



Analyzing temporal patterns of topic diversity using graph clustering

Takako Hashimoto¹ · David Lawrence Shepard² · Tetsuji Kuboyama³ ·
Kilho Shin³ · Ryota Kobayashi^{4,5} · Takeaki Uno⁶

Accepted: 11 September 2020 / Published online: 1 October 2020
© The Author(s) 2020

Abstract

During a disaster, social media can be both a source of help and of danger: Social media has a potential to diffuse rumors, and officials involved in disaster mitigation must react quickly to the spread of rumor on social media. In this paper, we investigate how topic diversity (i.e., homogeneity of opinions in a topic) depends on the truthfulness of a topic (whether it is a rumor or a non-rumor) and how the topic diversity changes in time after a disaster. To do so, we develop a method for quantifying the topic diversity of the tweet data based on text content. The proposed method is based on clustering a tweet graph using Data polishing that automatically determines the number of subtopics. We perform a case study of tweets posted after the East Japan Great Earthquake on March 11, 2011. We find that rumor topics exhibit more homogeneity of opinions in a topic during diffusion than non-rumor topics. Furthermore, we evaluate the performance of our method and demonstrate its improvement on the runtime for data processing over existing methods.

Keywords Social media analysis · Topic extraction · Graph clustering · Community detection · Data polishing

1 Introduction

After the East Japan Great Earthquake on 11 March, 2011, Twitter users reacted quickly and discussed a variety of topics both real and imaginary. An example is the rumor about an explosion at a petrochemical complex owned by Cosmo Oil Co., Ltd. Stories of oil tanks exploding and releasing harmful substances into the air caused widespread panic until official government announcement released on the following day. Social media has the potential to be a source of both help and trouble during

✉ Ryota Kobayashi
r-koba@edu.k.u-tokyo.ac.jp

Extended author information available on the last page of the article

disasters. Constructing strategies for disaster mitigation requires addressing issues that arise from social media as well.

Analysis and modeling of popularity dynamics of an online content has been an active area of research [1–11]. A popular method for extracting a topic is to collect all the tweets that mentioned a specific word (keyword) or hashtag and analyze the temporal patterns [1, 2, 7, 10]. While this approach makes it easy to extract the emergence of topics, we can often identify various “sub-topics” within a topic intuitively. The diversity of the content may vary greatly depending on topics.

We focus on a subtopic obtained by clustering the extracted tweet data related to a keyword (i.e., a topic). We study the *topic diversity* defined as the number of subtopics in a topic (discussed in more detail in Sect. 3.3). The topic models including Latent Dirichlet Allocation (LDA) [12] are popular method for discovering an abstract “topic” from a documents and so have been applied to social media analysis for discovering topics or sub-topics [2, 7, 10, 11]. In spite of their simplicity and usefulness, most topic model algorithms require users to specify the number of subtopics in advance. In this study, we cannot apply these models because we are interested in inferring the number of subtopics from data.

In this study, we investigate how topic diversity depends on the truthfulness of a topic (whether it is a rumor or a non-rumor) and how the topic diversity changes in time after a disaster. As a first step, we develop a method for quantifying the topic diversity of the tweet data based on text content. Our method is based on clustering a tweet graph using Data polishing [13, 14] that automatically determines the number of subtopics. Then, the proposed method is applied to a Twitter dataset before and after the Great East Japan Earthquake of 2011. We find that the temporal patterns in topic diversity differ between rumor and non-rumor topics. Finally, we evaluate the performance of the method and compare its performance with several baselines.

The contributions of this paper are as follows:

- We propose a method for analyzing topic diversity (i.e., homogeneity of opinions in a topic) based on graph clustering.
- We compare topic diversity between rumor and non-rumor topics by applying the method to Twitter dataset before and after the Great East Japan Earthquake of 2011.
- We compare the performance of the proposed method to other existing methods.

This paper is organized as follows. Section 2 introduces related work. Section 3 describes our proposed method. We apply the proposed method to a large Twitter dataset, evaluate our method and compare its performance with other existing method in Sect. 4. Finally, in Sect. 5, we conclude and discuss a direction for future research.

