example, we can realize that there are topics about the rumors, but it is not easy to detect each topic's meaning/purpose such as questioning, repeating, conflicting, and so on. Distinguishing real from fake, conflicting opinions, and change of opinion are difficult with conventional methods.

Sometimes these methods also identify false similarities between clusters over time. The accuracy of existing clustering techniques is not high, nor can these techniques deliver reasonable performance.

To alleviate these problems, we propose an efficient method for detecting time series topic transition by micro-clustering.

III. Proposed Method

This section presents our technique. Figure 1 shows our method, and Figure 2 visualizes the process of our method for graph generation and micro-clustering by a data polishing algorithm. In Figure 1, at first the nodes are created as Input Data, and similar nodes are connected by edges to form a graph. Then cliques are extracted as clusters by micro-clustering using data polishing. In our method, we define a topic as a set of one or more clusters. Topic transition is analyzed by considering the cluster diversity that constitute the corresponding topic.

A. Input Data

We group the tweets sequentially by a certain window (e.g. half an hour) when they were sent. We then create the sequence of $tweet_{id}$-$word_{id}$ count matrices, $\langle TW_0, TW_1 \ldots, TW_t, \ldots, TW_T \rangle$ that contains the words used in each tweet during each time period. To segment tweets that may not have used spaces to delineate word boundaries, we employed the Japanese morphological analyzer, MeCab [9].

$$TW_t = \begin{pmatrix} tw_{11} & tw_{12} & tw_{12} & \cdots & tw_{1n} \\ tw_{21} & tw_{22} & tw_{22} & \cdots & tw_{2n} \\ tw_{31} & tw_{32} & tw_{32} & \cdots & tw_{3n} \\ \vdots & \vdots & & \ddots & \vdots \\ tw_{m1} & tw_{m2} & tw_{m2} & \cdots & tw_{mn} \end{pmatrix} = (tw_{ij})_t$$

where $1 \leq i \leq m$ and $1 \leq j \leq n$. The index m is the number of tweets and n is the number of words during a time period. The element tw_{ij} shows 0 or 1 that i-th tweet did not use or used a particular word w_j during a time period. These time series matrices, $TW_0, \ldots, TW_T$, are obviously sparse.

B. Graph Generation

Next, a graph of tweets is formed during each time period. In the graph, each tweet tw_i in TW_t should be a node. Next, similar tweets that have similar words are connected by edges. To evaluate tweet similarity, we use the Jaccard coefficient [10], a popular measure for comparing the similarity and diversity of sample sets.

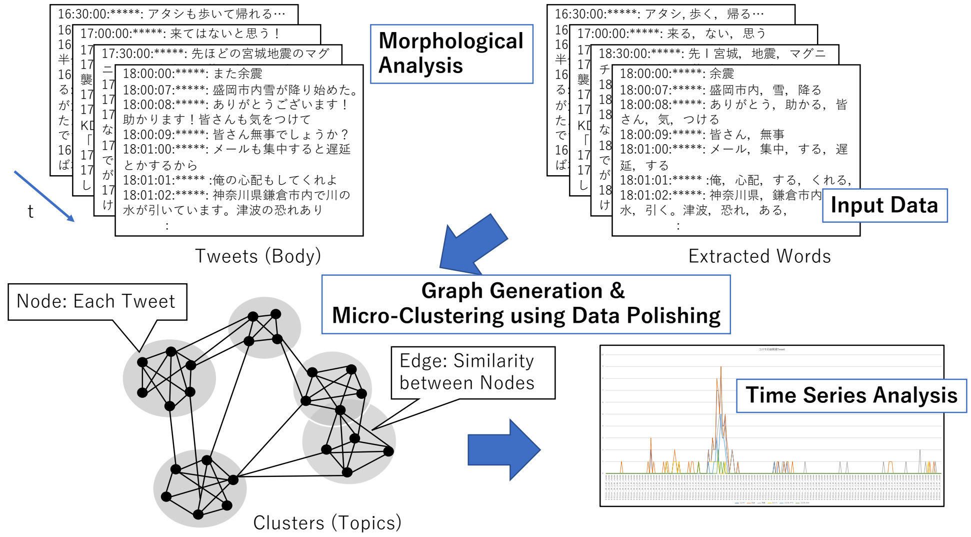
$$J(tw_i, tw_j) = \frac{|tw_i \cap tw_j|}{|tw_i \cup tw_j|}$$

We set the threshold s. If the Jaccard coefficient between nodes is larger than s, these nodes are connected by an edge.

C. Micro-Clustering using Data Polishing

Our method uses a data polishing algorithm for micro-clustering developed by one of the authors [2] [3]. This section describes the data polishing algorithm briefly; for more information, please see the cited paper. Micro-clusters are groups of data records that are similar or related, and have one meaning, or correspond to one group. Micro-clusters should satisfy the following conditions.

