J Vis (2023) 26:1405–1419
https://doi.org/10.1007/s12650-023-00936-0

REGULAR PAPER

Tiandong Xiao · Naoya Oda · Yosuke Onoue

Visualization of topic transitions in SNSs through document embedding and dimensionality reduction

Received: 1 September 2021 / Revised: 25 September 2022 / Accepted: 22 June 2023 / Published online: 27 July 2023

Abstract Social networking services (SNSs) have become the primary means by which individuals express themselves. Consequently, the thoughts of individuals could be explored by analyzing primary topics on SNSs. In this study, we proposed and developed a novel system for visual analytics to address the following intriguing questions. When do topics change? Do they ever resurface? What do people typically discuss? Using document embedding and dimensionality reduction approaches, we abstracted dynamic topics as several points in a two-dimensional space. In addition, we provided other charts depicting words that appeared at certain moments and their time series dynamics over entire periods. In addition, we created a novel text visualization technique called semantic-preserving word bubbles to visualize words at a specific time. In addition, we demonstrated the efficacy of the proposed system utilizing Twitter data regarding early COVID-19 trends, Fukushima nuclear disaster trends, and user ratings of system usability. In general, we have presented this novel system to aid users in exploring and comprehending the transitions of contents uploaded on SNSs.

Keywords Visual analytics · Text visualization · Time-series visualization · Social networking service · Topic visualization · Document embedding · Topic-transition analysis

1 Introduction

These days, social networking services (SNSs) such as Twitter and Facebook have become indispensable tools in our daily lives. They allow users to develop and exchange information at any time and location. Given this context, the collection and analysis of information shared by millions of SNS users have become both a challenge and a tremendous potential for addressing certain challenges, such as global disasters or pandemics.

Visual analytics of topical transitions is a popular direction of research on SNSs (Cao et al. 2012; Viégas et al. 2013; Zhao et al. 2014; Wu et al. 2014). Although some studies have examined topic transitions over time, they often portray time on a unidirectional axis, making it difficult for users to uncover complicated topic transitions, such as the resurfacing of a topic or the identification of related topics. Specifically, in

T. Xiao (✉) · N. Oda · Y. Onoue
Nihon University, Tokyo, Japan
E-mail: victorxiaotiandong@gmail.com

N. Oda
E-mail: oden6680@gmail.com

Y. Onoue
E-mail: onoue.yousuke@nihon-u.ac.jp

response to unforeseen complex events, such as the Fukushima nuclear disaster in Japan in 2011 and the COVID-19 pandemic in 2020, people's responses on SNSs changed drastically and significantly. Phenomena relating to a wide variety of occurrences, such as the above-mentioned topics, have exhibited intriguing patterns of people's interest in SNSs, especially when related topics resurface several times during distinct periods (e.g., multiple declarations of a state of emergency during the COVID-19 pandemic). However, these topical transitions are challenging to discover and analyze when using a visualization method that displays time along a unidirectional axis. Understanding such complex topical transitions can play a crucial role in the PR decision-making process of governments and businesses in an emergency.

Several studies in visual analytics enable the study of time series data through the dimensionality reduction projection process (Natsukawa et al. 2020,Van Den Elzen et al. 2016). It is known that snapshot-to-point approaches are effective for comprehending the transitioning of dynamic system states. However, these existing studies focus on time series numerical data and networks and have not been applied to text data. We, therefore, extend the snapshot-to-point methods such that they could be applied to text data for visualizing the dynamics of topics on SNSs.

This study proposes a novel visual analytics system for determining complex topic transitions on SNSs. We define a topic as the content of an entire SNS post at a given moment. As illustrated below, the proposed system has two primary characteristics. (1) Representing a topic at a certain period of time as a point in a two-dimensional (2D) space using the topic projection view. (2) Summarizing the sentences in a sequence of topics using the keyphrase view.