2 Related work

It is essential to discover a higher-level “topic” underlying the online content for summarizing and finding subtopics of collected data (e.g., tweets or hashtags). There are three approaches for topic identification: (1) topic models, (2) clustering

word features and (3) graph clustering. Topic models extract a latent topic from the frequency of words in a tweet by using a probabilistic model. LDA [12] and its extensions [15, 16] have been used to identify topics to social media data in a number of approaches [2, 7]. These algorithms have the additional disadvantage of requiring the user to specify the number of topics as a parameter. Clustering word features classifies online contents by applying clustering algorithms to feature vectors. In Rosa et al. [2], the authors developed a method for discovering a topic by applying K-means algorithm to TF-IDF vectors calculated from the tweets and showed their method performed better than LDA. Graph clustering classifies a large amount of online content by finding community structures. In Tanev et al. [17], the authors showed that the utilization of word co-occurrence graph clustering improved the performance of linking news events to tweets. While previous work identified a topic of each tweet and analyzed the temporal dynamics, our work additionally investigates the diversity of the content by calculating the number of topics. For this purpose, we exploit a graph clustering approach that does not require the number of topics to be specified in advance.

Recently, it has been argued that search engines and social networks potentially facilitate “filter bubble” effects, in which machine-learning algorithms amplify ideological segregation by recommending content targeted toward a user’s pre-existing opinions or biases [18, 19]. In Puschmann 2019, the author pointed out that a political party or political candidate is able to exert a great influence on search results, which decreases the diversity of the content people will see when they search for information about that candidate [20]. Some work has developed algorithms to address the diversity issue [21, 22]. Interestingly, Stoyanovich et al. [21] proposed ranking algorithms that achieved the diversity and fairness of the results, usually with modest costs in terms of quality. While these works focus on the diversity of the input to a user on social media, we focus on the topic diversity of a population of tweets.

In this study, we find the difference in the temporal pattern of the topic diversity between rumor and non-rumor topics, which is applicable to rumor detection algorithms. Rumor detection is an active research area, and several methods have been proposed [23–27]. While these methods are based on the machine learning methods, our method is based on clustering a graph of word co-occurrence. Unfortunately, it is difficult to select most relevant features in machine learning problems when there exist a large number of feature variables [28, 29]. The graph-based approach has three potential benefits: (1) it is simple to implement, (2) it is applicable to large data set, and (3) the result is easy to interpret.

Finally, we discuss the relation between this work and our previous one [10]. In the previous work, we proposed a method for visualizing temporal topic transition based on graph clustering. Since the method was proposed for visualization, we demonstrated the topic transition of a few topics. We neither analyzed the temporal patterns quantitatively nor examined the performance of the method. In this work, the method has been extended to quantify the content diversity of a topic by using correlation analysis and quantile regression. We also compare the temporal patterns of the topic diversity between rumor and non-rumor topics.

Furthermore, we evaluate the performance of the method and demonstrate its improvement on the runtime for data processing over existing methods.

3 Proposed method

Our method consists of three steps (Fig. 1): (a) Construction of tweet-word matrices, (b) Tweet graph generation, (c) Graph clustering using Data polishing.

3.1 Construction of tweet-word matrices: Fig. 1a

We analyze tweet data posted from 00:00 Japan Standard Time (JST) on March 11 to 0:00 JST on March 14, a total of 72 h. The tweet data were divided into 144 ($= 72 \div 0.5$) groups based on their posted time and a fixed time window of half an hour. We construct a sequence of tweet-word matrix for each window: $\langle W_1, W_2, \dots, W_{144} \rangle$. Tweet-word matrix (W) denotes the constituent word of the tweets

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{pmatrix}$$

where m and n are the number of tweets and words during a time window, respectively. The element w_{ij} is 1 if i th tweet contains the j th word, and 0 otherwise. We used the morphological analyzer MeCab [30] to perform Japanese word segmentation.

3.2 Tweet graph generation: Fig. 1b

We generate tweet graph from the tweet data in each time window. We define a tweet graph as an undirected graph, in which a node represents a tweet and an edge represents that the connected tweets are similar (Fig. 1): If the Jaccard coefficient [31] of two tweets is larger than the edge threshold θ_E , these nodes (tweets) are connected by an edge. The threshold was set to $\theta_E = 0.3$.

The main result (Fig. 3) is robust for the small change of the threshold: The result did not change qualitatively for different thresholds $\theta_E = 0.2$.

3.3 Clustering tweet graph: Fig. 1c

We here briefly describe our Data polishing algorithm for clustering a tweet graph (see [13, 14] for more detail). This algorithm iteratively increases the density of dense subgraphs and makes sparse subgraphs sparser. As a consequence, we obtain a graph whose dense subgraphs are all cliques and can thus easily be enumerated by a maximal clique enumeration algorithm.