1) quantity (the number of micro-clusters found should not be huge)
2) independence (micro-clusters should not be similar)
3) coverage (all micro-clusters should be found)
4) granularity (the granularity of micro-clusters should be the same)


16:30:00:*****: アタシも歩いて帰れる…
17:00:00:*****: 来てはないと思う！
17:30:00:*****: 先ほどの宮城地震のマグ
18:00:00:*****: また余震
18:00:07:*****: 盛岡市内雪が降り始めた。
18:00:08:*****: ありがとうございます！助かります！皆さんも気をつけて
18:00:09:*****: 皆さん無事でしょうか？
18:01:00:*****: メールも集中すると遅延とかするから
18:01:01:***** :俺の心配もしてくれよ
18:01:02:*****: 神奈川県鎌倉市内で川の水が引いています。津波の恐れあり
:

16:30:00:*****: アタシ, 歩く，帰る…
17:00:00:*****: 来る，ない，思う
18:30:00:*****: 先ｌ宮城，地震，マグニ
18:00:00:*****: 余震
18:00:07:*****: 盛岡市内，雪，降る
18:00:08:*****: ありがとう，助かる，皆さん，気，つける
18:00:09:*****: 皆さん，無事
18:01:00:*****: メール，集中，する，遅延，する
18:01:01:***** :俺，心配，する，くれる，
18:01:02:*****: 神奈川県，鎌倉市内
水，引く。津波，恐れ，ある，
:

Fig. 1. Proposed Method

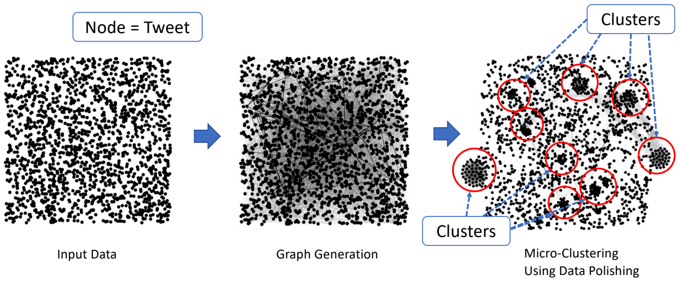


Fig. 2. Graph Generation and Micro-Clustering by Data Polishing

5) rigidity (the micro-clusters found should not change due to non-essential changes such as random seeds or indices of records)

In a graph, micro-clusters are considered to correspond to dense subgraphs, and the non-edges in the dense subgraphs are ambiguities. We also consider that edges included in no clusters are also ambiguities. The concept of data polishing for micro-clustering comes from this: add edges for these non-edges, and remove these edges from the graph. For identifying these non-edges and edges, we consider the following feasible hypothesis.

If nodes u and v are in the same clique of size k, u and v have at least $k-2$ common neighbors. Thus, we have $|N[u] \cap N[v]| \geq k$, and this is a necessary condition that u and v are in a clique of size at least k. We call this condition k-common neighbor condition. If u and v are in a sufficiently large pseudo clique, they are also expected to satisfy this condition. In contrast, if two nodes do not satisfy the condition, they belong to a pseudo clique with very small probability. Even though they belong to a pseudo clique, they actually seem to be disconnected in the clique, thus we may consider that they should not be in the same cluster. Let $P^k(G) = (V, E')$ where E' is the correction of edges connecting node pairs satisfying the k-common neighbor condition, and the polishing process is the computation of $P^k(G)$ from G. We call this process k-intersection polishing. To evaluate k-common neighbor condition, we also use the Jaccard coefficient. We set the threshold s' and if the Jaccard coefficient between nodes u and v considering their neighbors is larger than s', the edge is generated between them. The maximal clique enumeration is done by algorithms such as MACE [11].

D. Time Series Analysis

To analyze micro-clusters over time, we check the following for each time period:

- the number of tweets
- the number of micro-clusters
- the number of tweets in each micro-cluster
- frequent words in each micro-cluster
- tweet bodies in each micro-cluster

We focus on the specific topics such as the fake news, and confirm micro-clusters' quality is high and they can show the topic transition dynamics.

IV. Experimental Result

This section describes the results of an experiment we conducted with our method. First, we explain the input data and the target topic for the experiment. We compare our method with LDA to show that our method can present the diversity of clusters. Then we examine the characteristics of micro-cluster contents extracted by our method and topic transaction dynamics.

