Topic Life Cycle Extraction from Big Twitter Data based on Community Detection in Bipartite Networks

Takako Hashimoto*, Hiroshi Okamoto†, Tetsuji Kuboyama‡, and Kilho Shin§

*Chiba University of Commerce, Chiba, Japan 272-8512, Email: takako@cuc.ac.jp
†RIKEN Brain Science Institute, Saitama, Japan, Email: hiroshi.okamoto299792458@gmail.com
‡Gakushuin University, Tokyo, Japan, Email: kuboyama@tk.cc.gakushuin.ac.jp
§University of Hyogo, Kobe, Japan, Email: kilhoshin314@gmail.com

Abstract—This paper is showing a time series topic life cycle extraction from millions of Tweets using our original community detection technique in bipartite networks. We suppose that the authors role that means who belong to what topics is important to extract quality topics from social media data. We already proposed the topic extraction method that considers the relationship between the authors and the words as bipartite networks and explores the authors role by forming clusters as topics. As the next step, this paper applies our method to the time series topic life cycle detection. We extract topics in different time slots and analyze the time series of topic transition using the coherence measure that expresses the semantic accuracy of topics. The paper demonstrates that our method can detect the topic life cycle such as the growth, the conflicts and so on over time from millions of Tweets.

I. Introduction

This paper is trying a time series topic life cycle extraction from millions of Tweets using the community detection technique in bipartite networks [1]. We suppose that the authors role that means who belong to what topics is important to extract quality topics from social media data. If there are two topics that have similar words but have different authors, they may sometimes be considered as different topics. We already proposed the method that considers the relationship between the authors and the words as bipartite networks and explores the authors role by forming clusters as topics [2] and [3]. For topic extraction, the author's role can be considered, as well as the word's role based on bipartite network structures. As the next step, this paper further employ our already-proposed method for the time series topic life cycle detection. We extract topics in different time slots and analyze the time series of topic transition using the coherence measure [4] that expresses the semantic accuracy of topics. The paper demonstrates that our method can detect the topic life cycle such as the growth, the conflicts and so on over time from millions of Tweets. We also compare our method with LDA [5] that is one of major topic model and confirm our method's strength for the topic life cycle detection. By our experiment, we confirm that considering author's role is effective to show the lifecycle of topics.

This paper is organized as follows. Section II describes related work on time series topic extraction from SNS such as Twitter. Section III briefly introduces our already-proposed technique on community detection from bipartite networks [1]. Section IV summarizes our topic extraction method [2] [3] considering authors role based on bipartite networks and applies our method to the topic life cycle detection. Our experimental results are compared with results by LDA. Finally, Section V concludes this paper and offers directions for future research.

II. Related Work

Endoh et al. [6] proposed the emerging topic extraction method from Twitter based on Non-negative Matrix Factorization (NMF). They efficiently tried to extract quality topics related to a specific theme, however user should input a specified query of words in advance. Fujino et al. [7] and Wang et al. [8] analyzed Tweets over time based on LDA. Zhao et al. [9] analyzed Twitter and news articles using LDA. They don't consider author groups either, and define topic categories in advance to improve the quality of topics. They are still facing the accuracy problem. Kitada et al. [10] targeted 200 million Tweets related to the Great East Japan Earthquake, and tried to analyze them over time by LDA based technique. To analyze a topic transition over time, they computed the cosine similarity of adjoining topics in chronological order. They focus only on words similarity to analyze the time series of topic transition. These conventional topic extraction methods based on LDA require topic categories or corpus in advance to improve the topic extraction quality. Our method is based on our original community detection technique in bipartite networks. Considering the author groups helps more accurate time series of topic extraction.

Ferrara et al. [11] focused on user groups to detect social media campaigns on Twitter using DTW (Dynamic Time Warping) [12]. Hu et al. [13] also explored user groups on social media, because they supposed active users were essential for emerging topics. They were mainly focusing on user groups transition. On the other hand, our research aims to analyze the time series changes between authors and words.

Mucha et al. [14] proposed the time series community detection method in multiplex network. However it considered single networks but did not consider bipartite networks. For bipartite networks, we need to expand Mucha's method.

III. COMMUNITY DETECTION TECHNIQUE FROM BIPARTITE NETWORKS

This section briefly introduces the community detection technique from bipartite networks that was already proposed by [1].