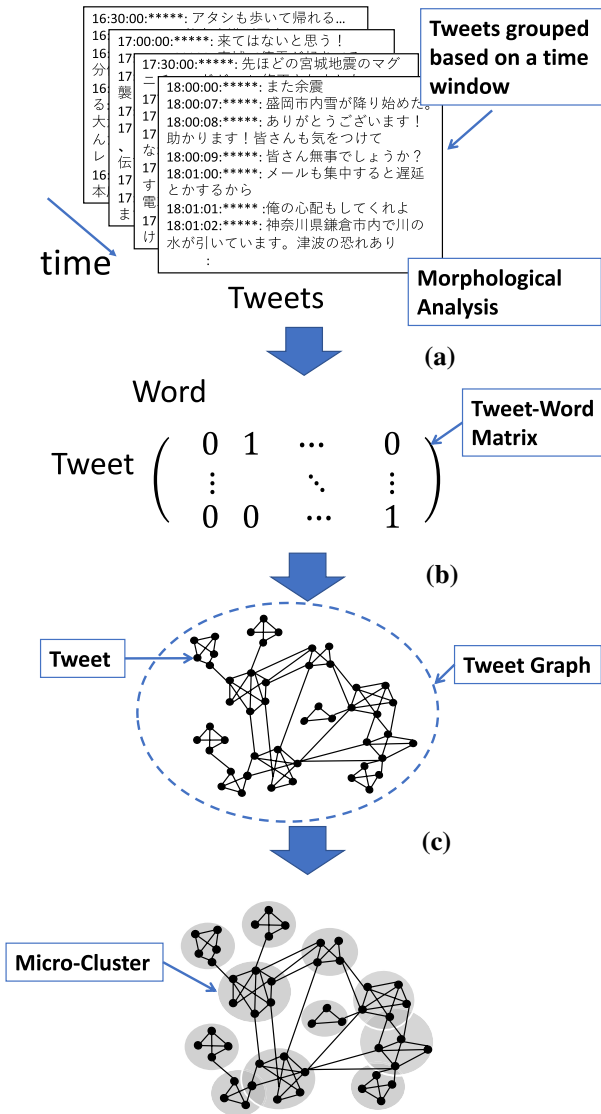


Fig. 1 Proposed method. Our method consists of three steps. **a** Construction of tweet-word matrices, **b** tweet graph generation, **c** clustering tweet graph

The iteration is described as follows. For arbitrary two vertices u and v , we consider the condition

$$|N[u] \cap N[v]| / |N[u] \cup N[v]| > \theta_p,$$

This intuitively means that u and v belong to the same cluster, since u and v 's neighbor sets must overlap considerably when they belong to the same cluster. We construct a new graph by using this condition so that two vertices are connected

when this condition is met in the original graph. In the other words, two vertices are connected in the new graph if and only if the vertices seem to belong to the same cluster. Data polishing applies this graph reconstruction iteratively until it does not change. Maximal clique enumeration is performed with an algorithm such as MACE[32], to obtain the clusters of the resulting graph. The threshold θ_p was set to 0.2. The threshold value affects the clustering result. The size of clusters increases as the threshold θ_p decreases.

In this paper, we analyze the Tweet graph related to a topic (e.g., non-rumor topic such as “I’m OK” and rumor topics). The resulting cluster is interpreted as a subtopic, and the number of clusters is interpreted as the diversity of the topic (topic diversity). For instance, suppose that there are two topics, and Topic A consists of 30 subtopics (clusters) and Topic B consists of 5 subtopics (Fig. 2). In this case, we interpret that Topic A is more diverse than Topic B.

4 Experiments

We first describe the Twitter dataset. Second, we analyze the temporal pattern of the topic diversity for rumor and non-rumor topics. Finally, we compare the performance of the proposed method with the existing clustering methods.

4.1 Dataset

Our data set consists of tweets posted around the time of the Great East Japan Earthquake that happened at 14:47 JST on March 11, 2011. This dataset was obtained from the social media monitoring company Hotto link Inc. [33], which tracked users who used one of 43 hashtags (for example, #jishin, #nhk, and #prayforjapan) or one of 21 keywords related to the disaster. Later, Hottolink collected all tweets posted by all of these users between March 9th (2 days prior to the earthquake) and March 29th. The total number of is around 200 million tweets, which offers one of a largest data set for users’ responses to a disaster. We focused on the dataset from 00:00 JST on March 11 to 0:00 JST on March 14, a total of 72 h.