A. Input data

Our target data set is the over 200 million tweets sent around the time of the Great East Japan Earthquake that happened at 14:47 on March 11, 2011. We obtained this dataset from the social media monitoring company Hotto

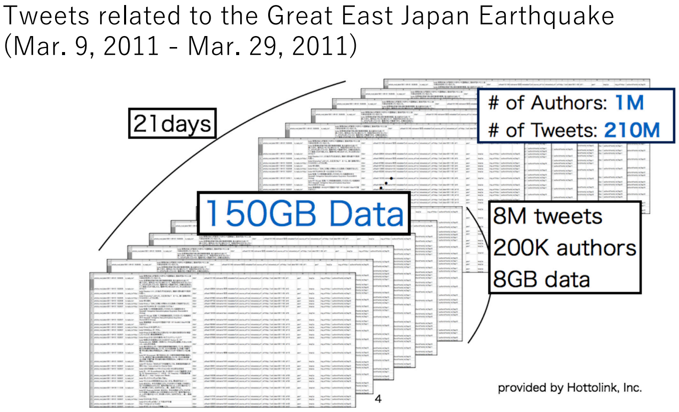

Tweets related to the Great East Japan Earthquake
(Mar. 9, 2011 - Mar. 29, 2011)
21days
# of Authors: 1M
# of Tweets: 210M
150GB Data
8M tweets
200K authors
8GB data
4
provided by Hottolink, Inc.

Fig. 3. Target Data: 200 million tweets related to the Great East Japan Earthquake

link Inc. [12], who tracked users who used one of 43 hashtags (for example, #jishin, #nhk, and #prayforjapan) or one of 21 keywords related to the disaster. Later, they captured all tweets sent by all of these users between March 9th (2 days prior to the earthquake) and March 29th (Figure 3). This dataset offers a significant document of users' responses to a crisis, but its size presents a challenge.

In subsequent subsections, we show our experimental result for tweets from 00:00 on March 11 to 24:00 on March 15, a total of 120 hours. We crated a sequence of $tweet_{id}$-$word_{id}$ count matrices for our dataset, one each 30 minutes, for a total of 240 slots. For example, each the 30-minute matrix on March 11 before 14:30 (before the earthquake), contains 60,000-80,000 tweets. On the other hand, each 30-minute matrix for 30 minutes on March 11 after 15:00 (after the earthquake) contains 300,000-500,000 tweets. The number of tweets increased dramatically after the earthquake. The size of each matrix after 15:00, March 11 is around 15MB and they were all quite sparse.

B. Target topic

As the target topic in this experiment, we selected the fake news about the petrochemical complex explosion. The fake news progressed in the following four stages:

- Stage-1: Fact: around 15:00 on March 11 (just after the quake), the Petrochemical Complex in Chiba was on fire
- Stage-2: Rumor: Around 19:00 on March 11, the following rumor was diffused.
  - Radiation and harmful chemicals are leaking into the air from the petrochemical complex. Be careful!
  - Don't go out! The rain contains radiation and harmful materials from the petrochemical complex explosion.
- Stage-3: Correction: Around 15:00 on March 12 (the day after the earthquake), the industry's Website and the local government's twitter officially corrected the rumor.
- Stage-4: Disappearance: At night on March 12, the topic disappeared.

The rumor about the oil tank emitting harmful substances into the air was diffused, and frightened people. Finally the rumor was officially corrected by the government and disappeared. To evaluate the progress of the target topic, we investigated micro-clusters that contain the word ”cosmo oil,” the name of company that owned the petrochemical complex, over time.

In this experiment, we examined the target topic transition and the diversity of clusters in each time period to show our method's effectiveness.

C. Graph Generation

In this experiment, we treated retweets (RT) as normal tweets. In the graph, retweets were also recognized as nodes. We set the Jaccard coefficient threshold $s = 0.3$. If we set the threshold $s = 0.5$, semantically similar tweets that use slightly different words are sometimes missed. If we set the threshold $s < 0.3$, we find many tweet pairs that incidentally use a couple of same words. We then form the graphs of tweets in all time periods.

D. Micro-Clustering using Data Polishing

We set the Jaccard coefficient threshold $s' = 0.2$. We experientially know that the threshold $0.1 <= s <= 0.4$ does not make much difference for the result. And if we set the threshold $s <= 0.1$, we gain too big clusters and if we set the threshold $s >= 0.5$, we may gain many small clusters that should be just one cluster. We then connect edges between nodes that have same friend nodes. Then the maximal clique enumeration should be done.

E. Comparison with LDA

As section III-C mentioned, our method is based on the maximal clique enumeration. On the other hand, the major topic model LDA is based on probabilistic model. The approaches of our method and LDA are different, therefore, it is not easy to simply compare. But to show that our method is able to present the diversity of clusters, we tried to compare our method with LDA.

We utilized the lda using collapsed Gibbs sampling [13]. As mentioned previously, the Great East Japan Earthquake caused twitter use to soar to 8 million tweets per day in Japan. Our dataset alone contains over 200 million tweets sent by nearly 1 million authors. This volume of data is typical for events of similar significance. However, conventional data mining methods like LDA do not scale well to this challenge. Therefore, we randomly selected 1800 tweets (hereinafter, written as ”sample tweets”) from tweets posted during 15:00-15:30 on March 12 (24 hours after the quake).

