Ranking of Importance Measures of Tweet Communities: Application to Keyword Extraction From COVID-19 Tweets in Japan

Ryosuke Harakawa, *Member, IEEE*, and Masahiro Iwahashi, *Senior Member, IEEE*

***Abstract*— This article presents a method that detects tweet communities with similar topics and ranks the communities by *importance measures*. By identifying the tweet communities that have high importance measures, it is possible for users to easily find important information about the coronavirus disease (COVID-19). Specifically, we first construct a community network, whose nodes are tweet communities obtained by applying a community detection method to a tweet network. The community network is constructed based on textual similarities between tweet communities and sizes of tweet communities. Second, we apply algorithms for calculating centrality to the community network. Because the obtained centrality is based on tweet community sizes as well, we call it the importance measure in distinction to conventional centrality. The importance measure can simultaneously evaluate the importance of topics in the entire data set and occupancy (or dominance) of tweet communities in the network structure. We conducted experiments by collecting Japanese tweets about COVID-19 from March 1, 2020 to May 15, 2020. The results show that the proposed method is able to extract keywords that have a high correlation with the number of people infected with COVID-19 in Japan. Because users can browse the keywords from a small number of central tweet communities, quick and easy understanding of important information becomes feasible.**

***Index Terms*— Community detection, coronavirus, coronavirus disease (COVID-19), network analysis, network centrality, semantic understanding.**

I. Introduction

THE outbreak of coronavirus disease (COVID-19) has seriously affected human health and economic activity around the world. During the COVID-19 epidemic, users search social media networks, such as Twitter,[1] Weibo,[2] and YouTube,[3] as well as the traditional media, such as television and radio for information. In particular, Twitter is a very popular social media network [1], [2] and has become an important source of information [3]. Therefore, we use Twitter as the platform for this research. In Twitter, various *tweets* (i.e., short text messages), including information (misinformation, in some cases), have been disseminated widely [4]–[6]. This makes it difficult for users to understand the situation and acquire the relevant knowledge, about COVID-19.

One of the effective solutions to this problem is the visualization of an overview of a large amount of content [7]–[12]. Qian *et al.* [7] explained that it is very time-consuming for users who are not familiar with a topic to browse large amounts of content and quickly gain a general understanding of it; therefore, it is important to automatically mine multiview opinions on the target topic. Our recent study [9] proposed a method for extracting tweet communities[4] from a tweet network that represents the similarity between tweets. In this article, we define a set of tweets with similar topics as a tweet community. The obtained tweet communities enable us to gain a general understanding of many tweets. However, there remains the problem that it is difficult for users to browse all tweet communities as the number of tweet communities increases.

To solve this problem, we propose a method that detects tweet communities, with similar topics, from a tweet network and ranks the communities by *importance measures*. By identifying the tweet communities that have high importance measures, it is possible for users to easily find important information about COVID-19. Inspired by reports that network representation is useful for multimedia content analysis, including clustering [13] and tweet community extraction [9], we also employ a network-based approach. We aim that the importance measure represents the importance of each tweet community. The importance measure should simultaneously evaluate: i) importance of topics in the entire data set and ii) occupancy (or dominance) of tweet communities in the network structure. If the centrality is large but the size is small, the tweet community may be trivial because of oversplitting. On the other hand, it is not guaranteed that the tweet community, whose size is large but centrality is small, includes important topics in the entire data set. As the importance measure, we develop new centrality considering a size of a

Manuscript received June 30, 2020; revised January 5, 2021 and February 22, 2021; accepted March 1, 2021. Date of publication March 17, 2021; date of current version August 2, 2021. This work was supported by the Adaptable and Seamless Technology Transfer Program through Target-Driven Research and Development (A-STEP) from Japan Science and Technology Agency (JST) under Grant JPMJTM20DJ. *(Corresponding author: Ryosuke Harakawa.)*

The authors are with the Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, Nagaoka 940-2188, Japan (e-mail: harakawa@vos.nagaokaut.ac.jp; iwahashi@vos.nagaokaut.ac.jp).

Digital Object Identifier 10.1109/TCSS.2021.3063820

[1]https://twitter.com/

[2]https://www.weibo.com/

[3]https://www.youtube.com/

[4]The terms *community* and *cluster* have the same meaning, in general; however, we use the term community in this article to avoid confusion with the term cluster used in information science and the word cluster that means a group of people infected with COVID-19.

tweet community because the centrality and size satisfy i) and ii), respectively.

Specifically, the core algorithms and the novelty are as follows.

1) We construct a *community network*, whose nodes are the tweet communities. Each node in the community network has a topic, which represents the meaning of the tweets in the corresponding tweet community. In the community network, a node that has edges with high weights includes a central topic in the entire data set, and the node is dominant in the network structure.
2) We calculate the centrality of each node in the community network. Because the centrality is calculated using tweet community sizes as well, we call it the importance measure in distinction to conventional centrality.
3) The novelty of our work lies in its calculation of the importance measure of a community network, rather than a tweet network. This helps us hierarchize the tweet communities to find important information about COVID-19, even as the number of tweet communities increases.

We conducted experiments by collecting 76 000 Japanese tweets about COVID-19 dated between March 1, 2020 and May 15, 2020. In these experiments, we defined words that have a strong correlation with the number of people infected with COVID-19 as keywords. The results show that keywords extracted from only the central tweet communities that were detected by our method were similar to those extracted from all collected tweets. This implies that users can gain a general understanding of many tweets about COVID-19 by browsing only a small number of central tweet communities. This is useful for users to understand the situation and acquire the relevant knowledge about COVID-19 even when there is a flood of information (including misinformation, in some cases).

The rest of this article is organized as follows. Section II describes related work, comprising existing information technologies (in particular, data mining methods) that target COVID-19. In Section III, the proposed method for extracting tweet communities and ranking them by importance measures is explained. Section IV presents the experimental results for real tweets about COVID-19 in Japan and discusses the effectiveness of our method for extracting keywords that are related to the number of infected people. Finally, conclusions and future work are discussed in Section V.

II. Related Work

This section describes information technologies for COVID-19—in particular, pioneering studies that utilize social media networks—to clarify the contribution of our study. Researchers have raised concerns about misinformation, myths, and conspiracies related to COVID-19 [4]–[6], [14], [15]. The term *infodemic* means the phenomenon characterized by a flood of information and misinformation. Cinelli *et al.* [4], Shahi *et al.* [5], and Medford *et al.* [6] reported that COVID-19 has caused an infodemic on social media networks, such as Twitter, Instagram,[5] YouTube, and Reddit.[6] Shahi *et al.* [5] highlighted the necessity of proposing actions for authorities to counter misinformation and hints for social media users on how to stop the spread of misinformation. Singh *et al.* [14] found myths in Twitter by manually defining myths, according to their frequency of appearance in different websites using the search phrase "Coronavirus Common Myths," and defining how dangerous they were. Ferrara [15] reported that accounts that automatically post tweets (namely, bots) are used to promote conspiracy theories in the United States, in stark contrast with human users, who focus on public health and welfare.

Tracking and predicting events about COVID-19 on social media networks have been studied [16]–[19]. Hamzah *et al.* [16] proposed a Web platform called *CoronaTracker*. CoronaTracker provides a predictive model to forecast COVID-19 outbreaks within and outside China, based on daily observations. Furthermore, it can classify the news related to COVID-19 into negative and positive sentiments, to understand the influence of the news on people's behavior, both politically and economically. Zhong *et al.* [17] proposed a susceptible–infected–removed (SIR) model-based method [20] for predicting the number of infected cases in China. Zheng *et al.* [18] also predicted the trend of COVID-19 in China. They combined an improved susceptible–infected model and a long short-term memory (LSTM) network [21] with news information extracted via natural language processing (NLP), to estimate the number of infected cases. Dynamic topic modeling [19] was proposed for analyzing the COVID-19 Twitter narrative among U.S. governors and presidential cabinet members, to track the evolution of subtopics related to risk, testing, and treatment.