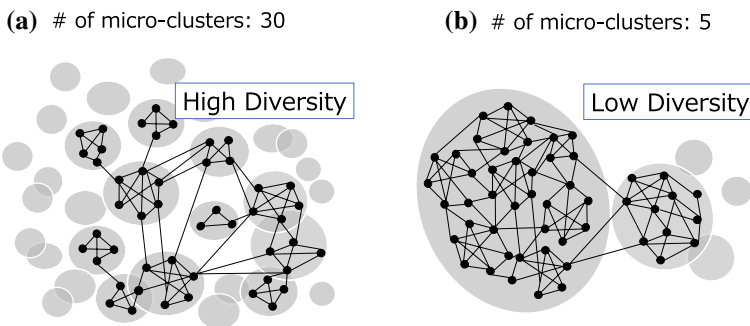


Fig. 2 Interpretation of the number of the clusters (# of clusters)

We picked out 10 topics that contain the following keywords: “do” (“suru” in Japanese), “I’m OK”, “go home”, “important”, “safe”, “damage”, “Fukushima”, “Miyagi”, “Cosmo Oil”, and “Isodine”. The collected topics consists of eight non-rumor topics (“do”, “I’m OK”, “go home”, “important”, “safe”, “damage”, “Fukushima”, and “Miyagi”) and two rumor topics (“Cosmo Oil” and “Isodine”). Note that the word “suru” (an auxiliary verb often translated as “do” or ignored when translating to English) is a common function word; a user’s usage of “suru” does not depend on the topic that they are tweeting about. Analyzing the usage of “suru” gives us a baseline for a general topic. The other words were heavily used in tweets after the East Japan Great Earthquake. Particularly, the two rumor topics (“Cosmo Oil” and “Isodine”) were related to well-known rumors spread after the earthquake [34]. A detailed description of the rumors follows.

“Cosmo Oil” An explosion at the Cosmo Oil plant released harmful substances into the air.

“Cosmo Oil” is the name of a Japanese oil company. The rumor about the oil tank explosion was diffused and frightened people. A rumor topic about the explosion at the Cosmo Oil petrochemical complex progressed through the four stages:

1. *Fact* Around 15:00 JST on March 11 (just after the quake), the petrochemical complex in Chiba caught fire.
2. *Rumor* Around 19:00 JST on March 11, the following two tweets were posted and retweeted:
 - 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.
3. *Correction* Around 15:00 JST on March 12 (the day after the earthquake), the industry’s Web site and the local government’s twitter officially corrected the rumor.
4. *Disappearance* At night on March 12, the topic disappeared.

“Isodine” Isodine was good for protecting from radiation.

“Isodine” is the brand name of a mouthwash that includes iodopovidone. A rumor emerged that isodine protects people from radiation. It progressed in the following four stages:

1. *Fact* After the nuclear plant explosion occurred, twitter users expressed fear of radiative contamination.
2. *Rumor* Around 7:00 JST on March 12 (the day after the earthquake), the rumors about isodine’s protective benefits emerged.
3. *Correction* Around 15:00 JST on March 12 (the day after the earthquake), the government and isodine’s manufacturer corrected the rumor.

4. *Disappearance* At night on March 12, the topic gradually disappeared.

4.2 Analyzing temporal patterns of topic diversity

We examine the temporal patterns of topic diversity of tweets including 10 keywords (e.g., topics) from 0:00 JST, 11 th March to 0:00 JST, 14 th March. We calculated the topic diversity for each 30 minutes widow by the following procedure. First, the word count matrices were constructed from the collected tweets (Sect. 3.1). Second, the tweet graph was generated (Sect. 3.2) and the topic diversity (i.e., the number of subtopics) was obtained by applying Data polishing algorithm to each graph (Sect. 3.3).

Figure 3 shows log–log scatter plot of the number of tweets and the topic diversity for 10 topics and for all the dataset. Each circle represents the number of tweets and the topic diversity at a time window. The color in a circle represents the time: The color changes from white to dark blue as the time passes. For all the dataset (Fig. 3a) and non-rumor topics (Fig. 3b–i), we observe that the topic diversity is highly correlated with the number of tweets (Pearson correlation coefficient ranged from 0.953 to 0.996). This result suggests the power law relationship between the topic diversity and the number of tweets. For rumor topics (Fig. 3j, k), we observe that the scatter plot disperses from the linear fit compared to the non-rumor ones. The topic diversity is less correlated with the number of tweets (Pearson correlation coefficient 0.958 and 0.938 for “Cosmo Oil” and “Isodine”, respectively).

To quantify the dispersion in the scatter plot, we applied the quantile regression [35, 36] that fits the top and bottom 5% data point. Figure 3 depicts the top and bottom 5% regression lines (in blue and red, respectively) and their slopes. These slopes are close for all the dataset (Fig. 3a) and non-rumor topics (Fig. 3b–i). In contrast, these slopes are different rumor topics (Fig. 3j, k): 0.73 vs 0.59 and 0.83 vs 0.60 for “Cosmo Oil” and “Isodine”, respectively. This result indicates that the scatter plot of the rumor topics disperses compared the non-rumor topics, which is consistent with the correlation analysis. In addition, we found that the slopes of the general topic “do” (“suru” in Japanese) was smaller than that of the specific ones. A possible reason could be that 1) there are a huge difference in the number of tweets between “do” and the other topic and 2) the slope tends to decrease when the number of tweets is large. Thus, the slope may not be useful for comparison of two topics when the number of tweets in each topic is not comparable.