First, we applied our method to the sample tweets, which produced 1273 micro-clusters. For LDA, we set the

TABLE I
Some tweets from the 1st topic derived by LDA (excerpt)

ID	Tweet body
132	Also the karaoke pavilion.
303	If it is true the karaoke hall is disappointing but there is confirmation that it is a fact and everyone is doing
529	About 25 minutes after the passage bridge I plan to pass near Hachi
781	I do not bombed it but I always emit exhaust gas during power generation
806	I am going to take a bath now it feels sick because of my heat ...
884	Well it took an hour to shop I bought some water and a cupan so I replaced the content of disaster prevention bags that had been expired.
950	It is strange.
986	Kawashima's thank you [diffusion] people want rushed to volunteer. After three days not a turn.

TABLE II
Some tweets from the 1st topic derived by our method (excerpt)

ID	Tweet body
523	Kanto If in a million households, 1 kW cooking can be done by 4 o'clock in the evening, we can reduce the peak electricity power supply crisis by 1 million kW. There is a life that can only be saved by that much. Everyone Kanto housewife please help us save lives Do not stop the hospital's electricity ...
260	please! Kanto 1 million kilowatt cooking in 1 million households by 4 o'clock in the evening we can reduce the peak electricity which electricity supply crisis will be 1 million kW the life which can be saved just by that alone We ask everyone to help save lives.It stop electricity at hospital ...
449	please! Kanto 1 million kilowatt cooking in 1 million households by 4 o'clock in the evening we can reduce the peak electricity which electricity supply crisis will be 1 million kW the life which can be saved just by that alone We ask everyone to help save lives.It stop electricity at hospital ...
1558	[Operational Guidelines for Operation of Yashima]? Kanto 1 million won cooking at 1 million households by 4 o'clock in the evening By merely finishing it by 4 o'clock we can reduce the peak power at which electricity supply crisis will be 1 million kW? Do not refrain because there is a possibility of secondary disasters such as fire? I understand everything with Yashima Strategy ...

number of topics as 1273 (the same number of topics derived by our method), the number of iterations as 3000, and the hyper parameters α and β as 0.1 and 0.01 respectively. As a result, LDA produced 600 clusters.

Figure 4 and Figure 5 show the number of tweets of each cluster derived by LDA and our method respectively. LDA made one big cluster (the number of tweets = 157) and a lot of small clusters (≤ 26). We can see the cluster size bias in LDA's result. However, unlike LDA, our method produced relatively similar sized clusters.

Next, we analyzed the contents of clusters. Table I shows excerpted tweets of the 1st (biggest) topic derived by LDA. Since there are too different tweets in Table I, it is not easy to understand the meaning of the cluster. Table II shows excerpted tweets of the 1st (biggest) topic derived by our method. All the tweets are basically same. Unlike LDA, we can easily understand the meaning of the cluster. This shows our method can extract quality clusters.

As for the target topic (the fake news about the petrochemical complex explosion), there were the following 16 tweets in the sample set (Table III). Table III also shows the clusters produced by LDA and our method. In LDA, we can see three different clusters ($\# = 58, 59, 441$) for 16 tweets, on the other hand, in our method, there are clusters ($\# = 440, 445, 558, 760, 793, 879, 989, 1046, 1047, 1170, 1206$). LDA cluster #441 has 14 tweets and LDA clusters #58 and #59 have one tweet respectively. However, our method's cluster #558 has 4 tweets, our method's cluster #1170 has 3 tweets, and the other our method's clusters ($\# = 440, 445, 760, 793, 879, 989, 1046, 1047, 1206$) have one tweet respectively. Our method's cluster #558 consists of four tweets baed on "Be careful not to go out and do not expose your skin! Explosion of Cosmo Oil causes harmful substances", but has two aspects such as a rumor and correction of a rumor. It can be said that our method cluster #558 shows the transition from a rumor to correction on tweets that include "Be careful not to go out and do not expose your skin! Explosion of Cosmo Oil causes harmful substances". Our method's cluster #1170 consists of 3 tweets based on "It is expected that rain will fall in the Chiba and the metropolitan area due to the fire at the ironworks in Chiba, including umbrellas and raincoats". This cluster shows the rumor diffusion. The other clusters show different aspects such as a family matter related to the oil tank explosion ($id = 440$), the official URL from the industry ($id = 445$), people's different opinions ($\# = 760, 793, 879, 1047, 1146, 1206$), and correction of the fake story ($\# = 989$). Our method can show different aspects of the target topic unlike LDA.

LDA is good for major topic extraction and it tends to pull all diverse subtopics into one topic. Through the experiment, we can confirm that our method can extract different opinions on a topic. Both LDA and our method have different characteristics respectively, so we can use them jointly according to the purpose for the analysis.