Some researchers have constructed COVID-19-related data sets [22], [23]. Chen *et al.* [22] constructed a multilingual Twitter data set for stimulating the research community. In [23], an Arabic data set of tweets on COVID-19 since January 1, 2020 was presented.

Research on inferring or classifying the topics behind Twitter or Weibo posts has been conducted [24], [25]. Wicke and Bolognesi [24] analyzed the discourse around COVID-19 by applying latent Dirichlet allocation [26], a well-known topic modeling method, to a corpus of tweets sent during March and April 2020. Li *et al.* [25] classified Weibo posts about COVID-19, according to seven types of situational information, to find specific features for predicting the reposted amount of each type of information.

In addition, some review articles have been published [27], [28], discussing information technologies, including artificial intelligence [27] and data science [28], for tackling the COVID-19 epidemic.

Our work is the first attempt to clarify topics (in particular, keywords that have a high correlation with the number of people infected with COVID-19 in Japan) on the basis of complex network analysis with NLP. As described in Section I, the technical novelty of our method is that we hierarchize the tweet communities by calculating the importance measures of each tweet community, rather than those of each tweet.

[5]https://www.instagram.com/

[6]https://www.reddit.com/

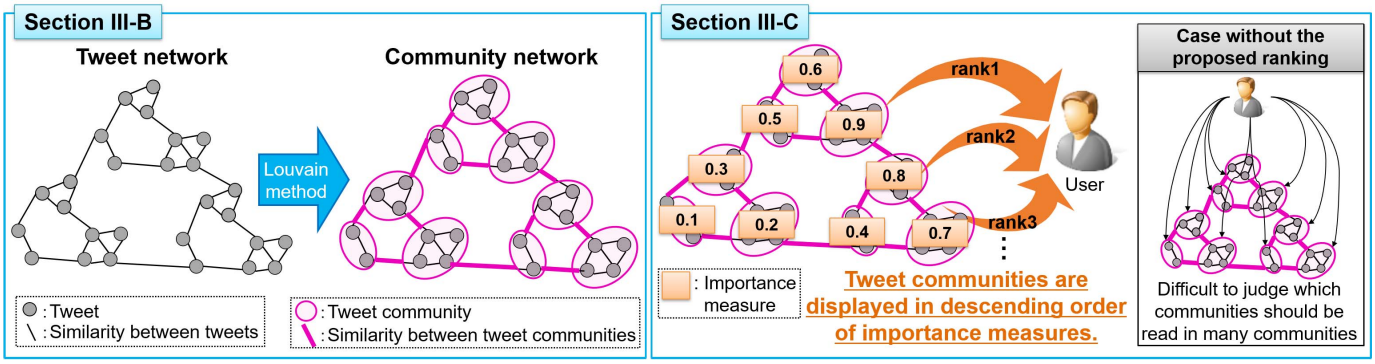


Fig. 1. Overview of Sections III-B and III-C. In Section III-B, two types of networks are constructed. First, we construct a tweet network whose nodes are tweets, which represents similarities between the tweets. Second, we construct a community network whose nodes are tweet communities, which represents similarities between the tweet communities. In Section III-C, importance measures of each tweet community are calculated. We display tweet communities in descending order of the importance measures. This can overcome the difficulty that users cannot judge which communities should be read in many communities.

III. Ranking of Tweet Communities

To gain a general understanding of many tweets about COVID-19, we present a method that detects tweet communities with similar topics and ranks these communities by importance measures. In Section III-A, our method for Twitter data acquisition is described. The proposed method consists of two phases: construction of a community network (Section III-B) and ranking of tweet communities (Section III-C) [see Fig. 1].

A. Data Acquisition

From March 1, 2020 to May 15, 2020, we collected 1000 Japanese tweets per day. In Japan, a state of emergency was declared by the government on April 7. Therefore, people's tension had been increased, especially during the above period. By using the query "a novel coronavirus" (新型コロナウイルス in Japanese), we collected tweets by a keyword search, via an open-source Twitter tool.[7] Moreover, because personal communication is not relevant to the task of tweet community detection, we removed *reply tweets*, as in our previous study [9]. Furthermore, we removed URL strings beginning with an "http" or "pic" prefix. (In Twitter, an attached image is represented as a shortened URL that starts with "pic.") In this way, we constructed 76 data sets for the experiment (one for each day).

B. Construction of Community Network

As in our previous work on tweet community detection [9], we employ a network-based approach. In the experiment presented in Section IV, we performed the subsequent processing on each of the 76 data sets separately.

First, for each data set, we represented each tweet as f_i ($i = 1, 2, \ldots, N$, where N is the number of tweets in one data set). Here, we collected only Japanese tweets and performed the following processing. Using a natural language processing tool called Janome (https://mocobeta.github.io/janome/en/), we performed the morphological analysis and extracted only nouns. Then, we removed stop words defined in https://www.kaggle.com/lazon282/japanese-stop-words. Also, we removed words that consist of only one character because they are likely to be trivial symbols and numbers. Note that Japanese nouns do not change inflection (for example, we do not separate singular nouns from plural ones). Thus, we do not perform lemmatization.

We then extracted textual features $\boldsymbol{t}_i$ that represented the semantics of each tweet f_i. Because there are tweets whose grammar and context are poor, features considering only word frequencies will be more suitable than embedding-based features considering the word order. In fact, the article [29] reports that the term frequency–inverse document frequency (TF–IDF) features [30] have more discriminative power than Doc2Vec [31]. Motivated this fact, we use TF-IDF features as $\boldsymbol{t}_i$.

Following the report that a k-nearest neighbors (k-NN) network is usually suitable for adapting to data set properties [13], we constructed a k-NN network using $\boldsymbol{t}_i$. Specifically, for each tweet f_i, we calculated cosine similarities between $\boldsymbol{t}_i$ (TF–IDF features of f_i) and $\boldsymbol{t}_j$ (TF–IDF features of the other tweets f_j ($j = 1, 2, \ldots, N$, $i \neq j$)). From f_j ($j = 1, 2, \ldots, N$, $i \neq j$), we selected k tweets in descending order of the similarities. By connecting unweighted edges between f_i and the selected f_j, we constructed the k-NN network. The obtained k-NN network represented the relationships between tweet semantics. The k-NN network based on TF–IDF features was also used for tweet community extraction in our recent study [9]. In this article, we define the obtained k-NN network as a tweet network $G = (V, E)$.

Using G, we detect tweet communities with similar topics. Following the reports that the Louvain method [32] works well for multimedia content clustering [9], [11], [33], [34], we apply the Louvain method [32] to G. The Louvain method is based on a quality measure of community detection results called modularity [35]. The modularity Q is defined as

$$Q = \frac{1}{2m} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(w_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij}. \tag{1}$$

Here

$$2m = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$$

[7] https://github.com/Jefferson-Henrique/GetOldTweets-python/

$$k_i = \sum_{j=1}^{N} w_{ij}$$

where δ_{ij} is 1 if f_i and f_j belong to the same tweet community and 0 otherwise. Also, w_{ij} denotes the existence of an edge between f_i and f_j; thus, w_{ij} is 1 if an edge between f_i and f_j in G exists and 0 otherwise. By recursively maximizing Q, we can successfully obtain tweet communities $C_1, C_2, \ldots, C_M$ (where M is the number of communities) containing semantically similar tweets. The details of the algorithm are shown in Algorithm 1.