Next, we observe how the topic diversity of the rumor topics changed before and after the earthquake. Figure 4a shows the time course of the topic diversity of the rumor about “Cosmo Oil”. We can see that the topic was not popular before the earthquake. While the number of tweets increased dramatically just after the earthquake, the topic diversity does not increase much, showing that the topic burst with relatively homogenous opinions, or low diversity. After the rumor correction was issued, both the number of tweets and the topic diversity increased and the topic diversity grew higher than that before the rumor correction. Figure 4b shows the time course of the topic diversity of the rumor about “Isodine”. We can see that the temporal pattern of the topic diversity is similar

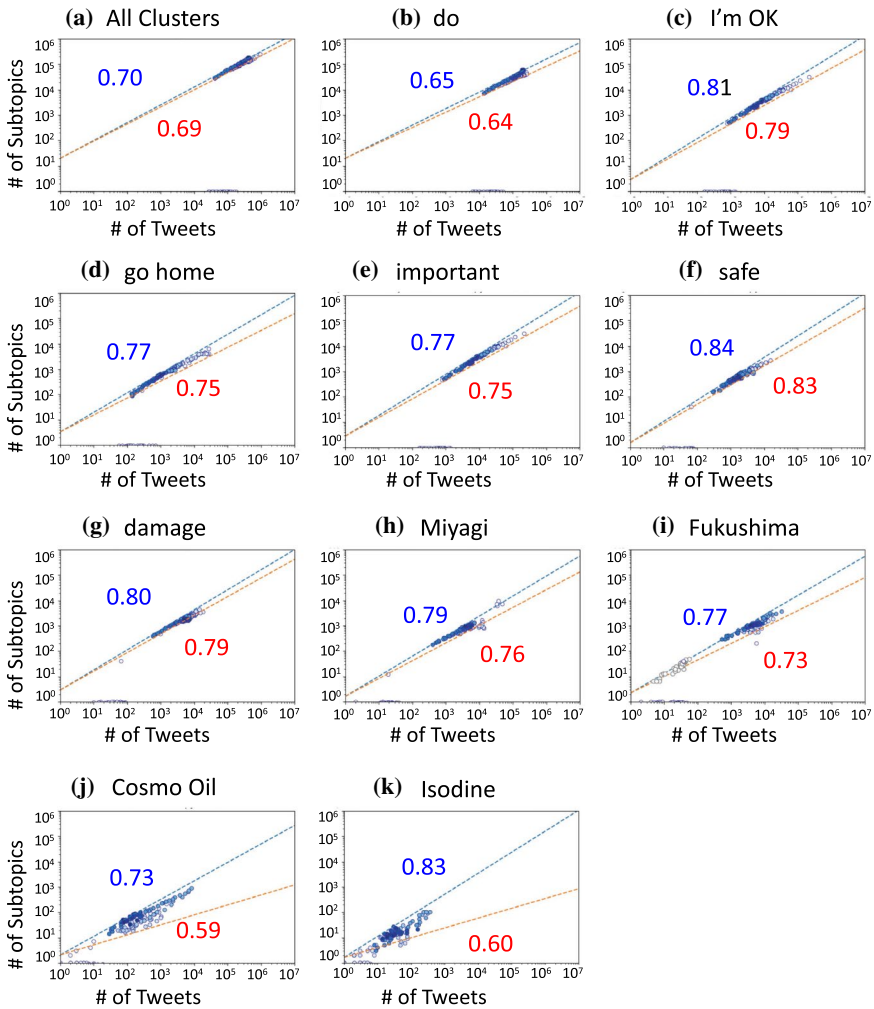


Fig. 3 The number of tweets (# of Tweets) vs the number of subtopics (# of Subtopics) for all the dataset (a) and for 10 keywords (b–k). The slope of the top 5% and bottom 5% regression lines are shown in blue and red, respectively (color figure online)

to that of “Cosmo Oil”. Interestingly, both rumor topics spread with low topic diversity and they were corrected with high topic diversity.

4.3 Performance evaluation

In this section, we examine the runtime and the size of subtopics generated by the proposed method and existing methods.