In network science, a "communities" refers to a group of nodes that are densely connected each other and are more sparsely connected with nodes outside the group [15]. Detecting communities in networks is of essential importance for finding functional modules of complex systems described by networks. To achieve soft clustering of words appearing in Tweets, therefore, the present study has conducted community detection in bipartite networks of users and words. Community detection in the present study has adopted a technique which allows for defining overlapping between communities. This technique exploits a random walk in the network from which we wish to detect communities, and is also applicable to bipartite networks such as those examined in the present study.

Individual communities are characterized by the conditional probabilities, because $p(n \mid k)$ can be viewed as the relative importance of node n in community k. In our method, community detection is solved by EM algorithm.

The $p(n \mid k)$ is generally positive for different k's; this means that node n belongs to more than one community. The algorithm has only one parameter, α. The magnitude of this parameter controls the resolution of community detection; the smaller its magnitude, the larger the number of detected communities. In addition, the number of communities k that should be detected is provided by users.

IV. TIME SERIES TOPIC EXTRACTION BASED ON COMMUNITY DETECTION IN BIPARTITE NETWORKS

At first, we summary our already-proposed method using our original community detection technique described in Section III to extract topics from big Twitter data. We first group the Tweets by a certain period according to the time they were sent. Then, we create the sequence of Author-Word Count Matrices. Author-Word Count Matrices can be considered as bipartite networks that consist of authors and words (See Figure 1).

From these bipartite networks, our method forms clusters as topics and selects feature authors and words that have high belongingness for each topic. For forming clusters (topics), our method allows intersections. To analyze the semantic accuracy of topics, we utilize the topic coherence measure [4] that shows the semantic accuracy.

As the next step, the paper applies our method to the topic life cycle detection. To show our method's effectiveness, we conducted the following experiment.
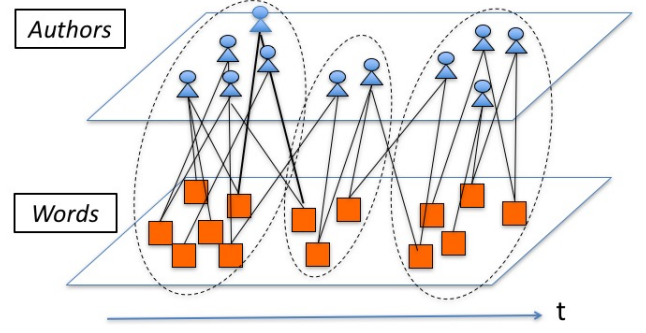

Fig. 1. Bipartite networks by authors and words

A. Target Data

Our target data is over 200 million Tweets sent during the time of the Great East Japan Earthquake that happened at 14:47 on March 11, 2011. The social media monitoring company Hottolink [16] 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 recorded all Tweets sent by all of these users between March 9th (2 days prior to the earthquake) and March 29th. This resulted in an archive of around 200 million Tweets, sent by around 1 million users. An average of about 8 million Tweets were posted by around 200 thousand authors per day. The average data size per day was around 8GB, and the total data size was over 150GB. This dataset offers a significant document of users' responses to a crisis, but its size presents a challenge.

B. Step 1: Creation of Author-Word Count Matrices

We began by creating Author-Word count matrices for our dataset. To segment Tweets that may not have used spaces to delineate word boundaries, we employed the fast and customized Japanese morphological analyzer, MeCab [17]. Then we created Author-Word count matrices for 30 minutes.

C. Step 2: Clustering and Step 3: Feature Selection

We set $\alpha = 0.001$ and $k = 100$ and run the program developed based on our technique that was explained in Section III. We also set the number of EM iteration as 600. Our method formed clusters (≤ 100) until they converges.

To compare our experimental result with LDA, we leveraged the `GibbsLDA++` library provided by Xuan-Hieu Phan and Cam-Tu Nguyen [18]. It is a `C/C++` implementation of Latent Dirichlet Allocation (LDA) using the Gibbs Sampling technique for parameter estimation and inference. 80 topics were extracted by 1000 Gibbs sampling iteration. As Step 3, feature authors and words were extracted by both our method and LDA.

Table I and Table II show the feature words of the top 10 topics on Mar. 11 16:00-16:30. The top 10 topics have high coherence values (as we discuss later) for our method and LDA respectively.