Under the topic projection view, sentences posted on SNSs within short timeframes were summarized, then vectorized using a document embedding technique. Using the dimensionality reduction method, a high-dimensional vector set was then plotted in a 2D space for visualization. In addition, to find and analyze topics with comparable categories, a group of topics (referred to as a topic group) that are significantly close were extracted using a clustering method, and a group of words from those topics were displayed in the keyphrase view. In the keyphrase view, the frequency of occurrence of certain words in the topics was displayed as word clouds. In this paper, we propose a novel semantic-preserving word bubble (SPWB) technique to address several challenges associated with word clouds. In the SPWB method, words are represented as circles and placed on the canvas using word vectors, dimensionality reduction, and force simulation while preserving their semantics.

In this study, we demonstrate the efficiency of the proposed system using two application examples and user evaluation. We demonstrated that our system can identify complex transitions between topics in the application examples. User evaluation demonstrated that our system possesses the requisite usability. We utilized Twitter data regarding COVID-19 and the Fukushima nuclear disaster for the application samples. The COVID-19 topic data were collected between January and April of 2020 from posts in Japan, whereas the Fukushima nuclear disaster data were collected between February 28, 2011, and March 31, 2013. Using these application examples, our proposed system successfully visualized topical transition patterns, including the topic's content and time series dynamics.

The following are the primary contributions of this work. (1) We extended the snapshot-to-point method to employ text data and visualize topic transitions in SNSs. (2) We present SPWB, a novel text visualization technique. (3) Using our proposed system to visualize real-world Twitter data, we demonstrated that complex topical transitions do, in fact, occur on SNSs.

2 Related work

This section examines prior research on the visual analytics of social media data. Due to the magnitude and complexity of social media datasets, researchers have started utilizing visualization tools to enhance the quality of data exploration and analysis. The temporal aspect of data is one of the most important parts of social media datasets. Herein, we summarized visual analytics studies focusing on the distribution patterns and topical transitions of social media topics.

2.1 Visual analytics of spread patterns

Researchers have investigated the relationship between information dissemination and users who repost such content for several years. Whisper (Cao et al. 2012), which represented and analyzed information dissemination on spatial and temporal scales, is one of the earliest research in this area. Whisper was created

to visualize the distribution pattern of SNS users and to identify the primary distributor and popular areas of trending topics. Due to Whisper's emphasis on people, it is challenging for users to simultaneously examine changes in multiple topics using the main system view.

2.2 Visual analytics of topic transitions

Social media can rapidly reflect real-world events. Therefore, the analytical system should identify pertinent events posted on social media. The visual analytics of topical transitions emphasize sequential trends, dynamic patterns, and relationships between event topics. FluxFlow (Zhao et al. 2014) investigated the magnitude and period of the explosion of anomalous information that began to disseminate, whereas OpinionFlow (Wu et al. 2014) aimed at visualizing opinion distribution on SNSs. These studies with sentiment analysis have allowed for exploring the dynamic characteristics of emotional opinions. Meanwhile, EvoRiver (Sun et al. 2014) focused on visualizing the interdependence and competition among SNS topics. EvoRiver demonstrated the relationship of the accounts that lead to topical transitions. Using the horizontal axis of the visualization system to represent time, as in the aforementioned study, can be challenging for users to identify relatively comparable topics in social media data, particularly when they occur across various periods.

2.3 Time-series visualization using state space

In the previous research, wherein time was represented in state space and not as the horizontal axis of the visualization system, to the best of our knowledge, time, as a continuous variable, was expressed as points on a timeline. Time curves (Bach et al. 2016) and the research by (Van Den Elzen et al. 2016) and (Natsukawa et al. 2020) are examples of these types of topical visualization in which users can discover the relationships between the various time frames of data by discretizing continuous variables, such as time, as points. Elzen et al. proposed the snapshot-to-point methodology. This approach is a multistep process developed to generate original real-world data (in their instance, network data) as 2D points. The data must first be modeled as a dynamic model and discretized if necessary. Second, the data are vectorized and normalized into high-dimensional vectors. The third step is to use dimensionality reduction to obtain 2D points. Finally, the 2D points can be visualized. In addition, we discovered that a visualization interface includes data features that might aid users in comprehending and analyzing the state of SNS data. In our investigation, we discovered that the aforementioned studies could be cited because the visualization research for SNS data exhibits a high correlation between text content and publication time. Based on this understanding, we developed an approach to handle the features of SNS data and discovered that it improves the performance outcomes of SNS visualization studies.