Finally, we construct a community network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which tweet communities with central topics (where topics are the meanings that represent tweets in the community network) are densely linked to other communities. Each node of $\mathcal{G}$ is one of the obtained tweet communities; therefore, we can write $\mathcal{V} = \{C_1, C_2, \ldots, C_M\}$. The edge weight $e_{ij} \in \mathcal{E}$, from node C_i to node C_j, is defined as follows:

$$e_{ij} = \frac{1}{\log |C_i||C_j|} \sum_{f_m \in C_i, f_n \in C_j} w_{mn} \tag{2}$$

where $|C_i|$ is the number of tweets contained in C_i and $f_m \in C_i$ denotes a tweet in the tweet community C_i. We do not place an edge between C_i and C_j if none of the tweets in C_i and C_j are connected by edges in the tweet network. Equation (2) can simultaneously evaluate: i) importance of topics in the entire data set and ii) occupancy (or dominance) in the network structure. Specifically, the numerator shows i). Although the denominator is a normalization term, the logarithm function reduces the overnormalization when the community sizes are large. Therefore, the denominator shows ii).

Algorithm 1 Detection of Tweet Communities by the Louvain Method [32]

Input: Tweet network G whose nodes are tweets f_i $(i = 1, 2, \ldots, N)$.
Output: Tweet communities $C_1, C_2, \ldots, C_M$.

```
Assign each node f_i (i = 1, 2, ..., N) to each tweet community.
while Improvement of Q (in Eq. (1)) of G is obtained do
  while Improvement of Q of G is obtained do
    /* Local maximization of modularity */
    for each node of G do
      Evaluate the gain of Q when a node is set to each tweet community including neighborhood nodes.
      Reassign a node to the tweet community for which the positive gain of Q is maximum.
    end for
    Calculate Q of G.
  end while
  Update the obtained tweet communities as C_1, C_2, ..., C_M.
  /* Updating a new network */
  Update G with a self-loop whose nodes are the obtained tweet community, where each edge weight is the sum of the edge weights of the original network.
end while
Return the tweet communities C_1, C_2, ..., C_M.
```

C. Ranking of Tweet Communities

Having obtained the community network $\mathcal{G}$, we hierarchize the tweet communities $C_1, C_2, \ldots, C_M$. The input of our algorithm is $\mathcal{G}$, and the output is the result of sorting $C_1, C_2, \ldots, C_M$ in descending order of their importance measures $\rho_1, \rho_2, \ldots, \rho_M$. As described above, performing the importance measure calculation on a community network, rather than a tweet network, is the novelty of this study.

Specifically, we calculate centrality, i.e., degree centrality, closeness centrality, betweenness centrality, and hyperlink-induced topic search (HITS) centrality [36], of $C_1, C_2, \ldots, C_M$. We call the obtained centrality importance measures and denote them by $\rho_1, \rho_2, \ldots, \rho_M$. They are calculated as follows.

Degree Centrality: The degree centrality is the most primitive centrality measure that is defined as

$$\rho_i = \kappa_i$$

where κ_i is a weighted degree of C_i in $\mathcal{G}$. C_i with a high degree centrality is similar to the neighbor nodes.

Closeness Centrality: The closeness centrality is defined as

$$\rho_i = \frac{M - 1}{\sum_{j=1,2,\ldots,M, i \neq j} d(i, j)}$$

where $d(i, j)$ is the shortest path distance from C_i to C_j. Thus, the closeness centrality represents the accessibility of each node in $\mathcal{G}$.

Betweenness Centrality: The betweenness centrality is defined as

$$\rho_i = \sum_{C_s, C_t \in \mathcal{V}} \frac{\sigma(s, t|i)}{\sigma(s, t)}$$

where $\sigma(s, t)$ denotes the number of shortest paths from C_s to C_t and $\sigma(s, t|i)$ denotes the number of such paths that pass through C_i. In this article, we calculate the shortest paths considering edge weights in $\mathcal{G}$. Thus, ρ_i represents the importance of C_i in information propagation in $\mathcal{G}$.

HITS Centrality: The HITS algorithm is equivalent to principal component analysis (PCA) of the network structure [37]. First, we represent $\mathcal{G}$ in the form of an adjacency matrix $\boldsymbol{L} \in \mathbb{R}^{M \times M}$. The elements of $\boldsymbol{L}$ are the edge weights of $\mathcal{G}$. The HITS algorithm calculates the principal eigenvector $\boldsymbol{u} \in \mathbb{R}^M$ of $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$. The ith element of $\boldsymbol{u}$ becomes ρ_i. Note that there is eigenvector centrality as well-known centrality. The eigenvector centrality is equivalent to the principal eigenvector of $\boldsymbol{L}$. In this study, $\mathcal{G}$ is an undirected graph; thus, $\boldsymbol{L}$ is a symmetric matrix. According to the basic linear algebra, eigenvectors of $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ and $\boldsymbol{L}$ are the same. Therefore, in this study, HITS centrality is equivalent to eigenvector centrality.

In $\mathcal{G}$, a node that has edges with high weights includes a central topic (where topics are meanings that represent tweets in the community network) in the entire data set, and it

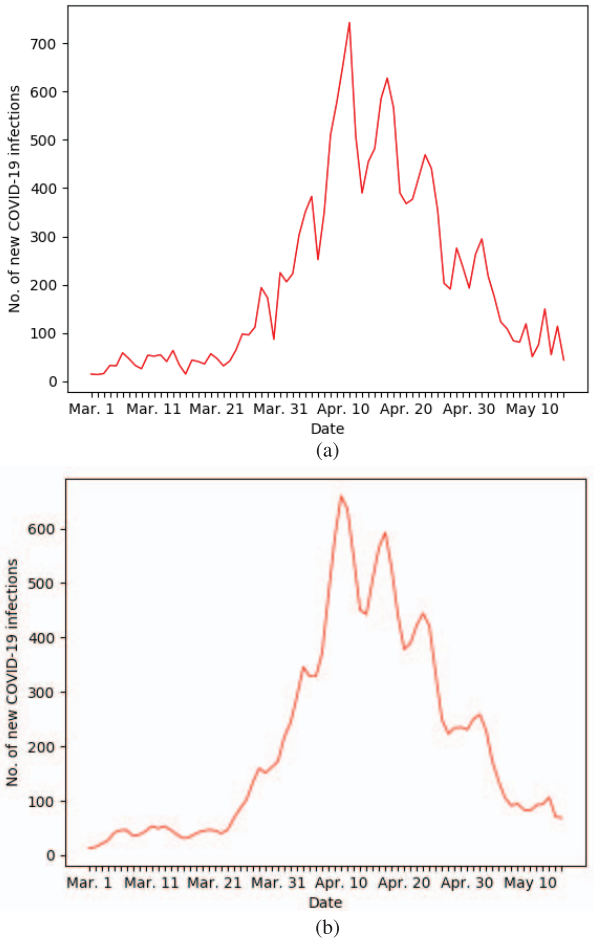


Fig. 2. Number of new COVID-19 infections per day in Japan from March 1, 2020 to May 15, 2020. (a) Raw data. (b) Three-day moving averages.

is dominant in the network structure. Therefore, displaying $C_1, C_2, \ldots, C_M$ in descending order of $\rho_1, \rho_2, \ldots, \rho_M$ enables users to easily find important information about COVID-19, even if many tweet communities are obtained.

IV. Experimental Results and Discussion

In this section, experimental results for real Twitter data are presented and discussed to verify the effectiveness of the proposed method.

A. Quantitative Discussion

We attempt to quantitatively discuss the point that our method enables users to easily find important information about COVID-19 from many tweets. To do this, we evaluate the accuracy of the extraction of keywords about COVID-19, as explained next.

1) Ground Truth: We define the keywords about COVID-19 by focusing on their correlation with the number of infected people. First, we collected the number of new COVID-19 infections per day in Japan from March 1, 2020 to May 15, 2020. The number of new COVID-19 infections is published by Google based on the Wikipedia statistics.[8] There is a case where reports from health centers in every place to the Ministry of Health, Labour and Welfare are delayed because of holidays of the health centers. This results in the fluctuation of the number of new COVID-19 infections depending on the day of the week. To remove the influence of this fluctuation on the subsequent analysis, we calculated three-day moving averages [see Fig. 2].