Topic Id	Coherence	
92	-525.19	appearance, siren, information, source, Sendai, key, scream, helicopter, angry, only
82	-533.31	blur, impossible, all, energization, zone, earth, leakage, breaker, electric, shop, advice, blackout
72	-595.62	Emergency, Embedded, Helicopter, Interview, The, Media, Site, Hope, Volume, Worker
35	-606.86	Narita, Line, Mito, Yamanote, Line, Driving, Joetsu, Match, Saikyo, Line, Keiyo, Mizu-cho, Line
22	-647.27	bath, gas, possibility, electricity, evacuation, securing, main, plug, place, water, outage
69	-669.22	bleeding, root, diffusion, wound, upper, arm, hope, head, case, inside
8	-739.17	diffusion, hope, tsunami, warning, blackout, people, Miyagi, Naruko, situation, earthquake
37	-764.77	Coastal, Pacific, Hokkaido, Chiba, Aomori, Miyagi, Tsunami, warning, Iwate, prefecture, Kujukuri
51	-773.18	number, message, reproduction, safety, disaster, message, dialing, confirmation, message, dialing, recording, home
52	-778.1	Enoshima, Niigata, Confirmation, Thailand, Morita, Tamurari, Ueda, Kana, Kusunoki, Dai, Onosaka, Masaya

TopicId	Coherence	
2	-748.30	coastal, Pacific, Hokkaido, Chiba, Aomori, Kujukuri, Miyagi, prefecture, outer, chamber
77	-831.48	safe, confirmation, okay, members, Morita, Tamurari, Ueda, Kana, Kusunoki, Daisuke
35	-1090.88	evacuation, coast, high, ground, big, tsunami, warning, neighborhood, estuary, renewal, tsunami, warning
0	-1120.85	Skype, phone, smartphone, official, contact, Tweet, attention, serious
70	-1141.94	confirmation, family, earthquake, photograph, mobile, phone, charger, fire, emergency, food
16	-1145.83	Tsunami, warning, Iwate, Ibaraki, Tohoku, Tsunami, warning, Fukushima, Miyagi, Kanto
39	-1171.53	Kyushu, Tsunami, warning, Tsunami, warning, Coastal, Kagoshima, Around, Present, Oita
7	-1177.87	gas, evacuation, occurrence, confirmation, door, serious, food, contact
11	-1226.61	driving, all, lines, matching, influence, bullet, train, closing, road, closure, Narita, line
21	-1274.39	Disaster, Message, Board, Mobile, Disaster, DoCoMo, Message, Board, Start, Operation, Access

D. Step 4: Topic Detection

We computed the coherence measure that was proposed by Mimno [4]. Table I and Table II show the top 10 topics on Mar. 11 16:00-16:30 that have high coherence values. From Table I and Table II, we understand both our method and LDA could extract popular topics. For example, topics about tsunami warning ($TopicId = 8$ and 37 in Table I and $TopicId = 35$ and 16 in Table II), energy problems ($TopicId = 82$ and 22 in Table I and $TopicId = 7$ in Table II) and sending messages

($TopicId = 51$ in Table I and $TopicId = 21$ in Table II) were extracted as the top 10 topics.

We then confirmed if specific topics such as fake news were extracted by our method. As an example, we selected the following rumor about the petrochemical complex explosion in Chiba prefecture after the Great East Japan Earthquake. It was processed as follows:

- March 11 3pm: The Petrochemical Complex in Chiba was on fire just after the earthquake. This news was true.
- March 11 6pm: The following fake news happened .
  - The Complex was exploded.
  - The atmosphere has radioactive particle and harmful materials from the petrochemical complex. Be careful!
  - Don't go out! Because the rain includes radioactive particles and harmful materials from the petrochemical complex explosion.
- March 12 3:30pm: A news to correct the fake news was officially delivered through the industrys Website and the local government's (Urayasu City) Twitter.

We checked the time series of topics that includes the word "Cosmo Oil" which was the Petrochemical Complex company name from the extracted topics and then evaluated their coherence measure transition.