2.4 Word clouds

Word clouds, also known as tag clouds, are a visual representation of text data. Word clouds can encode the significance of words through size or color and arrange them on 2D canvases. Wordle (Viegas et al. 2009) is a well-known word cloud system that generates compact layouts utilizing the greedy algorithm. Additionally, it can arrange words in various orientations. This technique does not, however, capture the relationships between words, such as semantic and temporal relationships. (Adä et al. 2010) and (Wu et al. 2011) proposed methods for placing comparable phrases in close proximity. Nonetheless, (Barth et al. 2014) demonstrated that creating semantically related word clouds is a nondeterministic polynomial-time hardness problem. Furthermore, a comparison between Wordle and the carving method (Wu et al. 2011) revealed that semantic word clouds are not as compact as those generated by Wordle. (Cui et al. 2010) proposed a system for demonstrating the relationship between temporal changes by providing a significant trend chart and generating multiple word tags representing different time frames. However, to investigate time-varying data, users must compare the differences between various word clouds, which might be perplexing. (Binucci et al. 2016) developed a method for constructing fully dynamic semantic word clouds, where all the necessary data for word clouds are not known in advance. The dynamic transitions in word clouds were depicted through animation so that users could easily comprehend the continuous evolution of the changes. However, it is difficult for users to comprehend the changes in time independently (e.g., to find a similar topic at different periods). Generally, supporting the representation of words at different times on the same chart while maintaining semantically and positionally correlated relationships remains a challenge.

3 Design requirement

On SNS platforms, people's interests change constantly and are beneficial to multiple users. It is necessary to have a system that can discover and analyze the topics of social media posts. We considered social scientists, government employees, students, public relations managers of corporations, etc., as our prospective users who may be interested in SNS subjects, topic-changing patterns, people's responses to announcements, etc. To address the requirements mentioned above of the users, we extracted our design objectives by examining prior works on visualization studies for SNSs, focused on the recurrence of similar topics, and set the following design requirements.

G1. Discover the differences between topics in various backgrounds and at various discrete stages The system should enable researchers to examine numerous topics over time and identify parallels and differences in their respective contents.

G2. Discover the topics that coappear Two topics often appearing at the same time may indicate a relationship between them or that people tend to worry about both topics simultaneously. The system should be able to easily discover topic groups, as mentioned above.

G3. Discover people's reactions to a piece of news (1) The time is between when a piece of news is published and when people react to it on SNSs; (2) the amount of time individuals spend discussing a piece of news. People react to news and share their opinions on SNSs. Such reactions should be summarized by the system for analysts.

G4. Visualize the changes in the popularity of certain topic discussions over time In addition to the general popularity of a topic, users are curious about its popularity at a particular period. The system should be able to visually represent the dynamics in the popularity of various topical discussions over time.

G5. Interactive visual exploration This is particularly critical due to the complexity of SNS data, which includes various crucial information, such as publication time and content details. In addition, when analyzing and visualizing SNS data, large data sizes are regularly encountered. The visualization system must be dynamic and allow users to simply compare and discover intriguing topical transitions or relatively invariant SNS posts. Therefore, the system should have rich interactions, such as summarizing, focusing, and filtering, with content visualization to aid researchers in gaining insight into the change patterns of topics and their details.

4 System overview

Based on our design requirements, we present a topic projection view for modeling complicated topic transitions using the document embedding technique, Doc2Vec (Le and Mikolov 2014). A similar approach was presented for visualizing dynamic networks (Van Den Elzen et al. 2016). Using network vectorization and dimensionality reduction at each point in time, this method visualizes the time evolution of networks. We used two extra views, namely the keyphrase and timeline views, to demonstrate topical details. Figure 1 provides an overview of our proposed visual analytics system utilizing these views. The interface comprises three parts: (1) the topic projection view on the left displays topical transitions by time, (2) the keyphrase view on the right reveals topic details at specific times, and (3) the timeline view at the bottom of the image shows the time series change in the number of SNS posts and word occurrences. The proposed system is a single-page web application written in JavaScript utilizing React (https://reactjs.org/) and D3.js (Bostock et al. 2011).