F. Micro-Cluster Contents Characteristics

Here, we examine the characteristics of micro-cluster contents. Figure 6 shows the number of tweets related to the Great East Japan Earthquake for 240 time periods (from 00:00 March 11 to 24:00 March 15). Obviously, just after the earthquake, the number of tweets dramatically increased. From Figure 6, we could easily see that something happened, but not exactly what. The number of micro-clusters overlapped on the graph (Figure 6) and we can find that the number of micro-clusters is significantly correlated to the number of tweets.

Since the number of tweets is around 200K-500K for each period, it is not easy to understand the each period's contents (opinions and reactions). However, the number of micro-clusters is around 10K - 50K and has important

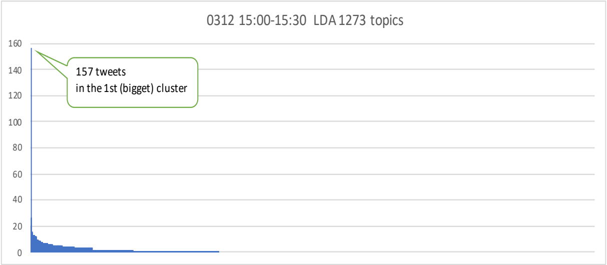


Fig. 4. # of tweets of each cluster (LDA)

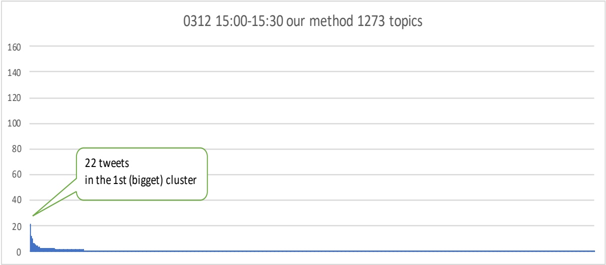


Fig. 5. # of tweets of each cluster (Our Method)

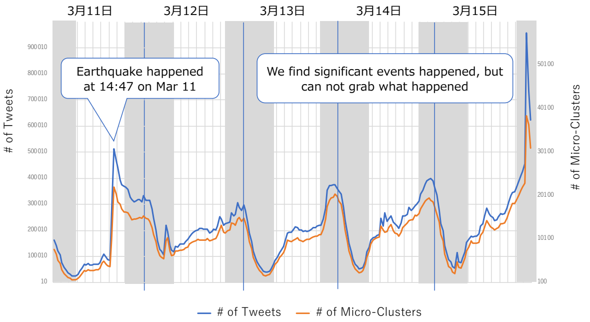


Fig. 6. # of Tweets and # of Micro-Clusters

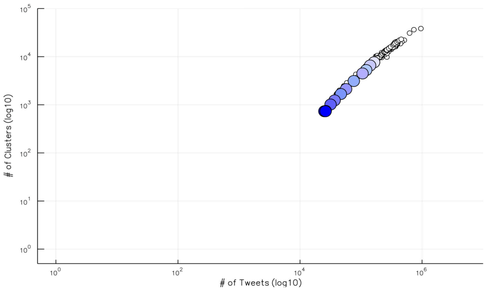


Fig. 7. # of Tweets vs # of Micro-Clusters (base 10 log-log plot)

words in each cluster, so that it is relatively easier to understand contents in each slot.

Table IV shows micro-clusters about the target fake news. In Table IV, cluster #1 is the fact that it there was a ”petrochemical complex fire” and #2 shows a rumor occurred afterwards. In #2, the phrase such as ”harmful substance”, ”rain”, ”rain coat” can be observed. Also, we can see that the phrase ”spreading hope” was retweeting the rumor information. #3 is also the fake information in the same way. On the other hand, #4 is a cluster showing attempts to correct the rumor. Tweets containing the word ”fake” were retweeted. It is noteworthy that #3 and #4 belong to the same time period, and it turns out that topics of different aspects can be extracted in the same time period. #5 is also a rumor correction topic, but it can be seen that it is not a retweet like #4 but a kind of free discussion by posters. If we utilize a conventional topic extraction technique such as LDA, we would only discover a few huge clusters and many small clusters. We would not get clusters that show different viewpoints.

G. Topic Transaction Dynamics

Figure 7 shows the relationship between the number of tweets and the number of micro-clusters. Each circle show one time period. The nearly linear relationship observed on the log-log plot implies a power law relationship between the number of tweets and the number of micro-clusters. Here, we make the hypothesis that the line is the upper bound of the topic diversity; that is, when each topic emerges independently, the total number of topics is equal to this upper bound. On the other hand, when some intensive diffusion occurs in a topic, the number of topics moves away from the upper bound.