In Table III (our method results), we can see the following:

- #1 topic that includes the words "fire" without the word "explosion", "harmful material" and so on. It shows the fact that the Petrochemical Complex in Chiba was on fire.
- #2, #3 and #4 topics have words "explosion", "harmful material", "rain" and so on. they represent the fake news that the Complex had exploded and the rain includes radioactive particles and harmful materials.
- #5 and #6 topics are in the same time slot (Mar.12 15:30-16:00). Although #5 topic has words "explosion", "harmful material", "rain" and so on which show the fake news, #6 topic has "Confirmation", "precise", "Chain Mail", "false" and "fake" to corrected the fake news upon referring to the government city name (Urayasu City).

It turned out that our method can extract the fake news and the correct news at the same time and show the topic transition over a period of time from the first to the last. Our method can show the conflict between the fake news and the correct news.

On the other hand, in Table IV (LDA results), we can see the following:

- #1 topic includes words "fire" without the word "explosion", "harmful material" and so on. It showed the fact situation that the Petrochemical Complex in Chiba was on fire as our method did.
- #2 topic with the words "explosion", and #3 topic has "explosion" and "harmful substance". They represented the fake news that the Complex was exploded. Although they didnt have the word "rain" which was the keyword in the fake news, they could find the fake news earlier than our method.

TABLE III
TOP FEATURE WORDS OF TOPICS THAT INCLUDE "COSMO OIL" DERIVED BY OUR METHOD DURING MAR, 11 AND 12

#	Date	Coherence	
1	Mar.11 16:00-16:30	-1821.75	Chiba, fire, dangerous, Ichihara, South, Inland, complex, gas Tank, petrochemical Complex, Factory, Cosmo Oil, Flame, Kesennuma, Rapid, Steel Plant, Chiba Prefecture, Northern part, still, chemistry, black smoke, Chiba city, Yamagata prefecture, refinery, middle street,
2	Mar.11 18:00-18:30	-1856.75	Chiba, Explosion, Ichihara, Ironworks, Complex, Fire, Danger, Cosmo Oil, Factory, Food, Fire, Tank, Explosive Sound, Operation, Tent, Petrochemical Complex, Petroleum, Mito Art Center, Gas Tank, Refinery, Urgent , Fire pillar, devastation, Ichibara shi, Chiba Refinery, Chiba city, Mashi, Makuhari, bright red, Soga,
3	Mar.11 19:30-20:00	-1081.04	evacuation center, neighborhood, Chiba, city, explosion, factory, residence, Shinagawa station, contamination, wearing, chemistry, harmful, chemicals, everyone, fight, influence, rain, blankets, drinks, Complex, Cosmo Oil, Shinagawa Prince Hotel, Google, Light snack, Ichihara, General Affairs Bureau, Tomomi Yokohama, Ochanomizu Station, Studio Penta,
4	Mar.11 23:00-23:30	-1612.18	near, explosion, body, living, Cosmo oil, going out, Chita city, hazardous substances, contact, adhesion, rain coat, copy-paste, reprint, Chiba, memory, oil, wish, active, pretty, dawn, wear, chemistry, tools, rain, cardboard, diffusion, contamination, city,
5	Mar.12 15:30-16:00	-1326.80	Cosmo Oil, Hazardous Substance, Fact, Attachment, Explosion, Totting, Confusion, Melt, Many, Distribution, Presentation, Description, Fire, Information, Identification, Chiba, Text, Report, Denial, Preventive, Impact, Corporation, Human Body, Rain It falls, title, harmless, secondary disaster, fire suppression, important, relationship,
6	Mar.12 15:30-16:00	-1407.63	Confirmation, together, Chiba, precise, fire, living, disaster prevention, chain mail, tank, neighborhood, Chiba Refinery, explosion, harmful substance, Cosmo Oil, Chiba Prefecture, earthquake, scattering, Ichihara, false, Ichihara, Urayasu , Information, mischief, diffusion, request, publicity, energy saving, safe, electricity,