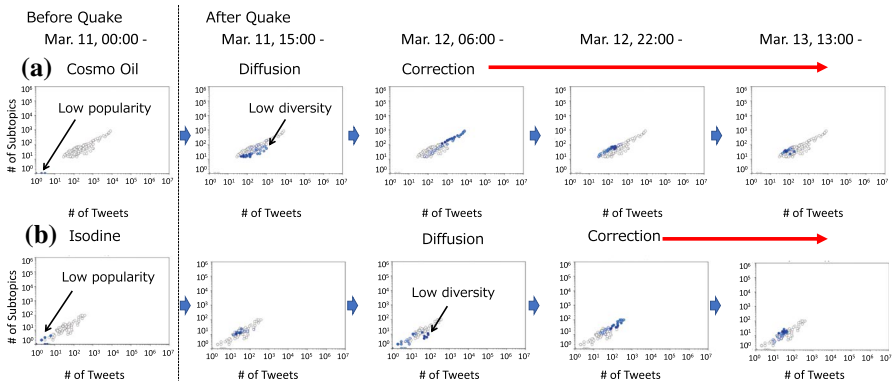


Fig. 4 Time course of topic diversity. Cosmo oil (a) and isodines (b)

4.3.1 Runtime

We examine the runtime for processing the tweet dataset by the proposed method based on Data polishing (Sect. 3.3) and compare the performance with four existing methods for clustering subtopics: LDA [12], K-means [37], MeanShift [38] and Agglomerative clustering [39]. LDA is the most popular topic model algorithm based on word frequency across documents. K-means, MeanShift and Agglomerative clustering are general clustering algorithms that were applied to word vectors $\mathbf{w}_j = (w_{j1}, w_{j2}, \dots, w_{jn})$ ($j = 1, 2, \dots, m$) (Sect. 3.1). Data polishing is a graph clustering approach and was applied to the tweet graph (Sect. 3.2). We used Nysol Python [40]'s implementation of Data polishing, and scikit-learn's implementations of K-means, MeanShift, and Agglomerative Clustering [41], and LDA implemented in Python Gensim [42]. All our experiments were performed on a 2018 13-inch MacBook Pro, with a 2.7 GHz Intel Core i7 with 16 GB 2133 MHz of memory.

Figure 5 illustrates the performance of the algorithms, as the number of tweets was increased from 2000 to 30,000. These tweets were randomly extracted from the Tweet data posted during 18:30–19:00 JST on March 11. The runtime was measured by executing each method for three times and calculating their average. Note that we stopped the measurement if the runtime reached 2 h (7200 s). While Data polishing and MeanShift can automatically determine the number of subtopics from the data, the other methods cannot do it. For these methods, the number of subtopics was assigned as the number that Data polishing determined. The numbers of subtopics estimated by Data polishing were smaller than those by MeanShift. The result shows that Data polishing is more efficient than the existing algorithms. The second most efficient one is LDA. The proposed method is five times faster than LDA for the sample size of 20,000 and at least six times faster than LDA for the sample size of 30,000. LDA code was terminated because the runtime exceeded 2 h.

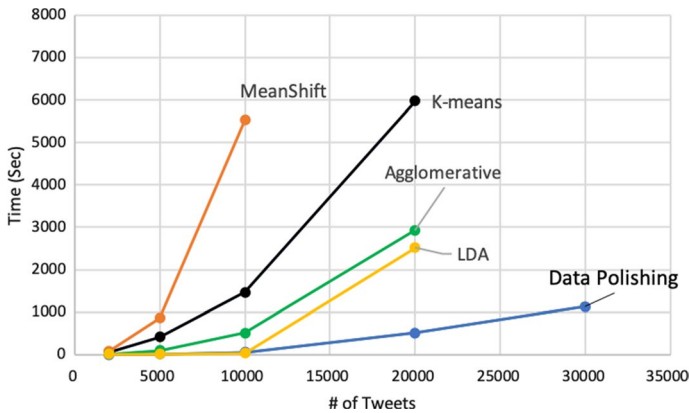


Fig. 5 Performance evaluation. The runtime of the proposed method (data polishing: blue) was compared with four existing methods (MeanShift: orange, K-means: black, agglomerative: green, and LDA: yellow). We stopped the measurement if the runtime reached 2 h (7200 s). Please note that the color version is available at the online version of this article (color figure online)

4.3.2 Number of tweets in top 5 largest subtopics

We randomly extracted 10,000 tweets posted during 18:30–19:00 JST on March 11 and analyzed the Tweet data by using the proposed method and the existing methods (LDA, K-means, MeanShift, and Agglomerative clustering). Table 1 shows the number of tweets in the top 5 largest subtopics. K-means and MeanShift each generated one large cluster (more than 14% of the total tweet) as a largest subtopic. Typically, these method tend to result in one huge subtopic and many small subtopics, which often leads to a trivial clustering or difficulty for interpreting the results. In contrast, the proposed method (Data polishing) generated the smallest cluster as the largest subtopic (~ 4% of the total tweets). Though LDA and Agglomerative clustering’s largest subtopics were smaller than K-means’ and MeanShift’s, they require specifying the number of subtopics ahead of time.