4.1 Data preprocessing

First, we present our proposed system's data processing procedure. The minimum data requirement of our system is a group of short-text data together with their publication timestamps. We obtained both the topic and word vectors and visualized them using Doc2Vec, a neural network-based approach for document embedding. Doc2Vec encodes documents and words into high-dimensional vectors (generally hundreds of dimensions). For topic visualization, we employed the snapshot-to-point method used for network data by Elzen et al., where our data are short-text-based.

As demonstrated in Fig. 2, we defined a group as a collection of k SNS posts and a topic document as a collection of multiple groups. We used the document vectors obtained from Doc2Vec as the topic vectorization by assigning the same topic document index to multiple SNS posts within a given period. Similar

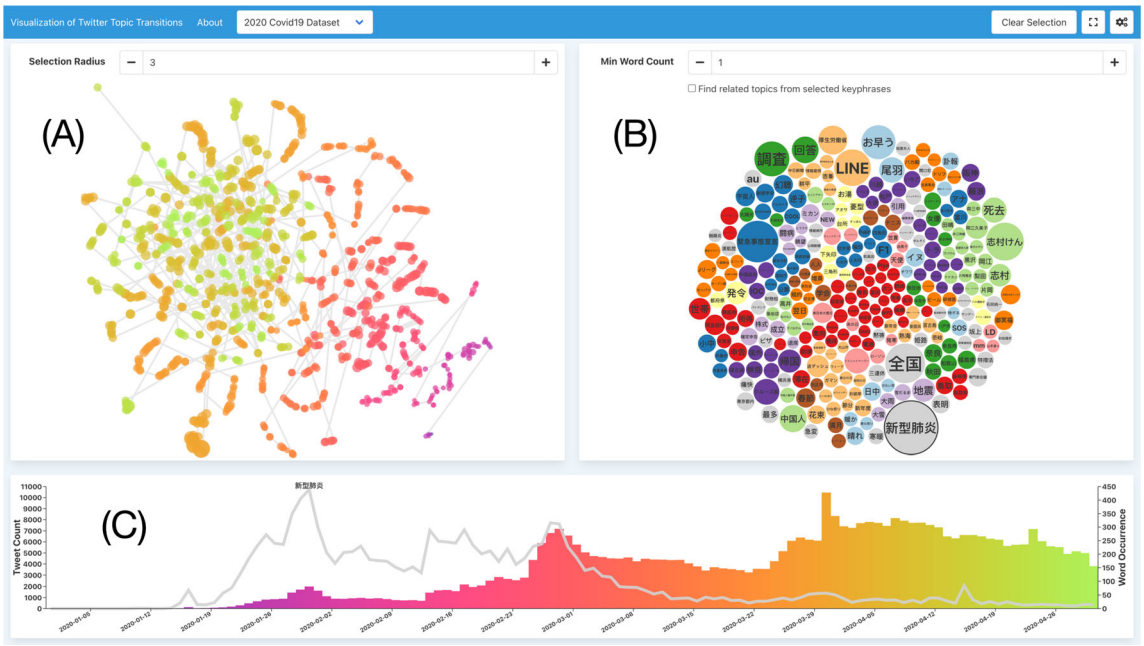

Fig. 1 Overview of the proposed system. The projection view **A** displays the changing pattern of topics. The keyphrase view **B** displays the details of the topics. The timeline view **C** displays the daily numbers of SNS posts and words' occurrences