TABLE IV
TOP FEATURE WORDS OF TOPICS THAT INCLUDE "COSMO OIL" DERIVED BY LDA DURING MAR, 11 AND 12

#	Date	Coherence	
1	Mar.11 16:00-16:30	-1562.10	Chiba, Fire, Sendai Airport, Tsunami, Ichihara, Complex, Fire, Gas Tank, Tank, Submerged, Petrochemical Complex, Reliable, Factory, Cosmo Oil, Explosion, Fire, Sendai, Runway, One Wave, Airport, Oil, rooftop, still, video, chemistry, Okushiri Island, Hokkaido Southwest Offshore Earthquake, city,
2	Mar.11 17:00-18:30	-1695.30	Chiba, explosion, fire, Ichihara, ironworks, complex, cosmo oil, factory, fire, earlier, tank, explosion sound, oil industry, oil, direction, gas tank, Odaiba, big site, Urayasu, Ichihara city, refinement Place, fire pillar, liquefaction, urgent, congested, cheap, Chiba Refinery, accident, campus,
3	Mar.11 19:30-20:00	-1600.26	Information, necessary, earthquake information, news, message, going out, Sos, program, character, situation, blog, radio, collection, Fuji TV, body, reporter, manual, talk, reflection, chiba city, harmful substance, Circulation, Cosmo Oil, Hoax, Resident, Exact, Kappa,
4	Mar.11 23:00-23:30	-1196.75	Information, necessary, earthquake information, news, message, going out, Sos, program, character, situation, blog, radio, collection, Fuji TV, body, reporter, manual, talk, reflection, chiba city, harmful substance, Circulation, Cosmo Oil, Hoax, Resident, Exact, Kappa,
5	Mar.12 15:30-16:00	-1838.06	Confirmation, together, precise, resident, disaster prevention, explosion, fire, Chiba, Chain Mail, tank, harmful substance, Chiba Refinery, neighbor, Cosmo Oil, Chiba, scatter, Ichihara, false, earthquake, Urayasu City, Hara City, Transmission, Mischief, fake, Mama, Vietnamese, deep breathing, Nationality, SKYPE, Congestion,,

- #5 and #6 are topics that have words "Hoax" and "fake". They present the news that corrected the fake news. However, unlike our method, LDA could not extract two types of topics such as the fake news and the correct news in the same time slot.

LDA cannot show the conflict between the fake news and the correct news as our method did. We suppose that the author's group that spreads the fake news is different from the author's group that delivers the correct news, and our method can detect that by considering the author's role. Our method is better than LDA for detecting the conflict that is essential for the topic life cycle.

Figure 2 and Figure 3 show the time series transition of topics that have the word "Cosmo Oil". The graph shows a horizontal axis representing time and a vertical axis for the values of coherence. Through Figure 2, we can see the transition of the topic related to "Cosmo Oil". At first, Just after the earthquake (Mar. 11 3pm), the fact topic that the complex was on fire appeared. And then the fake topics about the complex explosion were created around 7pm. The coherence values of fake topics were higher than the fact

topics. This means the fake topics were more popular than the fact. On Mar 12, topics for correcting the fake appeared. Especially, just after the governmental official tweet at 3:31pm, the correct topics' coherence values were high. We can see the conflict between two types of topics such as fake topics and correct topics at the same time, and the progress of the topics were finally converging. Figure 2 is well represented by the life cycle of the topic such as the grows and the conflict related to Cosmo Oil and peoples reaction as well.

On the other hand, in Figure 3, we can not understand the topic transition well. Coherence values of all topics were almost the same and unlike our method, we can not see two types of topics such as fake topics and correct topics at the same time. It does not present the topic life cycle clearly.

