Revealing Trend Clusters Influenced by a Social Event from COVID-19 Tweets in Japan

Ryosuke Harakawa and Masahiro Iwahashi, Nagaoka University of Technology

The spread of coronavirus disease (COVID-19) has affected human health and economic activity around the world. The analysis of tweets (short text messages on Twitter) would give an insight into COVID-19 and other infection diseases [1], [2]. This study aims to reveal trend clusters influenced by a social event (*intervention*) from Japanese tweets related to COVID-19. In this study, we regard daily changes in the frequencies of each word in tweets as *trends* and define trends with similar wave forms as *trend clusters*.

Interrupted time series analysis (ITSA) [3] is useful for this aim. ITSA can quantify the influence of the intervention at a known point on the increase or decrease of trends. ITSA can be used in the case of the *counterfactual* (trends if the intervention does not exist) cannot be observed.

For previous work with content analysis, Huerta et al. [4] evaluated the public sentiment in Massachusetts during the COVID-19 pandemic. Zhang et al. [5] investigated teens' emotions during the COVID-19 pandemic using Reddit. By combining ITSA with latent Dirichlet allocation (LDA), Kobayashi et al. [6] investigated the evolution of the public opinion on COVID-19 vaccination in Japan. By manually grouping topics extracted by LDA, they summarized many tweets into four themes: personal issue, breaking news, politics, and conspiracy and humor. This made it possible to apply ITSA to big data.

However, if there is another event on a day close to the the intervention, ITSA cannot exclude trends influenced by such an event. If we can prepare a controlled group that is not exposed to the intervention, controlled ITSA [7] is useful. However, this study addresses the challenging task in which we cannot prepare such a controlled group.

This paper presents graphical lasso-guided principal component analysis (GLIPCA) with ITSA, that is, GLIPCA-ITSA. To overcome the above problem, we newly combine ITSA with our previously proposed GLIPCA [2]. GLIPCA-ITSA identifies trends that are likely to be influenced by the intervention and extracts the intrinsic cluster structure from them. Although indirect correlations between trends make it difficult, the sparse structural learning [8] solves the difficulty. Furthermore, we modify the promising network clustering algorithm [9] for extracting trend clusters with different peaks. Even if there is another event on a day close to the intervention, we can judge the trend cluster strongly influenced by the intervention.

For data crawling, the period was set from March 8, 2020, to May 7, 2020. This corresponds to 30 days before and after the state-of-emergency declaration in Japan on April 7, 2020. In the reference[1], words that frequently appeared on the target day but did not frequently appear on other days were defined using Yahoo! News. We selected nouns from them and defined the nouns as important words. Moreover, we set the Japanese corresponding to "$*$ (corona OR pneumonia)" to queries. Here, $*$ is each of the important words. Using Twitter API, we collected up to 500 tweets for each day and each query. Consequently, we collected about $9,459$ tweets on average per day and $576,971$ tweets during the whole period. For each tweet, we extracted the nouns and removed stop words. Furthermore, we removed words whose document frequencies were less than 1% of the total number of tweets from the dataset for each day. The number of the remaining words is denoted by M.

The proposed GLIPCA-ITSA aims to clarify the trend cluster strongly influenced by the intervention without using the controlled group. To this aim, we first calculate a vector $\boldsymbol{f}_i \in \mathbb{R}^N$ ($i = 1, 2, \cdots, M$, $N = 61$) that aligns the daily frequencies of each word. The tth element of $\boldsymbol{f}_i$, that is, $f_i(t)$, represents the frequency of the ith word on the tth day. Here, we normalize $\boldsymbol{f}_i$ by the empirical distribution function and smooth it by a moving average filter whose width is B. Moreover, we apply ITSA to $\boldsymbol{f}_i$ ($i = 1, 2, \cdots, M$) as follows:

$$f_i(t) = \beta_{i,0} + \beta_{i,1}T + \beta_{i,2}d(t) + \beta_{i,3}Td(t) + e_i(t). \tag{1}$$

Here, $d(t)$ is a dummy variable that becomes 0 before the intervention and 1 after it, $e_i(t)$ is the error term, and T is time elapsed since the intervention. Importantly, $\beta_{i,2}$ and $\beta_{i,3}$ represent the level change and slope change by the intervention, respectively. In this study, we utilize $\beta_{i,2}$ as the effect of the intervention because we aim at finding trends directly influenced by the intervention. From the M trends, we select trends in which $\beta_{i,2}$ are more (less) than 0 as trends whose frequencies are amplified (attenuated). Hereafter, the number of the selected trends are denoted by M'.