Table 1 The number of tweets for top 5 largest subtopics

Rank	Data polishing	LDA	K-means	MeanShift	Agglomerative
1	428	780	1466	1745	790
2	277	266	276	65	102
3	176	196	209	52	101
4	144	188	155	28	63
5	123	172	145	23	59

Data polishing (proposed method), MeanShift, and their results are shown by bold letters, because these methods can automatically determine the number of subtopics

5 Conclusion

In this paper, we have proposed a method for analyzing the topic diversity using Graph Clustering. After generating the tweet graph based on the similarity between tweets, we use Data polishing algorithm to obtain the clusters in the graph. We interpret a cluster as a subtopic and the number of clusters as the topic diversity, i.e., the number of subtopics in a target topic. This method was applied to a dataset of millions of tweets posted before and after the Great East Japan Earthquake of 2011. The proposed method is useful for detecting a low topic diversity situations, and the rumor diffusion associates with the burst with low topic diversity or homogeneous opinions. We have confirmed that our approach outperformed other existing clustering approaches (LDA, K-means, MeanShift, and Agglomerative clustering) in running time and results. Our method has significant applications in that it could, for example, companies or celebrities about newly circulating rumors, and help them avoid bad publicity on social media.

The main limitation is that we focused on only 10 topics in this study. We have performed a case study of rumor diffusion on Twitter during a disaster rather than a general study of rumor diffusion. We are planning to apply the proposed method to another rumor dataset [43] and systematically investigate temporal patterns in topic diversity toward the development of a practical rumor detection algorithm.

Acknowledgements This work was partially supported by JST CREST JPMJCR1401, JST ACT-I JPM-JPR16UC, JST PRESTO JPMJPR1925, JSPS KAKENHI JP17H03279, JP18K11560, JP19H01133, JP19K12125, JP18K11443 and JP17H00762.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp 591–600
2. Rosa KD, Shah R, Lin B, Gershman A, Frederking R (2011) Topical clustering of tweets. In: Proceedings of the ACM SIGIR: SWSM, pp 63–70
3. Matsubara Y, Sakurai Y, Prakash BA, Li L, Faloutsos C (2012) Rise and fall patterns of information diffusion: model and implications. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 6–14
4. Hashimoto T, Shepard D, Kuboyama T, Shin K (2015) Event detection from millions of tweets related to the great East Japan earthquake using feature selection technique. In: 2015 IEEE International Conference on Data Mining Workshop. IEEE, pp 7–12
5. Ikeda K, Sakaki T, Toriumi F, Kurihara S (2016) An examination of a novel information diffusion model: considering of twitter user and twitter system features. In: International Conference on Autonomous Agents and Multiagent Systems. Springer, pp 180–191