Overall, our method can present the time series topic transition on Twitter better than LDA. In addition, our method can control the resolution of community detection according to the magnitude of parameter α mentioned in Section III. By changing the parameter α, we can understand the topic phylogenetic structure in a certain time slot. The graph shows number of extracted topic horizontally and the parameter α that shows the resolution of community detection vertically. If α is small, the number of extracted topics becomes large. There are steady states before the number of extracted topics changes. Through Figure 4, we can understand how topic structures were formed. It is also our methods strong point that shows the topic structure.
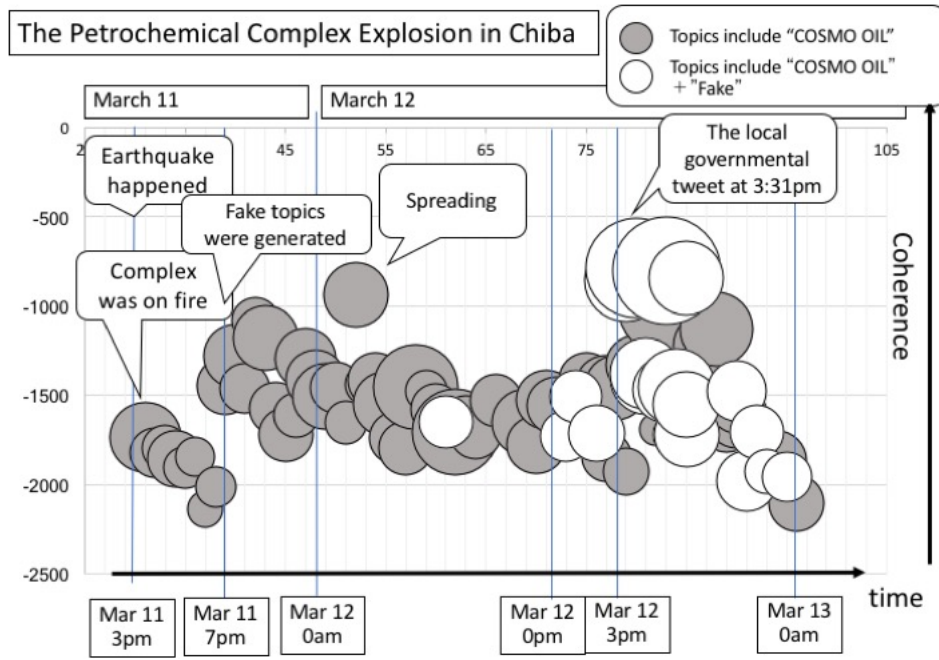
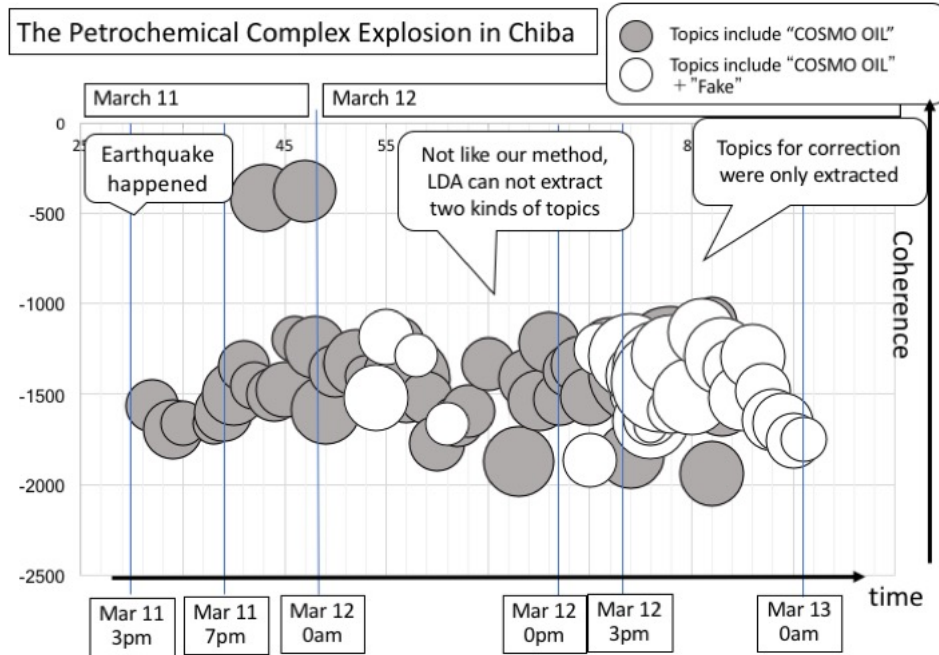Fig. 2. Time Series Topic Transition of the Fake news by Our Method

Fig. 3. Time Series Topic Transition of the Fake news by LDA

V. CONCLUSION

This paper explored the time series topic transition from millions of tweets after the Great East Japan Earthquake by considering the authors role and the semantic accuracy of topics. We used our original method Community Detection Technique in Bipartite Networks that can take into account the bipartite network structure that consists of the authors, words.

Our method can extract quality topics and show the life cycle of topics. To show our methods effectiveness, we compared our experimental results with LDA.