Here, $\boldsymbol{c}$ denotes the 76-D vector that contained the number of infections (after the moving average) for each day. Second, for each day, we counted the number of times that each word appeared in tweets. If a word appeared multiple times in one tweet, we counted it only once to reduce the influence of tweets in which the same word is repeated many times. Also, we ignored the query words used in data acquisition because they appeared in all tweets. Thus, for each word, we obtained a 76-D vector $\boldsymbol{w}$ that contained the number of times that the word appeared in tweets each day.

Furthermore, we calculated the Pearson correlation coefficient (CC) between $\boldsymbol{c}$ and each $\boldsymbol{w}$. The article [38] reports that $|\mathrm{CC}| > 0.4$ (where $|\mathrm{CC}|$ is the absolute value of CC) shows substantial correlation. In the medicine field, $|\mathrm{CC}| > 0.4$ can be interpreted as "Fair" correlation among "None," "Poor," "Fair," "Moderate," "Very Strong," and "Perfect" correlations. In the psychology field, we can interpret $|\mathrm{CC}| > 0.4$ as "Moderate" correlation among "Zero," "Weak," "Moderate," "Strong," and "Perfect" correlations. In the politics field, $|\mathrm{CC}| > 0.4$ can be interpreted as "Strong" correlation among "None," "Negligible," "Weak," "Moderate," "Strong," "Very Strong," and "Perfect" correlations. According to this report, we defined words with $|\mathrm{CC}| > 0.4$ as the ground truth of the keywords. Hereafter, this set of keywords is denoted by $\mathcal{W}_{\mathrm{GT}}$. We defined the keywords in this way because we considered that words with a high correlation with the number of infections contained semantics relevant to the surrounding situation and necessary knowledge, such as the countermeasures.

2) Comparative Methods: In this experiment, we compared the following ten cases.

Cases 1–4: The cases in which tweet communities are displayed in descending order of the proposed importance measures. Cases 1–4 use degree centrality, closeness centrality, betweenness centrality, and HITS centrality, respectively.

Cases 5–8: The cases in which tweet communities are displayed in descending order of comparative measures. The comparative measures calculate (2) by replacing $\log |C_i||C_j|$ with $|C_i||C_j|$. Thus, these cases only consider the importance of topics in the entire data set. Cases 5–8 use degree centrality, closeness centrality, betweenness centrality, and HITS centrality, respectively.

Case 9: The case in which tweet communities are displayed in descending order of tweet community sizes. Thus, this case only considers the occupancy of tweet communities in the network structure.

Case 10: The case in which tweet communities are displayed in a random order.

For each case, we extracted as keywords the words that appeared in tweets contained in the displayed tweet communities, in the same manner as the extraction of $\mathcal{W}_{\mathrm{GT}}$. Furthermore, we denote the keywords obtained in

[8]https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data

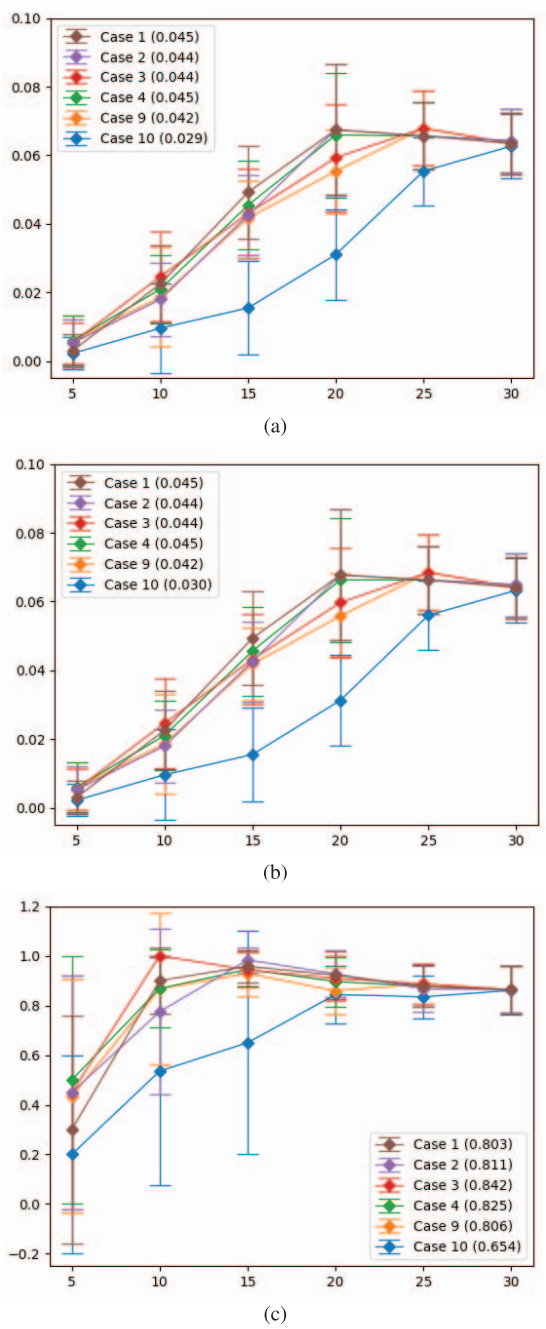


Fig. 3. (a) Jaccard index. (b) Recall. (c) Precision. The horizontal axis shows the number of tweet communities displayed to users. The data points and error bars show the mean and standard deviation of the results of ten validations. The means for each case are shown in parentheses. Seven most frequent keywords in each tweet community are displayed to users. We show results when k was set to 3 in the k-NN network construction in Section III-B.

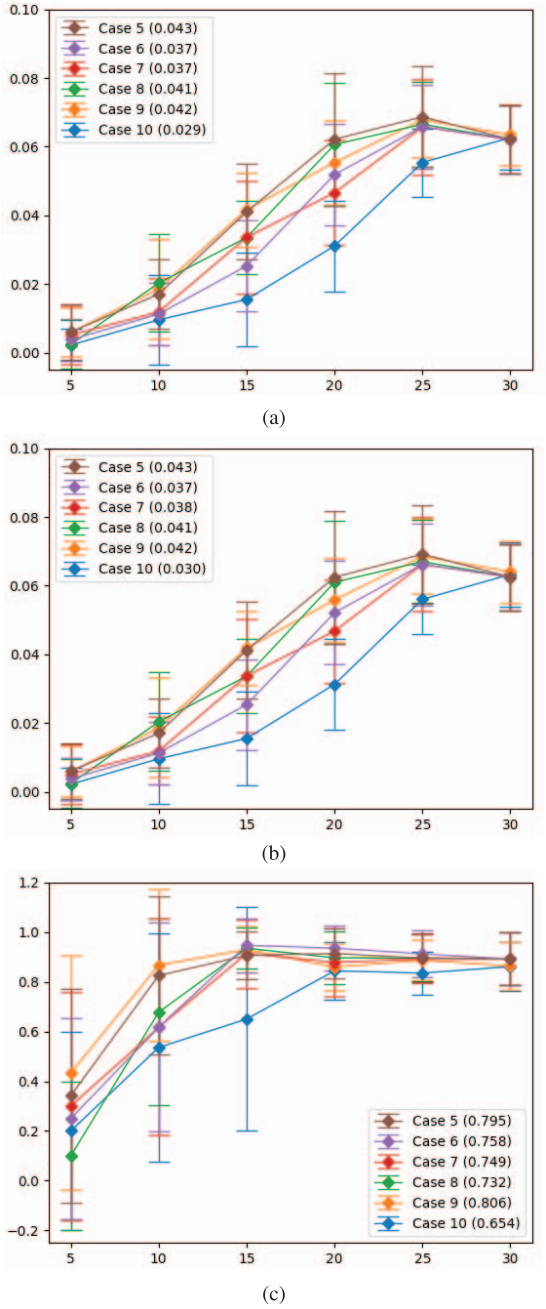


Fig. 4. (a) Jaccard index. (b) Recall. (c) Precision. The notation of these figures is the same as in Fig. 3. Seven most frequent keywords in each tweet community are displayed to users. We show the results when k was set to 3 in the k-NN network construction in Section III-B.

cases 1–10 by $\mathcal{W}_{C1}$, $\mathcal{W}_{C2}$, $\mathcal{W}_{C3}$, $\mathcal{W}_{C4}$, $\mathcal{W}_{C5}$, $\mathcal{W}_{C6}$, $\mathcal{W}_{C7}$, $\mathcal{W}_{C8}$, $\mathcal{W}_{C9}$, and $\mathcal{W}_{C10}$, respectively.