Figure 8 shows the relation between the number of tweets that include the word ”cosmo oil” and the number of micro-cluster that also include the word ”cosmo oil”. Each circle show each time period as in Figure 7. However, unlike Figure 7, it does not seem to show a correlation between the number of micro-clusters and the number of tweets. In some time periods of Figure 8, the number of micro-clusters is lower than the number that is expected. That means that for the topic about the petrochemical complex fake news, there are some slots that show low diversity of topics.

By comparison, Figure 9 shows the relation of between the number of tweets that include the function word ”that” (in Japanese - koto) and the number of micro-cluster that also include the function word ”that” (in Japanese - koto). By analyzing the function word ”that” (in Japanese - koto), we suppose we can get the baseline for random topics. Just as in Figure 7, the graph shows a significant correlation between the number of tweets and the number of micro clusters. Unlike in Figure 8, the fake topic transition dynamics are different.

We evaluate the progress of the fake news over time (Figure 10 and Figure 11). In Figure 10 (a), at first, the fact topic occurred. Then the rumor was diffused in Figure 10 (b). The rumor diffusion obviously show low diversity. We can see that the diffusion occurred in low diverse topics. In Figure 10 (c), the rumor correction happened. At that time, the number of tweets and the number of clusters

TABLE III
16 tweets related to the target topic and cluster id of LDA and our method

Tweet id	Body	Cluster id *LDA*	Cluster id *Ourmethod*
180	Chiba the metropolitan area It is expected that rain will fall in the Chiba and the metropolitan area due to the fire at the Cosmo Oil tank in Chiba including umbrellas and raincoats Spread it! Try it for the time being! Do not hit the rain! It looks really dangerous	441	1170
544	Information from people working for the Cosmo Oil Chiba Refinery It is said that there is a high possibility of chemical rain falling in the Kanto area such as Tokyo and Chiba due to the fire.In case of rain Be sure to use raincoat and make sure not to rain on your skin! It is very likely that carcinogenic substances are included ...	441	1170
630	Chiba it is going to rain containing chemicals from the Cosmo Oil in the metropolitan area! It seems that it should rain with an umbrella and raincoat. It seems that it is really dangerous rain. Everyone in Kanto Though you are really careful! My wife's family home is Chiba and relatives of Oi are also full of Kanto ... Is it better to believe than to worry ...	441	440
644	RT Attention to mail related to Chiba Refinery Cosmo Oil Co. Ltd. Corporate Communications Department Public Relations Office http //www.cosmo-oil.co.jp/information/110312/index.html	58	445
729	It is a message from an acquaintance of the prefectural office staff Please inform the nearby people and inform it of the transfer. people living in Chiba prefecture and the neighboring area. Cosmo Oil explosion causes harmful substances In case of going out please carry an umbrella or kappa etc. so that the body does not come into contact with the rain in case it goes out because it falls with clouds and etc and falls with rain etc ..	441	558
886	RT [Cosmo Oil Official Opinion] Although there is a statement that harmful substances adhere to clouds and the like and falls with rain etc. there is no such fact. We think that the influence of the atmosphere generated by combustion on the human body is very small. Source: http ...	441	558
1064	If the combustion gas is poisonous it can not be used in general households even if it flows out without burning there is no effect on the human body although there is concern about the greenhouse effect. The level of the hoax is low. So Cosmo Oil: LP gas stored in the tank the influence of the atmosphere generated by combustion on the human body is very small	441	760
1109	Oh it does not contain harmful things from the Cosmo Oil in the rain	59	793
1229	Residents in the vicinity of Cosmo Oil fire Nearly being extinguished it is sunny I heard it on the radio last night but it was that there was no harmful substance.In the evening last night the explosion continued The influence of the plume or the sky was covered with smoke and lightly rained temporarily but now ...	441	879
1380	Chiba people Cosmo Oil fire is currently being extinguished Sometimes harmful substances come down occasionally About RT that said umbrellas ? but announced that it was harmless last night at NHK News There was a peace of mind as a Cosmo Oil stakeholder ...	441	989
1474	The email containing chemicals from the Cosmo Oil and rain falling in Tokyo has come around like a hoax	441	1047
1514	Tokyo japan Diffuse hope Information from the factory worker. Be careful not to go out and do not expose your skin! Explosion of Cosmo Oil causes harmful substances to adhere to clouds etc. Because I get together so when I go out I carry an umbrella or kappa and my body rains ...	441	558
1598	Although there is a statement that harmful substances adhere to clouds and the like due to the explosion of Cosmo Oil and it falls with rain etc. there is no such fact. Http //is.gd/FE5P6E (Original source: http //is.gd/YxGPMN	441	558
1621	Attention should be paid to the serious matter that substances of the Cosmo Oil explosion come into contact with the skin as it rains ... From a certain source	441	1146
1656	It is informal! It is expected that rain will fall in the Chiba refinery the steel factory fire caused by fire in Chiba in the metropolitan area scientific chemicals use umbrellas and raincoats! Please spread it! Please do not rain! It seems really dangerous.	441	1170
1705	The matter of harmful rain of Cosmo Oil source This thing! There seems to be little danger ./ It was LPG gas that burned with Cosmo Oil so it's not affected	441	1206