Next, we divide the obtained M' trends into trend clusters with different peaks. To this aim, we first remove indirect correlations between the obtained trends. Specifically, we use the sparse structural learning [8]. This enables us to obtain the partial correlation matrix $\boldsymbol{L} = (l_{ij})$ ($i = 1, 2, \cdots, M'$, $j = 1, 2, \cdots, M'$). We can consider $\boldsymbol{L}$ as a network whose nodes are trends, and it connects the trends with direct correlations.

Finally, we extract trend clusters with direct correlations to help us judge the trend cluster strongly influenced by the intervention. To this aim, we adopt the promising network clustering algorithm [9]. Because it is equivalent to PCA, we can preserve the reproductivity required for decision-making.

This research is partially supported by JSPS KAKENHI Grant Number JP23H03503 and SECOM Science and Technology Foundation.

[1] http://agora.ex.nii.ac.jp/crisis/covid-19/mass-media/word/

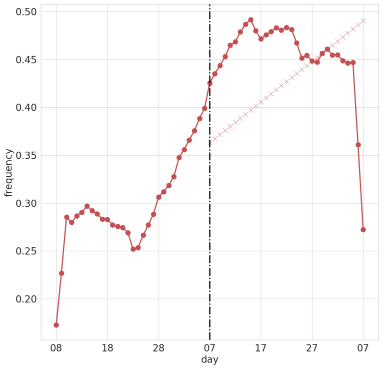

Fig. 1. Results for GLIPCA-ITSA. We set the state-of-the emergency declaration on April 7, 2020 to the intervention (the vertical line). One trend cluster positively influenced by the intervention was revealed. The dot lines represent the estimated counterfactual. The horizontal and vertical axes show days and means of frequencies of trends in each cluster, respectively.

TABLE I
Examples of words (English translations) in the trend cluster revealed by GLIPCA-ITSA shown in Fig. 1.

PCR tests, sales, online, business suspension, the state-of-emergency declaration, favor, benefits, compensation, convergence, sense of smell

However, it generates duplicated members of different trend clusters, decreasing the interpretability. According to [2], we improve the original algorithm to overcome this drawback.

In the experiment, we set $B = 5$ and $\rho = 0.01$ (a sparsity parameter when calculating $\boldsymbol{L}$). We took the state-of-emergency declaration on April 7, 2020 as the intervention. In Fig. 1 and Table I, the trend cluster whose frequencies were amplified by this intervention are shown. We can see that topics such as PCR tests, business suspension, and compensation increased after the state-of-emergency declaration.

In contrast, Fig. 2 and Table II show the trend clusters negatively influenced by this intervention. The first trend cluster suggests that international discussion on the Tokyo Olympics and the pandemic news around the world were attenuated by the domestic news. However, the second trend cluster is not reasonable as topics attenuated by the intervention. Notably, GLIPCA-ITSA easily makes us aware of this fact because the trend cluster has a peak after the intervention. Moreover, if we strengthened the sparsity of $\boldsymbol{L}$, the second trend cluster disappeared (see Fig. 3). This indicates that the trends in the second trend cluster were connected with weak direct correlations. Note that the original ITSA used in previous work [4]–[6], [10], [11] may mislead us. This is because the previous methods cannot divide multiple trends, which may have been influenced by the intervention, into trend clusters with different peaks.