3) Evaluations: For the quantitative discussion, we use the Jaccard index, recall, and precision. The Jaccard index is a frequently used metric that represents the overlap between two sets. The recall represents the comprehensiveness. The precision represents the ratio of correct keywords (i.e., keywords included in the ground truth) to keywords that are displayed to users. They are defined as follows:

$$\text{Jaccard} = \frac{|\mathcal{W}_{\text{GT}} \cap \mathcal{W}_{\text{M}}|}{|\mathcal{W}_{\text{GT}} \cup \mathcal{W}_{\text{M}}|}$$

$$\text{Recall} = \frac{|\mathcal{W}_{\text{GT}} \cap \mathcal{W}_{\text{M}}|}{|\mathcal{W}_{\text{GT}}|}$$

$$\text{Precision} = \frac{|\mathcal{W}_{\text{GT}} \cap \mathcal{W}_{\text{M}}|}{|\mathcal{W}_{\text{M}}|}$$

$$\text{M} \in \{\text{C1}, \text{C2}, \text{C3}, \text{C4}, \text{C5}, \text{C6}, \text{C7}, \text{C8}, \text{C9}, \text{C10}\}.$$

Following the principle of tenfold cross validation, we randomly extracted 900 tweets from each data set, calculated the

TABLE I

EXAMPLES OF KEYWORDS EXTRACTED AS THE GROUND TRUTH ($\mathcal{W}_{GT}$), KEYWORDS IDENTIFIED BY THE PROPOSED METHOD ($\mathcal{W}_{C4}$), AND KEYWORDS FOUND BY RANDOM SELECTION ($\mathcal{W}_{C10}$). THE NUMBER OF DISPLAYED TWEET COMMUNITIES, THE NUMBER OF DISPLAYED KEYWORDS FROM EACH TWEET COMMUNITY, AND THE VALUE OF k ARE 15, 7, AND 3, RESPECTIVELY

Ground truth

English translation of keywords	Keywords in Japanese	CC (correlation coefficient)
business suspension	休業	0.67
official announcement	発令	0.61
hand sanitizer gel	ハンドジェル	0.61
hospital	病院	0.57
emergency	緊急	0.46
declaration	宣言	0.46
state	事態	0.46
China	中国	-0.58
Italy	イタリア	-0.48
cruise	クルーズ	-0.48
princess	プリンセス	-0.44
diamond	ダイヤモンド	-0.41

$\mathcal{W}_{C4}$: The asterisk shows incorrectly detected keywords.

English translation of keywords	Keywords in Japanese	CC (correlation coefficient)
business suspension	休業	0.52
emergency	緊急	0.44
declaration	宣言	0.43
official announcement	発令	0.41
* society	* 社会	* 0.40
influence	影響	-0.40

$\mathcal{W}_{C10}$: The asterisk shows incorrectly detected keywords.

English translation of keywords	Keywords in Japanese	CC (correlation coefficient)
* digital	* デジタル	* 0.44

above metrics for the extracted tweets, and repeated this ten times. By calculating the mean and standard deviation of the ten validations, we attempt to assess the effectiveness of the proposed method accurately.

Next, we show the Jaccard index, recall, and precision for cases 1–10. Note that the size-based method (case 9) displays large tweet communities; therefore, the number of displayed keywords is likely to be larger than other methods. This may make the comparison unfair. Based on the report on the human short-term memory [39], we assume that users memorize only seven frequent keywords in each tweet community. To perform the fair comparison based on this practical assumption, Fig. 3 shows the evaluation results for cases 1–4, 9, and 10. For the k-NN network construction in Section III-B, we should avoid k that erroneously detects tweet communities. A large k is not suitable because it results in a too dense network to reveal the community structure. For this reason, we here set k to 3. From Fig. 3, we can observe that the proposed method (cases 1–4) achieves better results than the size-based method (case 9) and the random method (case 10), for every metric. In particular, the superiority of the proposed method to the random method is statistically significant. For the Jaccard index, the p-values of Welch's t-test when comparing

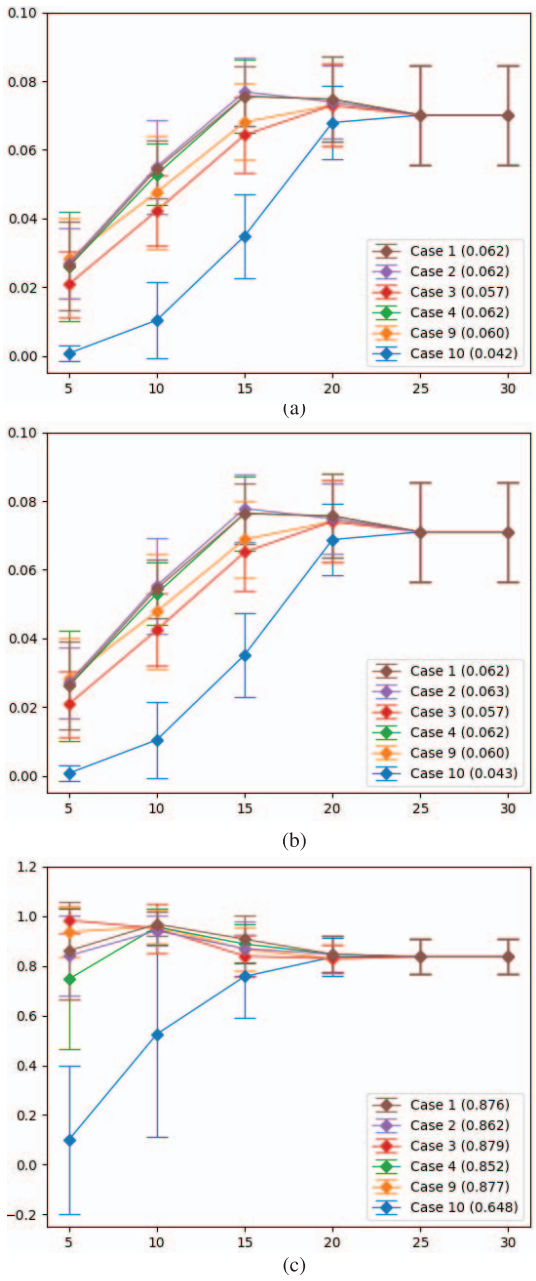


Fig. 5. (a) Jaccard index. (b) Recall. (c) Precision. The notation of these figures is the same as in Fig. 3. Seven most frequent keywords in each tweet community are displayed to users. We show the results when k was set to 6 in the k-NN network construction in Section III-B.