are high and they show high diversity. We can see that the rumor correction occurred in high diversity clusters. Finally, In Figure 10 (d), the disappearance happened while diversity remained high, but gradually the number of tweets decreased. Figure 11 shows the progress of the target fake news. In the topic transition that diffuses the rumor, the topic diversity was reduced, then rose to correct the rumors. The progress shows the dynamics of the topic transition and a kind of topic life cycle.

Through the experiment, we confirmed that our method can extract quality micro-clusters by data polishing. In addition, we realized that micro-clusters can show dynamics of the topic transition.

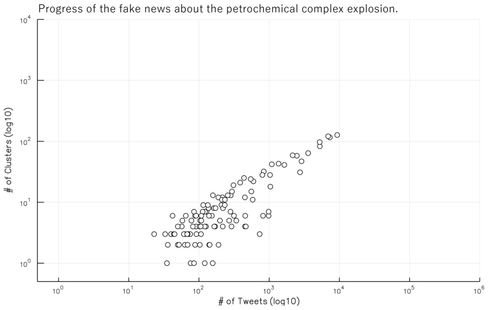


Fig. 8. # of Tweets vs # of Micro-Clusters that include the word ”cosmo oil” (base 10 log-log plot)

V. Conclusion

This paper proposed a topic transition analysis method based on micro-clustering using data polishing. Micro-clusters extracted from big data clearly show peoples opinions and reactions to topics. By evaluating micro-clusters over time, our method could present topic transitions that help people understand social situations. For one topic, different micro-clusters were made and each micro-cluster could show different opinions and or aspects of the topic.

As our future work we intend to apply our method to tweets about a greater variety of events. We also plan to develop parameters to characterize social situations. Finally, we plan to propose a model for a topic life-cycle in social media.

TABLE IV
Micro-Cluster example (excerpt) extracted from the fake news about the petrochemical complex explosion

ID	date	Time	Important words	Co-occurring words	Correspondence tweets (excerpt, quoted from Twitter)
#1	3/11	16:00-16:30	Cosmo Oil	Tank Chiba Burning Odor Factory Ichihara Petrochemical Complex Oil Complex Oil Chiba Oil Refinery What a fire Causes refinery	*Chiba's oil tank is burning. The fire at the chemical factory in Chiba is smelly. *Oil tank is burning. What are you going to do?
#2	3/11	21:30-22:00	Rain Chiba City Explosive When going out Umbrella near rain-coat Cosmo Oil harmful substance Cloud Attachment Carrying Personal Body Contact Contact Copy	Payment Notice Everyone Spread Hope	*Please [Reprint] If you live near Chiba City! Rain containing harmful chemical substances due to the explosion of petroleum industrial complex ... *[Reprint] Person who lives in the vicinity of Chiba city! ... ... Opposite with rain coat .... *[Diffusion hope] By explosion of oil tank ... (...) Do not let your body touch the rain ...
#3	3/12	12:30-13:00	Contact falling out when going out umbrella living cosmo oil hazardous substance explosive cloud adhesion rain coat	Chiba prefecture provided close together with neighborhood area informing city government office near the Ministry of Health, Labor and Welfare public announcement informed Please spread On request Chiba City everyone	* Chiba City Hall notified to the nearby people by the Ministry of Health, Labor and Welfare, Transferring ... *Those residing in Chiba prefecture and the neighboring area ... *When going out, carry an umbrella or rain coat, etc ... *Diffuse rapidly! ... (...) ... ... so that the body does not come into contact with the rain ...
#4			Harmful substances Rain says touched body bad influence spreading I'm getting off Cosmo employee staff involved Burning burning explosive blending	hope leaf honey Fake carbon dioxide	*Do not spread as it is demo. .... *It seems that rain of harmful substances in Chiba, fake. Even at NHK ... (short) Employees and stakeholders are in trouble.
# 5	3/12	17:30-18:00	Fake Cosmo Oil	Explosion Chain Mail Kuru Fire Accident Danger Caution Talking LP Gas Storage Cosmo Officially Negotiable Harmful Friendly Mail	*Take care of the Chain Mail of Fake ! ... *Cosmo Oil caution against chain mail of fake concerning fire accident *Talking about Cosmo Oil is a fake. ... *It is a fake. Cosmo officially denied.