6. Kobayashi R, Lambiotte R (2016) Tideh: time-dependent hawkes process for predicting retweet dynamics. In: Tenth International AAAI Conference on Web and Social Media, pp 191–200
7. Grabowicz PA, Ganguly N, Gummadi KP (2016) Distinguishing between topical and non-topical information diffusion mechanisms in social media. In: Tenth International AAAI Conference on Web and Social Media, pp 151–160
8. Jin H, Toyoda M, Yoshinaga N (2017) Can cross-lingual information cascades be predicted on twitter? In: International Conference on Social Informatics. Springer, pp 457–472
9. Proskurnia J, Grabowicz P, Kobayashi R, Castillo C, Cudré-Mauroux P, Aberer K (2017) Predicting the success of online petitions leveraging multidimensional time-series. In: Proceedings of the 26th International Conference on World Wide Web, pp 755–764
10. Hashimoto T, Uno T, Kuboyama T, Shin K, Shepard D (2019) Time series topic transition based on micro-clustering. In: 2019 IEEE International Conference on Big Data and Smart Computing. IEEE, pp 1–8
11. Ifeoluwa Adelani D, Kobayashi R, Weber I, Grabowicz PA (2020) Estimating community feedback effect on topic choice in social media with predictive modeling. *EPJ Data Sci* 9(1):25
12. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
13. Uno T, Maegawa H, Nakahara T, Hamuro Y, Yoshinaka R, Tatsuta M (2015) Micro-clustering: finding small clusters in large diversity. arXiv preprint [arXiv:1507.03067](https://arxiv.org/abs/1507.03067)
14. Uno T, Maegawa H, Nakahara T, Hamuro Y, Yoshinaka R, Tatsuta M (2017) Micro-clustering by data polishing. In: 2017 IEEE International Conference on Big Data. IEEE, pp 1012–1018
15. Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp 113–120
16. Wang Y, Agichtein E, Benzi M (2012) TM-LDA: efficient online modeling of latent topic transitions in social media. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 123–131
17. Tanev H, Ehrmann M, Piskorski J, Zavarella V (2012) Enhancing event descriptions through twitter mining. In: Sixth International AAAI Conference on Weblogs and Social Media, pp 587–590
18. Pariser E (2011) *The filter bubble: what the Internet is hiding from you*. Penguin, London
19. Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. *Public Opin Q* 80(S1):298–320
20. Puschmann C (2019) Beyond the bubble: assessing the diversity of political search results. *Digital J* 7(6):824–843
21. Stoyanovich J, Yang K, Jagadish HV (2018) Online set selection with fairness and diversity constraints. In: Proceedings of the EDBT Conference, pp 241–252
22. Charkov M, Gupta S (2019) Re-ranking search results for location refining and diversity. US Patent App. 16/356,811
23. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining. IEEE, pp 1103–1108
24. Derczynski L, Bontcheva K, Liakata M, Procter R, Hoi GWS, Zubiaga A (2017) SemEval-2017 task 8: RumourEval: determining rumour veracity and support for rumours. In: Proceedings of the 11th International Workshop on Semantic Evaluation, pp 69–76
25. Poddar L, Hsu W, Lee ML, Subramaniyam S (2018) Predicting stances in twitter conversations for detecting veracity of rumors: a neural approach. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence. IEEE, pp 65–72
26. Ma J, Gao W, Wong K-F (2018) Rumor detection on twitter with tree-structured recursive neural networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp 1980–1989
27. Yoshida Z, Aritsugi M (2019) Rumor detection in twitter with social graph structures. In: Third International Congress on Information and Communication Technology. Springer, pp 589–598
28. Shenkai G, Cheng R, Jin Y (2018) Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Comput* 22(3):811–822
29. Piotrowski AP, Napiorkowski JJ (2018) Some metaheuristics should be simplified. *Inf Sci* 427:32–62
30. Kudo T (2006) Mecab: yet another part-of-speech and morphological analyzer. <http://mecab.sourceforge.jp>
31. Jaccard P (1912) The distribution of the flora in the alpine zone. 1. *New Phytol* 11(2):37–50

32. Makino K, Uno T (2004) New Algorithms for Enumerating All Maximal Cliques. In: Proceedings of 9th Scandinavian Workshop on Algorithm Theory. Springer, pp 260–272
33. Hottolink, inc (2020) <http://www.hottolink.co.jp/english/>
34. Nabeshima K, Mizuno J, Okazaki N, Inui K (2014) Mining false information on twitter for a major disaster situation. In: International Conference on Active Media Technology, pp 96–109
35. Koenker R, Hallock KF (2001) Quantile regression. *J Econ Perspect* 15(4):143–156
36. Quantile Regression (2019) <https://cran.r-project.org/web/packages/quantreg/quantreg.pdf>
37. Arthur D, Vassilvitskii S (2007) K-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp 1027–1035
38. Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans Inf Theory* 21(1):32–40
39. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
40. NYSOL Python (2020) <https://www.nysol.jp/>
41. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
42. Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp 45–50
43. Li Q, Zhang Q, Si L, Liu Y (2019) Rumor detection on social media: datasets, methods and opportunities. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pp 66–75

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Takako Hashimoto¹  · David Lawrence Shepard² · Tetsuji Kuboyama³ · Kilho Shin³ · Ryota Kobayashi^{4,5} · Takeaki Uno⁶

Takako Hashimoto
takako@cuc.ac.jp

David Lawrence Shepard
shepard.david@gmail.com

Tetsuji Kuboyama
ori-bigdata2019@tk.cc.gakushuin.ac.jp

Kilho Shin
kilhoshin314@gmail.com

Takeaki Uno
uno@nii.jp

¹ Chiba University of Commerce, Chiba, Japan

² Evidation Health, California, USA

³ Gakushuin University, Tokyo, Japan

⁴ The University of Tokyo, Tokyo, Japan

⁵ JST, PRESTO, Saitama, Japan

⁶ National Institute of Informatics, Tokyo, Japan