cases 1–4 with case 10 are 0.001, 0.003, 0.002, and 0.002, respectively. For the recall, those are 0.002, 0.004, 0.002, and 0.002, respectively. For the precision, those are 0.024, 0.017, 0.003, and 0.007, respectively. Furthermore, Fig. 4 shows the evaluation results for cases 5–10. The performance of minor versions of the proposed method (cases 5–8) is worse than the proposed method (cases 1–4 in Fig. 3). Although the performance in cases 5–8 is superior to the random method (case 10), the performance is worse than the size-based method (case 9). This shows the necessity of simultaneously evaluating the importance of topics in the entire data set and occupancy of tweet communities in the network structure. Thus, the validity

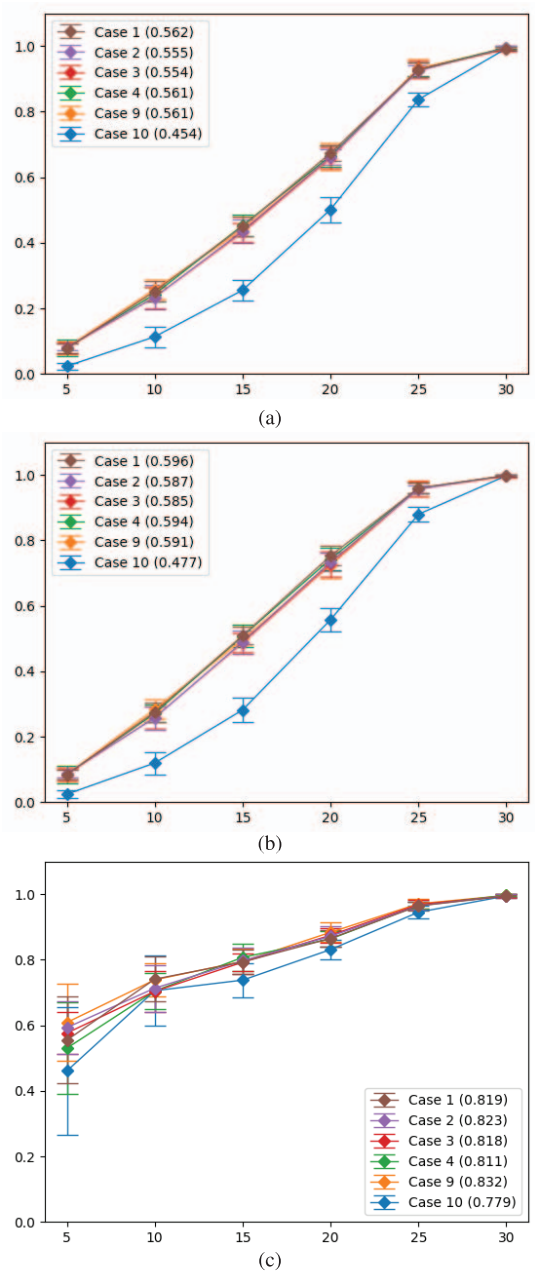


Fig. 6. (a) Jaccard index. (b) Recall. (c) Precision. The notation of these figures is the same as in Fig. 3. All keywords in each tweet community are displayed to users. We show the results when k was set to 3 in the k-NN network construction in Section III-B.

of the proposed method can be confirmed. Also, we can observe that results with degree centrality, closeness centrality, betweenness centrality, and HITS centrality are almost the same. This may be because the global structure and the local structure are similar due to the small size of the community network. If the community network becomes large and dense, HITS centrality would become powerful because it is equivalent to PCA and can exploit the global structure even for the large and dense network.

Table I shows the examples of the keywords extracted as the ground truth ($\mathcal{W}_{\mathrm{GT}}$), those identified by the proposed method ($\mathcal{W}_{\mathrm{C4}}$), and those found by random selection ($\mathcal{W}_{\mathrm{C10}}$).

TABLE II

CORRESPONDENCE BETWEEN THE ENGLISH TRANSLATION AND THE ORIGINAL JAPANESE FOR THE EXTRACTED KEYWORDS IN FIG. 7

Keywords shown in Fig. 7(b)

English translation	Original Japanese
infection	感染
expansion	拡大
cancellation	中止
prevention	防止
schedule	予定
influence	影響
notification	お知らせ

Keywords shown in Fig. 7(c)

English translation	Original Japanese
pneumonia	肺炎
infection	感染
misinformation	デマ
information	情報
countermeasure	対策
welfare	厚生
influence	影響

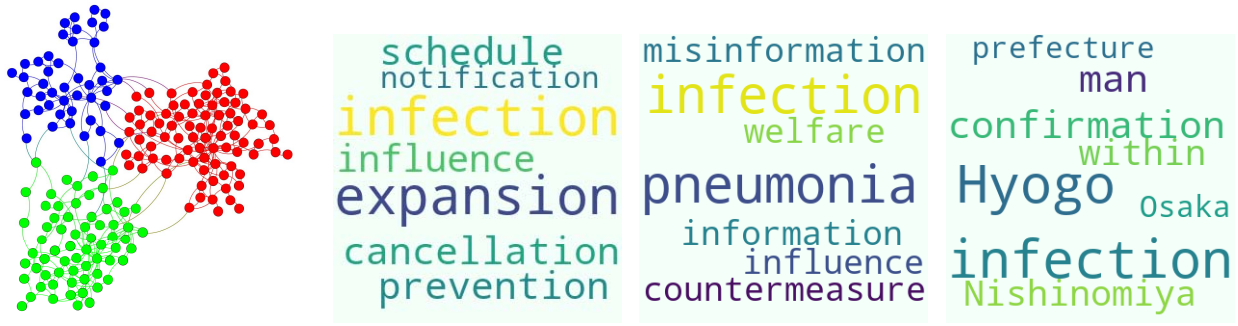
Keywords shown in Fig. 7(d)

English translation	Original Japanese
infection	感染
Hyogo (a prefecture in Japan)	兵庫
confirmation	確認
man	男性
Nishinomiya (a city in Hyogo Prefecture)	西宮
within the prefecture	県内
Osaka	大阪

Our method extracted keywords about the declaration of a state of emergency ("official announcement," "declaration," and "emergency"). As explained above, a state of emergency was declared by the government in Japan in April 7. "Business suspension" may appear because of the request for business suspension by the government, to prevent the spread of COVID-19 infection. Conversely, our method and the random method incorrectly detected "society" and "digital" as keywords. These words seem to be too general to capture COVID-19-related topics. We notice that the random method cannot detect any correct keywords. From this fact, we can confirm the superiority of our method for ranking tweet communities in descending order of the importance measures.

4) Verification Using Another k Value: To test another value of k, Fig. 5 shows the evaluation results where k was set to 6. Even in this setting, we can observe the effectiveness of the proposed method (cases 1–4). In particular, we can confirm the statistical significance of the proposed method to the random method (case 10). For the Jaccard index, the p-values of Welch's t-test when comparing cases 1–4 with case 10 are 0.000, 0.000, 0.005, and 0.000, respectively. For the recall, those are 0.000, 0.000, 0.005, and 0.000, respectively. For the precision, those are 0.000.

In general, it is difficult to find the best k for the topic extraction. To overcome this difficulty, we previously proposed a method [9] that collaboratively integrates community



(a) (b) (c) (d)

Fig. 7. (a) Visualization of three tweet communities, in descending order of importance measures (case 4), on March 1, 2020. The dots represent tweets, and the colors represent the tweet communities to which the tweets belong. Red, blue, and green colors show the tweet communities with the largest, second largest, and third largest importance measures, respectively. (b) Seven most frequent words in the tweet community with the largest importance measure. (c) Seven most frequent words in the tweet community with the second largest importance measure. (d) Seven most frequent words in the tweet community with the third largest importance measure. The correspondence between the original Japanese and the English translation is shown in Table II.

(a) (b) (c) (d)

Fig. 8. (a) Visualization of three tweet communities, in descending order of importance measures (case 4), on April 1, 2020. The notation in (b)–(d) is the same as in Fig. 7. The correspondence between the original Japanese and the English translation is shown in Table III.