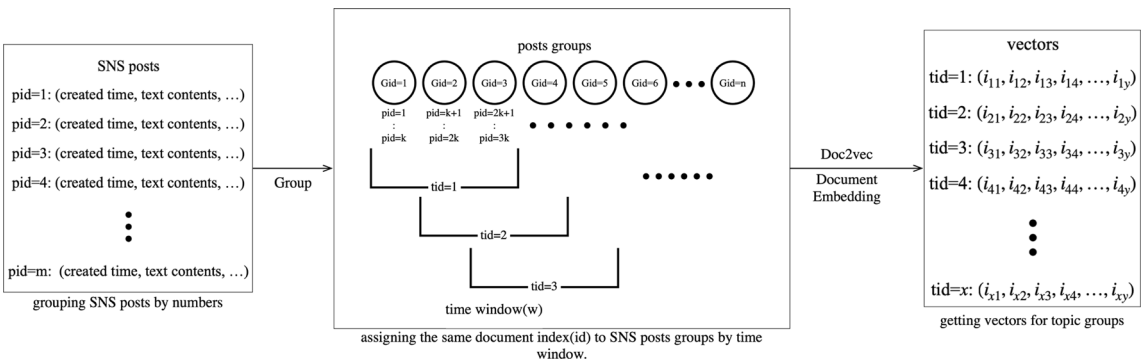

Fig. 2 Vectorization of topics from SNS posts. *m* Posts are combined with *n* groups, after which a total *x* topic document index (TID) is assigned to them using the time window(w). Doc2vec is used to compute high-dimensional vectors of each document

to the snapshot-to-point method, consecutive SNS posts were grouped while using specific intervals (window sizes) to analyze the topical transitions. Two parameters must be specified for the algorithm. The first parameter is the group size(k), the number of SNS posts in a group, and the second one is the window size(w), the number of groups assigned to the same topic document index. When the window size is w, the same index is assigned to the w groups before and after a group, i.e., the same index is assigned to the $2w + 1$ groups. The parameter values such as the group and window sizes are determined by the temporal length and size of the data.

The values of the two parameters employed in our approach depend on the details of the SNS data. The group size influences the number of resultant topic points and depends on the number of SNS posts. The window size primarily affects the accuracy, and difficulty in determining whether a set of topic points are similar. The values of these two parameters are chosen empirically based on the characteristics of each SNS dataset.

In the keyphrase view, keyphrases are retrieved from SNS posts and trained with Doc2Vec to identify the words that best characterize the topics. In the following two steps, we define the keyphrases (Fig. 3). Initially, keyphrases are extracted from each group of SNS posts using term frequency-inverse document frequency(TF-IDF) (Salton and Buckley 1988), a standard method for determining the significance of words

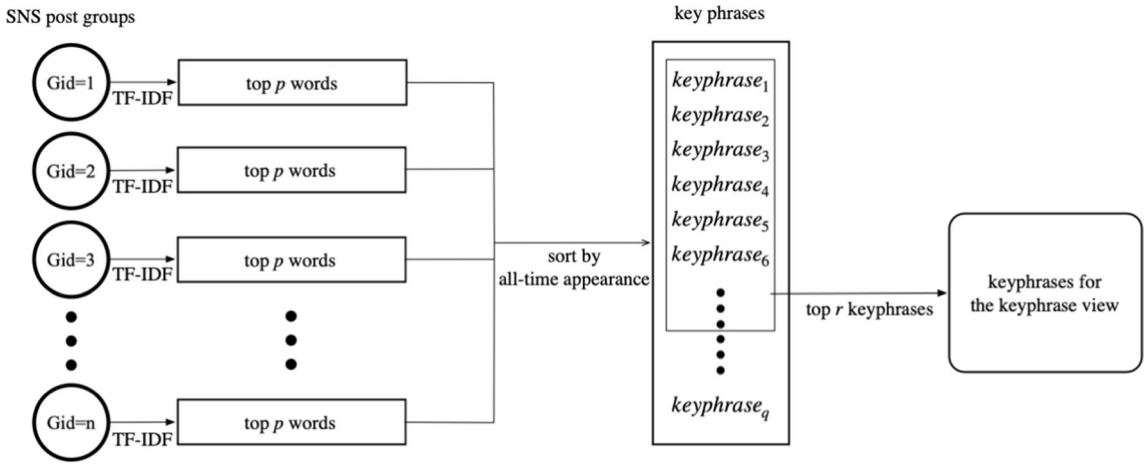

Fig. 3 Keyphrases selected from SNS posts. p keyphrases are obtained for each post group using TF-IDF, and then, the most r keyphrases of the all-time appearance are selected from all keyphrases for realizing the keyphrase view

in documents. Next, the number of occurrences of each group's keyphrases is tallied. Frequent keyphrases are extracted as the keyphrases of the entire topic.