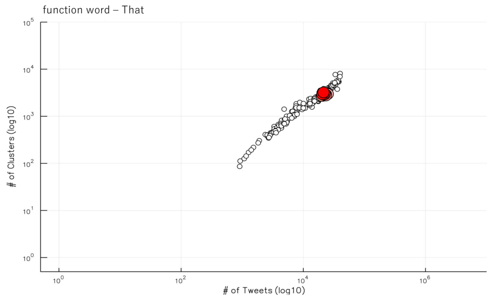


Fig. 9. # of Tweets vs # of Micro-Clusters that include the function word "Koto (That)" (base 10 log-log plot)

Acknowledgment

This work partially was supported by JST CREST, JSPS KAKENHI 18K11443 and 17H00762.

References

[1] D.M. Blei, , A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," ournal of machine Learning research 3.Jan (pp.993-1022) ,2003.

[2] T. Uno, , H. Maegawa, T. Nakahara, Y. Hamuro, R. Yoshinaka and M. Tatsuta, "Micro-clustering: finding small clusters in large diversity," arXiv preprint arXiv:1507.03067 ,2015.

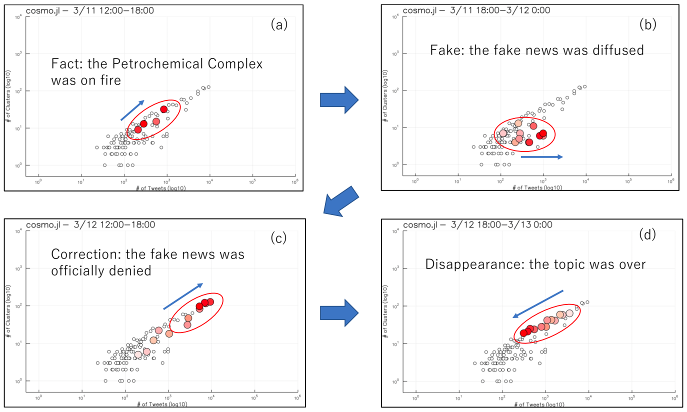


Fig. 10. Dynamics of topic transition changes of the fake news about "cosmo oil" (base 10 log-log plot)

[3] T. Uno, , H. Maegawa, T. Nakahara, Y. Hamuro, R. Yoshinaka and M. Tatsuta, "Micro-Clustering by Data Polishing," Proc, of 2017 IEEE International Conference on Big Data (BIGDATA), pp.1012-1018, 2917.

[4] H. Jin, M. Toyoda and N. Yoshinaga, "Can Cross-Lingual

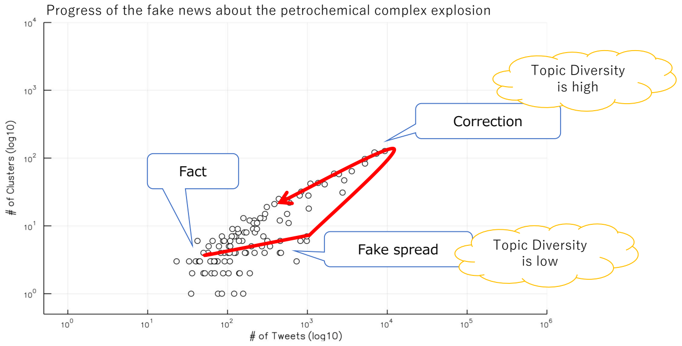


Fig. 11. Progress of the fake news about ”cosmo oil”

Information Cascades Be Predicted on Twitter?,” Proc. of the International Conference on Social Informatics, pp. 457-472, Springer, Cham, 2017.

[5] H. Kwak, C. Lee, H. Park, and S. Moon. “What is twitter, a social network or a news media?,” Proc. of the 19th International Conference on World Wide Web, pp. 591–600, ACM, 2010.

[6] T. Kitada, K. Kazama, T. S. F. Toriumi, A. Kurihara, K. Shinoda, I. Noda and K. Saito, “Analysis and Visualization of Topic Series Using Tweets in Great East Japan Earthquake,” Proc. of the 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2B3-NFC-02a-1, 2015.

[7] I. Fujino, and Y. Hoshino, “A Method for Identifying Topics in Twitter and its Application for Analyzing the Transition of Topics,” Proc. of DEIM 2014, C4-2, 2014.

[8] Y. Wang, E. Agichtein, and M. Benzi, “TM-LDA: efficient online modeling of latent topic transitions in social media,” Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 123-131, ACM, 2012.

[9] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/

[10] P. Jaccard, “The distribution of flora in the alpine zone,” New Phytologist, 11(2), 37–50, 1912.

[11] C. D. Manning, P. Raghavan, and H. Schutze, “Introduction to Information Retrieval,” Cambridge University Press, 2008.

[12] Hottolink, Inc., http://www.hottolink.co.jp/english.

[13] “lda: Topic modeling with latent Dirichlet Allocation,” https://lda.readthedocs.io/