(a) (b) (c) (d)

Fig. 9. (a) Visualization of three tweet communities, in descending order of importance measures (case 4), on May 1, 2020. The notation in (b)–(d) is the same as in Fig. 7. The correspondence between the original Japanese and the English translation is shown in Table IV.

detection results by multiple k values. Our future work includes the investigation of suitable k values.

5) Performance Limitation in the Proposed Method: In Sections IV-A3 and IV-A4, evaluations were performed when only seven most frequent keywords in each tweet community were displayed to users. Here, we perform evaluations when all keywords in each tweet community were displayed to users. This condition is not practical because we assume that users take a long time to read many keywords. Therefore, the evaluations here aim at verifying the performance limitation in the proposed method. Fig. 6 shows the evaluation results (where k was set to 3). We can observe that the performance of the proposed method (cases 1–4) is almost the same as that of the size-based method (case 9). As described in Section IV-A3, the size-based method displays more keywords than the proposed method. Thus, the correct keywords are likely to be included in the displayed many keywords. This results in the performance that is comparable with the proposed method. In summary, the proposed method is especially effective in the practical condition where users can read a limited number

TABLE III
CORRESPONDENCE BETWEEN THE ENGLISH TRANSLATION AND THE ORIGINAL JAPANESE FOR THE EXTRACTED KEYWORDS IN FIG. 8

Keywords shown in Fig. 8(b)

English translation	Original Japanese
mask	マスク
infection	感染
countermeasure	対策
distribution	配布
prevention	予防
Abe (name of the Prime Minister in Japan at that time)	安倍
government	政府

Keywords shown in Fig. 8(c)

English translation	Original Japanese
mask	マスク
Abe (name of the Prime Minister in Japan at that time)	安倍
government	政府
infection	感染
household	世帯
Prime Minister	総理
countermeasure	対策

Keywords shown in Fig. 8(d)

English translation	Original Japanese
infection	感染
confirmation	確認
news	ニュース
announcement	発表
within the prefecture	県内
NHK (the abbreviation of Japan Broadcasting Corporation)	NHK
man	男性

TABLE IV
CORRESPONDENCE BETWEEN THE ENGLISH TRANSLATION AND THE ORIGINAL JAPANESE FOR THE EXTRACTED KEYWORDS IN FIG. 9

Keywords shown in Fig. 9(b)

English translation	Original Japanese
infection	感染
deceased	死亡
patient	患者
Tokyo	東京
hospital	病院
occurrence	発生
announcement	発表

Keywords shown in Fig. 9(c)

English translation	Original Japanese
infection	感染
news	ニュース
NHK (the abbreviation of Japan Broadcasting Corporation)	NHK
confirmation	確認
deceased	死亡
Tokyo	東京
Hokkaido	北海道

Keywords shown in Fig. 9(d)

English translation	Original Japanese
infection	感染
information	情報
video	動画
relevance	関連
self-restraint	自粛
publication	公開
countermeasure	対策

of keywords. In the case where users can all keywords, the size-based method is substantially effective as well.

B. Examples of Displayed Tweet Communities

In this section, we show the examples of the tweet communities that are displayed to users. Figs. 7–9 shows three tweet communities, in descending order of importance measures, for March 1, April 1, and May 1, respectively. Here, we show the results by the importance measures with HITS centrality (case 4).

On March 1, Fig. 7(b) shows the news about cancellation of many events that attract large crowds for preventing the spread of infection of COVID-19. Fig. 7(c) shows the concern and warning to misinformation about COVID-19. In Fig. 7(d), the news and concern about the first COVID-19 infection in Nishinomiya City in Hyogo Prefecture appear.

Around April 1, the shortage of masks was a serious concern in Japan. To deal with this situation, Prime Minister Shinzo Abe declared that the government would issue two masks per household [see Fig. 8(b) and (c)]. Fig. 8(d) shows the news and people's concern about infection spread all over the country.

On May 1, Fig. 9(b) and (c) shows the news about deaths due to COVID-19. More specifically, in Fig. 9(b), we can observe the report that those in their sixties or older account for about 90% of the total. The curation of COVID-19-related information of various regions, and video messages from celebrities, was confirmed [see Fig. 9(d)]. This tweet community includes tweets about countermeasures, including opinions from experts for COVID-19. This may show people's wish, after the long period of self-restraint, to avoid the spread of infection of COVID-19.

From these results, we find that people's attention changed over time, from concern about the infection to the countermeasures and the wish for the ending of the spread of COVID-19. As a consequence of this section, we confirmed that our method is useful for understanding the situation, and acquiring the relevant knowledge, through ranking of tweet communities by importance measures.

V. CONCLUSION AND FUTURE WORK

This article presented a method that detects tweet communities with similar topics and ranks the communities by importance measures. By identifying only the communities with high importance measures, it becomes possible for users to easily find important information about COVID-19. Specifically, we construct a community network whose nodes are tweet communities, obtained by applying a community detection method to a tweet network. We then calculate the centrality to the community network as importance measures, to detect the most central tweet communities. We conducted experiments by collecting Japanese tweets about COVID-19 sent between March 1, 2020 and May 15, 2020. The results

show that our method can successfully extract keywords, that is, words that are strongly correlated with the number of people infected with COVID-19. Because users can browse the keywords from a small number of central tweet communities, quick and easy understanding of important information became feasible.

We discuss how to use the proposed method for fighting COVID-19. The proposed method will be beneficial for quick and objective decision-making based on public opinions. A small number of tweet communities detected by our method help individuals and organizations find meaningful keywords like Figs. 7–9. This makes it possible to quickly make decisions without taking a long time to manually search a flood of information. It is also notable that such decision-making is based on objective data. If individuals and organizations read a small number of tweets selected subjectively, they may make wrong decisions contrary to public opinions. Our method helps solve this problem. Moreover, our method will accelerate various data mining research for fighting COVID-19, such as opinion mining and sentiment analysis. In such research, tweets that are irrelevant to COVID-19 may increase computational cost and may cause noisy results. Because our method can extract relevant tweets from many tweets, our method helps overcome these drawbacks.

Furthermore, we focus on misinformation that is unique to specific regions and/or time periods. In fact, misinformation that 5G networks are the cause of COVID-19 was observed in specific regions such as Europe, America, and the Middle East. The proposed method will be used to handle such type of misinformation in Twitter as follows.

1) If misinformation is mentioned in many tweets, the proposed method can visualize it as tweet communities. Thus, a user can browse tweet communities, including misinformation.
2) The user selects a tweet community in which they would like to verify whether misinformation is included or not.
3) The proposed method is applied to tweets in different regions and/or time periods. Then, tweet communities that are similar to the user's selected tweet community are extracted.
4) We display the difference between the extracted tweet communities and the user's selected one. It will be useful to visualize the difference of most frequent words like Figs. 7, 8, and 9. In reference to the visualized difference, the user judges whether misinformation is included or not.

In the future, we will develop this system and evaluate the effectiveness. Note that this system will be useful for only misinformation that is mentioned in many tweets and are unique to specific regions and/or time periods. Thus, future work includes the detection of other type of misinformation.

The scope of this study is within COVID-19. Thus, the future work includes the application of the proposed importance measures to topics other than COVID-19. In our previous study [40], we confirmed that the community network is beneficial for efficient grouping of similar Web videos for retrieval. Specifically, we formulate grouping of similar Web videos as community detection in a network, whose nodes are Web videos and edges are hyperlinks weighted by video similarities. Then, we construct a community network in which each node includes multiple Web videos. By applying the community detection method [41] to the community network, we can efficiently group similar Web videos, while the accuracy of retrieval can be preserved. In this way, the versatility of the community network is confirmed. In the future, we will evaluate the proposed importance measures as well as the community network for other topics.