4.2 Topic projection view

The topic projection view was created to visualize transitions between topics. As illustrated in Fig. 4, topics, i.e., the groups of successive SNS posts, are represented by circles whose coordinates reflect their relationships with other topics. The successive circles representing topics and connected by a line indicating the timeline of the transitional patterns. As time passes, the color of the circles changes according to the diverging color scheme. The line connects continuous circles, and the color coding allows users to distinguish topics that are related by content but separated by time (i.e., similar topics but temporally separated) or vice versa. The size of the circles represents the number of SNS posts per hour, which is determined based on the number of posts between the first and last posts in that topic group.

Utilizing dimension reduction techniques, the 2D coordinates of high-dimensional topic vectors were determined. Several dimensionality reduction techniques can be classified as linear (e.g., principal component analysis (Wold et al. 1987)) and nonlinear (e.g., multidimensional scaling (Kruskal 1978) and t-distributed stochastic neighbor embedding (t-SNE) (van der and Hinton 2008)). Due to the requirement of our system to reduce the density of topic points while keeping the local data structure, we chose the nonlinear dimensionality algorithm t-SNE. We chose t-SNE to project high-dimensional topic vectors into

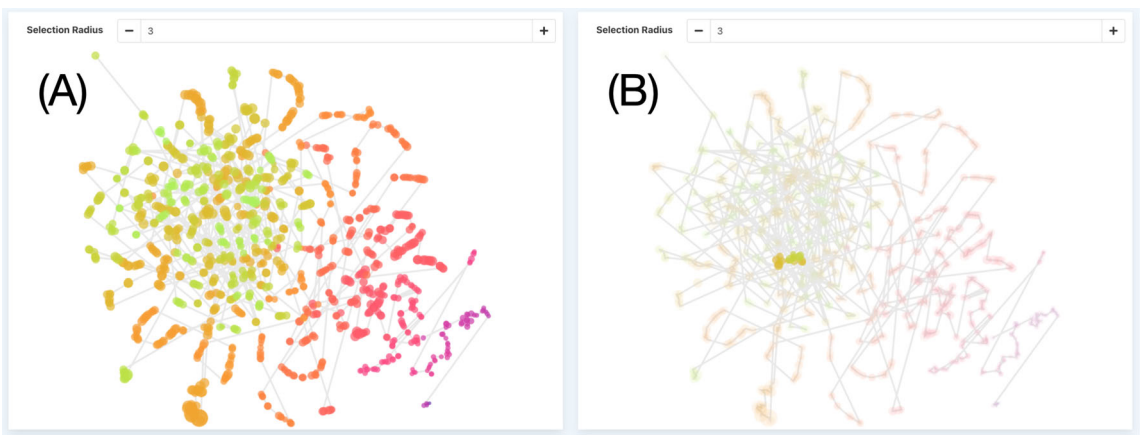

Fig. 4 Default topic projection view (**A**) and topic clustered visualization result (**B**). The topic projection view shows the topics as circles

two dimensions for the three reasons listed below. (1) t-SNE offers some advantages in terms of maintaining the spacing between high-dimensional points. (2) The projected coordinates are appropriate for displaying subject clusters and anomalies. (3) The integration of t-SNE and DBSCAN has good properties.

4.3 Keyphrase view

We built the keyphrase view (Fig. 5), which is situated on the right side of the topic projection view, to allow users to explore the characteristics of topics using the extracted keyphrases discussed in Section 3.1. This view utilizes the novel SPWB text visualization technique. Compared with general word clouds, SPWB can visualize the numerical values assigned to a word using a circle size while ordering words with comparable contexts or characteristics (e.g., the name of a person) to be close together. The keyphrase text is displayed at the center of a circle representing a keyphrase in the SPWB (called word bubbles). The size of the bubbles corresponds to the total number of times the words corresponding to the keyphrase recurred. This is advantageous because the size of an element (i.e., the number of words) can be read more accurately than in conventional word clouds, which express the quantity by font size. In addition, we color coded the groupings of keyphrases to assist users in locating keyphrases with comparable contexts.