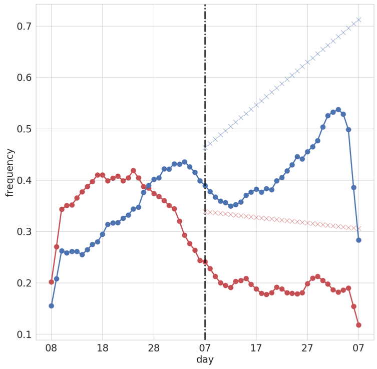

Fig. 2. Results for GLIPCA-ITSA. We set the state-of-the emergency declaration on April 7 to the intervention. Two trend clusters negatively influenced by the intervention were revealed. The first and second trend clusters are denoted by red and blue, respectively. The horizontal and vertical axes show days and means of frequencies of trends in each cluster, respectively.

TABLE II
Examples of words (English translations) in the trend clusters revealed by GLIPCA-ITSA shown in Fig. 2.

The first trend cluster
event, Olympics, ban, limit, trip, overseas, Italy, the United States, IOC, pandemic
The second trend cluster
Avigan, Takashi Okamura, woman, Yamanashi, hospitalization, hospital discharge, extension, comment, family, NHK

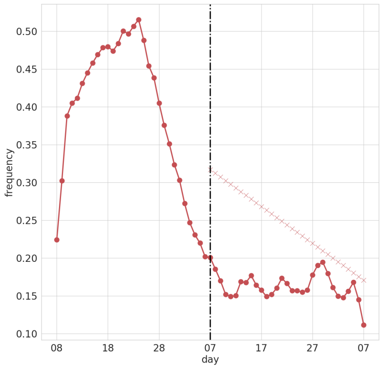

Fig. 3. Results for GLIPCA-ITSA when changing ρ to 0.02 from 0.01 in Fig. 2. The trends in this cluster and those in the first cluster in Fig. 2 were duplicated. The Simpson coefficient was 1, and the former cluster was completely included in the latter one.

References

[1] R. Harakawa and M. Iwahashi, "Ranking of importance measures of tweet communities: Application to keyword extraction from COVID-19 tweets in japan," *IEEE Trans. Computational Social Systems*, vol. 8, no. 4, pp. 1030–1041, 2021.

[2] R. Harakawa, T. Ito, and M. Iwahashi, "Trend clustering from COVID-19 tweets using graphical lasso-guided iterative principal component analysis," *Scientific Reports*, vol. 12, no. 5709, pp. 1–13, 2022.

[3] D. McDowall, R. McCleary, and B. J. Bartos, *Interrupted time series analysis*, Oxford University Press, 2019.

[4] D. T. Huerta et al., "Exploring discussions of health and risk and public sentiment in Massachusetts during COVID-19 pandemic mandate implementation: A Twitter analysis," *SSM - Population Health*, vol. 15, pp. 100851, 2021.

[5] S. Zhang, M. Liu, Y. Li, and J. E. Chung, "Teens' social media engagement during the COVID-19 pandemic: A time series examination of posting and emotion on Reddit," *Int. J. Environmental Research and Public Health*, vol. 18, no. 19, pp. 10079, 2021.

[6] R. Kobayashi et al., "Evolution of public opinion on COVID-19 vaccination in Japan: Large-scale Twitter data analysis," *J. Medical Internet Research*, vol. 24, no. 12, pp. e41928, 2022.

[7] J. L. Bernal, S. Cummins, and A. Gasparrini, "The use of controls in interrupted time series studies of public health interventions," *Int. J. Epidemiology*, vol. 47, no. 6, pp. 2082–2093, 2018.

[8] J. Friedman et al., "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[9] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[10] J. Gu et al., "The impact of Facebook's vaccine misinformation policy on user endorsements of vaccine content: An interrupted time series analysis," *Vaccine*, vol. 40, no. 14, pp. 2209–2214, 2022.

[11] W. Yang, Z. Wu, N. Y. Mok, and X. Ma, "How to save lives with microblogs? Lessons from the usage of Weibo for requests for medical assistance during COVID-19," in *Proc. ACM CHI Conf. Human Factors in Computing Systems*, 2022, pp. 1–18.