We believe that this study is one of the pioneering works on data mining for tackling the difficulty caused by COVID-19. However, we will develop more sophisticated methodologies. Future work includes the improvement of tweet community detection by using multimodal features, such as deep-learning-based text features, sentiment features, and visual features of attached images. After this improvement, we will develop a method for predicting the trend of people's attention to COVID-19 over time; this is required because the method proposed in this article does not include a time series modeling scheme.

ACKNOWLEDGMENT

The authors would like to thank Edanz Group (https://en-author-services.edanzgroup.com/) for editing a draft of this article.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. ACM Int. Conf. World Wide Web (WWW)*, 2010, pp. 591–600.
[2] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD 1st SNA-KDD Workshop Web Mining Social Netw. Anal.*, 2007, pp. 56–65.
[3] N. Alnajran, K. Crockett, D. McLean, and A. Latham, "Cluster analysis of Twitter data: A review of algorithms," in *Proc. 9th Int. Conf. Agents Artif. Intell.*, 2017, pp. 1–11.
[4] M. Cinelli *et al.*, "The COVID-19 social media infodemic," 2020, *arXiv:2003.05004*. [Online]. Available: http://arxiv.org/abs/2003.05004
[5] G. Kishore Shahi, A. Dirkson, and T. A. Majchrzak, "An exploratory study of COVID-19 misinformation on Twitter," 2020, *arXiv:2005.05710*. [Online]. Available: http://arxiv.org/abs/2005.05710
[6] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, "An 'infodemic': Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak," *Medrxiv*, vol. 7, Oct. 2020, Art. no. ofaa258, doi: 10.1101/2020.04.03.20052936.
[7] S. Qian, T. Zhang, and C. Xu, "Multi-modal multi-view topic-opinion mining for social event analysis," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 2–11.
[8] Q. Fang, C. Xu, J. Sang, M. S. Hossain, and G. Muhammad, "Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2281–2296, Dec. 2015.
[9] R. Harakawa, S. Takimura, T. Ogawa, M. Haseyama, and M. Iwahashi, "Consensus clustering of tweet networks via semantic and sentiment similarity estimation," *IEEE Access*, vol. 7, pp. 116207–116217, 2019.
[10] R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of Web video groups based on sentiment-aware signed network analysis," *IEEE Access*, vol. 5, pp. 16963–16973, 2017.
[11] R. Harakawa, T. Ogawa, and M. Haseyama, "Tracking topic evolution via salient keyword matching with consideration of semantic broadness for Web video discovery," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 20297–20324, Aug. 2018.
[12] R. Harakawa, T. Ogawa, and M. Haseyama, "A Web video retrieval method using hierarchical structure of Web video groups," *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 17059–17079, Dec. 2016.
[13] I. Pitas, *Graph-Based Social Media Analysis*. London, U.K.: Chapman & Hall, 2015.

[14] L. Singh *et al.*, "A first look at COVID-19 information and misinformation sharing on Twitter," 2020, *arXiv:2003.13907*. [Online]. Available: http://arxiv.org/abs/2003.13907
[15] E. Ferrara, "What types of COVID-19 conspiracies are populated by Twitter bots?" *1st Monday*, vol. 25, no. 6, May 2020, doi: 10.5210/fm.v25i6.10633.
[16] F. B. Hamzah, C. Lau, H. Nazri, D. V. Ligot, and G. Lee, "Coronatracker: World-wide COVID-19 outbreak data analysis and prediction," *Bull. World Health Org.*, vol. 2, pp. 1–31, Dec. 2020.
[17] L. Zhong, L. Mu, J. Li, J. Wang, Z. Yin, and D. Liu, "Early prediction of the 2019 novel coronavirus outbreak in the mainland China based on simple mathematical model," *IEEE Access*, vol. 8, pp. 51761–51769, 2020.
[18] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, and Z. Kang, "Predicting COVID-19 in China using hybrid ai model," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 2891–2904, May 2020.
[19] H. Sha, M. Al Hasan, G. Mohler, and P. J. Brantingham, "Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. Governors and cabinet executives," 2020, *arXiv:2004.11692*. [Online]. Available: http://arxiv.org/abs/2004.11692
[20] T. W. Ng, G. Turinici, and A. Danchin, "A double epidemic model for the SARS propagation," *BMC Infectious Diseases*, vol. 3, no. 1, p. 19, Dec. 2003.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set," *JMIR Public Health Surveill.*, vol. 6, no. 2, May 2020, Art. no. e19273.
[23] S. Alqurashi, A. Alhindi, and E. Alanazi, "Large arabic Twitter dataset on COVID-19," 2020, *arXiv:2004.04315*. [Online]. Available: http://arxiv.org/abs/2004.04315
[24] P. Wicke and M. M. Bolognesi, "Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter," 2020, *arXiv:2004.06986*. [Online]. Available: http://arxiv.org/abs/2004.06986
[25] L. Li *et al.*, "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 556–562, Apr. 2020.
[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[27] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact," *IEEE Access*, vol. 8, pp. 90225–90265, 2020.
[28] S. Latif. (Apr. 2020). *Leveraging Data Science to Combat COVID-19: A Comprehensive Review*. [Online]. Available: https://www.techrxiv.org/articles/Leveraging_Data_Science_To_Combat_COVID-19_A_Comprehensive_Review/12212516
[29] S. Almatarneh, P. Gamallo, and F. J. R. Pena, "CiTIUS-COLE at SemEval-2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 387–390.
[30] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002.
[31] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," 2014, *arXiv:1405.4053*. [Online]. Available: http://arxiv.org/abs/1405.4053
[32] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.
[33] D. Takehara, R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of content groups from different social media platforms using multiple social metadata," *Multimedia Tools Appl.*, vol. 76, no. 19, pp. 20249–20272, Oct. 2017.
[34] R. Harakawa, T. Ogawa, and M. Haseyama, "Accurate and efficient extraction of hierarchical structure of Web communities for Web video retrieval," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 1, pp. 49–59, 2016.
[35] A. Arenas, J. Duch, A. Fernandez, and S. Gomez, "Size reduction of complex networks preserving modularity," *New J. Phys.*, vol. 9, no. 176, pp. 604–632, 2007.
[36] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.
[37] M. Saerens and F. Fouss, "HITS is principal components analysis," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Sep. 2005, pp. 782–785.
[38] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emergency Med.*, vol. 18, no. 3, pp. 91–93, Sep. 2018.
[39] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97, 1956.
[40] R. Harakawa, T. Ogawa, and M. Haseyama, "[Paper] an efficient extraction method of hierarchical structure of Web communities for Web video retrieval," *ITE Trans. Media Technol. Appl.*, vol. 2, no. 3, pp. 287–297, 2014.
[41] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.

Ryosuke Harakawa (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Sapporo, Japan, in 2013, 2015, and 2016, respectively.

He is currently an Assistant Professor with the Department of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology, Nagaoka, Japan. His research interests include multimedia information retrieval and Web mining.

Dr. Harakawa is a member of the ACM, the IEICE, and the Institute of Image Information and Television Engineers (ITE).

Masahiro Iwahashi (Senior Member, IEEE) received the B.Eng., M.Eng., and D.Eng. degrees in electrical engineering from Tokyo Metropolitan University, Tokyo, Japan, in 1988, 1990, and 1996, respectively.

In 1990, he joined Nippon Steel Company Ltd. Since 1993, he has been with the Nagaoka University of Technology, Nagaoka, Japan, where he is currently a Professor with the Department of Electrical, Electronics and Information Engineering. His research interests include the areas of digital signal processing, multirate systems, and image compression.

Dr. Iwahashi is a Senior Member of the IEICE and a member of the Asia Pacific Signal and Information Processing Association (APSIPA) and the Institute of Image Information and Television Engineers (ITE).