The positions of the bubbles in SPWB are determined using the word vectors of the keyphrases defined in Section 3.1. Initially, t-SNE is used to project word vectors into two dimensions, similar to how topic vectors are described in Section 3.2. However, the projected coordinates are typically sparse and frequently result in many overlaps. This affects the space efficiency of drawing results and makes them difficult to read, thus we performed a force simulation inspired by force-directed graph drawing (Fruchterman and Reingold 1991) using the coordinates of the dimensionality reduction results as the initial placement of bubbles. We employed d3-force, which offers an algorithm for simulating physical forces on bubble particles based on velocity Verlet numerical integration (Verlet 1967). We employed three types of forces: (1) a spring force that induces attraction between linked bubbles, (2) a collision force that repels bubbles so that they do not overlap, and (3) a centering force that attracts all bubbles to the center. Concerning the spring force, bubbles that are positioned within a particular distance are linked through an initialization process.

We utilized clustering algorithms to classify the keyphrases into distinct semantic groups. Due to the difficulties of establishing a topic cluster, several clustering methods (such as k-means (MacQueen 1967) and DBSCAN (Daszykowski and Walczak 2009)) have been proposed. For data with a nonflat geometry and uneven cluster sizes, such as the topic point set generated by t-SNE in the previous step, DBSCAN, a density-based clustering algorithm, is among the well-evaluated algorithms that perform well in locating data regions with high density of observations versus the data regions with lower densities. In comparison to other approaches such as k-means, DBSCAN's ability to categorize data into arbitrarily-shaped clusters is a significant benefit. DBSCAN can connect two points within a specific radius and put them in the same cluster; this operation is spatially iterated. Consequently, numerous distant but continuously distributed points in the space can be integrated into the same cluster. In our scenario, we use DBSCAN to cluster points that are continuously distributed in close proximity to one another in space, regardless of whether

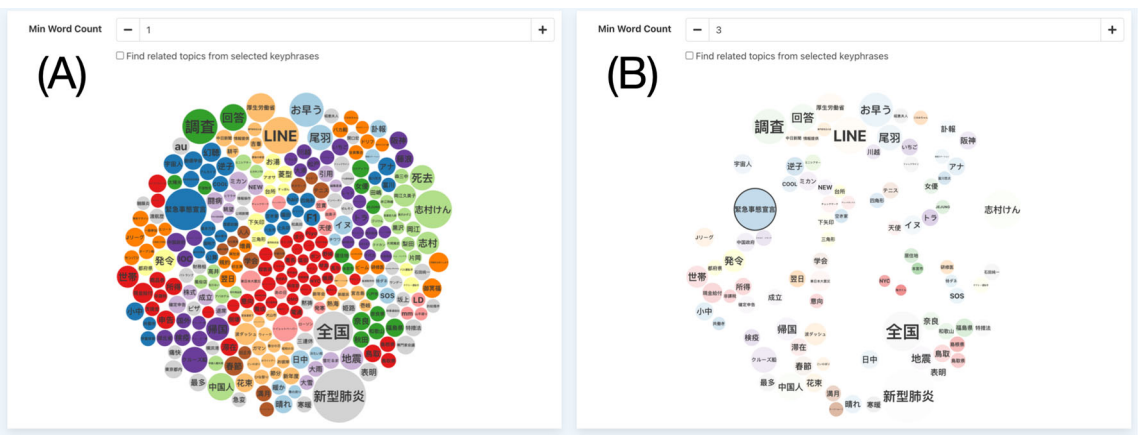

Fig. 5 Default keyphrase view (**A**) and the topic and word selected visualization result (**B**). The keyphrase view shows the keyphrases described in Section 3.1 as circles

they are likewise close to one another in time. Other clustering algorithms, such as k-means, may include additional irrelevant points.

In addition, we choose DBSCAN to cluster the keyphrases because of the combined advantages of t-SNE and DBSCAN.

4.4 Timeline view

The timeline view (Fig. 6) depicts the change in the number of SNS posts as a bar chart, with the horizontal axis representing time. In the timeline view, the occurrence of certain user-selected words is displayed as an overwrapped line chart. This view assists users in comprehending the temporal links between topics in a topic projection view. The colors of the bars and lines correspond to the time and the word group in the topic projection and keyphrase views, respectively.

4.5 User interactions

The proposed system includes a variety of user interactions that connect each view.