For future works, we plan to improve our time series analysis method on social media adopting Muchas multiples network technique to automatically detect fake topics. In addition, we conduct the experiment with other fake topics that happened after the East Great Japan Earthquake.
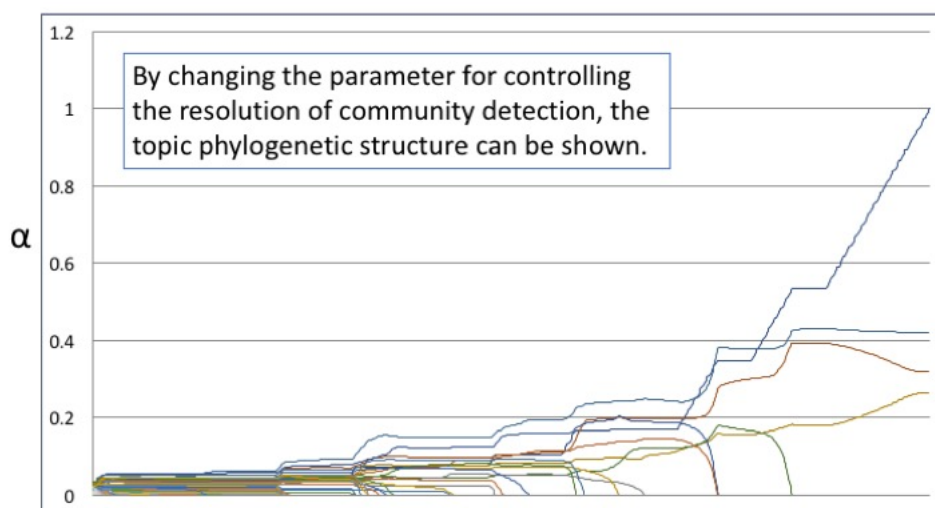
March 12 16:30-17:00

By changing the parameter for controlling the resolution of community detection, the topic phylogenetic structure can be shown.

Fig. 4. Topic Phylogenetic Structure during 16:30-17:00 on March 12

REFERENCES

[1] X. Qiu, A. S. Inagi, S. Nukui, T. Murata and H. Okamoto, *Random Walk based Community Detection from Bipartite Networks*, Proc. of The 30th Annual Conference of the Japanese Society for Artificial Intelligence, 2016.

[2] T. Hashimoto, T. Kuboyama, H. Okamoto and K. Shin, *Topic Extraction from Millions of Tweets based on Community Detection in Bipartite Networks*, Proc. of 27th International Conference on Information Modelling and Knowledge Bases (pp.409-424), 2017.

[3] T. Hashimoto, T. Kuboyama, H. Okamoto and K. Shin, *Topic Extraction on Twitter Considering Author's Role based on Bipartite Networks*, Proc. of Discovery Science 2017 (DS 2017) (pp.239-247), 2017.

[4] D. Mimno, H. M. Wallach, E. Talley, M. Leenders and A. McCallum, *Optimizing semantic coherence in topic models*, Proc. of the Conference on Empirical Methods in Natural Language Processing (pp. 262-272). Association for Computational Linguistics, 2011.

[5] D. M. Blei, A. Y. Ng and M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, (3 (4-5) pp. 993-1022), doi:10.1162/jmlr.2003.3.4-5.993, 2003.

[6] Y. Endo, H. Toda, and Y. Koike, *What's Hot in The Theme: Query Dependent Emerging Topic Extraction from Social Streams*, Proc. of the 24th International Conference on World Wide Web (pp. 31-32). ACM, 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 Forum 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] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan and X. Li, *Comparing Twitter and Traditional Media Using Topic Models*, Proc. of the 33rd European Conference on Information Retrieval(ECIR 2011), LNCS 6611 (pp. 338-349), 2011.

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

[11] E. Ferrara, O. Varol, F. Menczer, and A. Flammini, *Detection of Promoted Social Media Campaigns*, Proc. of ICWSM (pp. 563-566), 2016.

[12] D. J. Berndt, and J. Clifford, *Using dynamic time warping to find patterns in time series*, Proc. of AAAI Workshop on Knowledge Discovery in Databases (pp.359-370), 1994.

[13] Z. Hu, J. Yao, and B. Cui, *User group oriented temporal dynamics exploration*, Proc of AAAI 2014 (pp. 66-72), 2014.

[14] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter and J. P. Onnela, *Community Structure in Time-Dependent, Multiscale, and Multiplex Networks*, science 328.5980 (pp,876-878), 2010.

[15] M. E. J. Newman, *Communities, modules and large-scale structure in networks*, Nature Physics, 8(1) (pp.25-31), 2012.

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

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

[18] http://gibbslda.sourceforge.net/