The selection of topics in the topic projection view constitutes the initial interaction. The proposed system automatically extracts the topics related to a specific topic to assist users in discovering related topics. Similar to the keyphrase view mentioned in Section 3.3, clustering is conducted using DBSCAN for the 2D coordinates of the topics after dimensionality reduction to extract similar topics. DBSCAN's radius can be modified to control the scope of extracted topics. A wider radius was used to extract a greater number of relevant topics, whereas a smaller one was used to extract fewer related topics. In the timeline view, the bar that holds the extracted topics' dates is highlighted. The opacity of a keyphrase in the keyphrase view corresponds to its frequency in the extracted topics.

Using the timeline view, the second interaction involves selecting the topics that appeared at a specific moment. Similar to the topic projection view selection procedure, when users pick some bars in a timeline view, the topics in the time represented as bars are selected.

The third interaction involves filtering keyphrases in the keyphrase view by adjusting the minimum word occurrence value. The keyphrases whose frequencies are below the threshold value are concealed. When users filter topics by selecting a topic group in the topic projection view, the total word occurrence within the selected topics is calculated. This aids users in locating significant keyphrases that appear in all the topics or user-selected topics.

The final interaction permits displaying the time series variation in the frequency of occurrence of the user-selected keyphrases. A time series line appears in the timeline view when a keyphrase is selected. This facilitates comprehension of word occurrence and topical transitions. In addition, the option "Find-related topics from selected keyphrases" has been added above the keyphrase view. When this option is enabled if a user selects a keyphrase by clicking the word bubble, the topics that include more instances of the selected keyphrase than the minimum word occurrence value are highlighted. This facilitates the discovery of the relationships between topics using keyphrases.

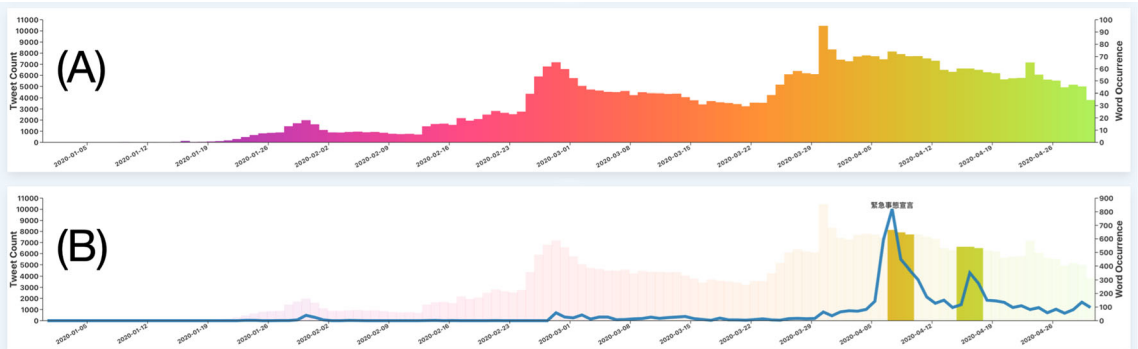

Fig. 6 Default timeline view (**A**), and visualization result of the selected topics and words (**B**). The timeline view shows the overview number of all SNS posts and user-selected posts

5 Application example

This section demonstrates two application examples of the proposed system using real-world SNS data. For both application instances, we gathered short-text SNS data from Twitter. The text of the obtained data contains continuous Japanese tweets filtered using a certain keyword. For each application example, we tested several values and decided the group and window sizes based on the duration of each dataset. Before the data preprocessing stage, we segmented words using the morphological analysis engine, MeCab (Taku 2005), and the dictionary mecab-ipadic-NEologd (Toshinori 2015). Using the collected data to train a Doc2Vec model, we obtained 300-dimensional topic and word vectors.

For the first dataset, we gathered 424,904 tweets from Japan containing COVID-19-related terms between January 1, 2020, and April 30, 2020. We set the group and window sizes for the topic points to 400 and 3, respectively. The two parameters' values are empirically determined based on the quantity of SNS posts and the four-month duration of the dataset. In addition, we extracted five keyphrases for each group, resulting in a list of the 300 frequently occurring keyphrases. A first glance at the visualization result using our system is depicted in Fig. 1 in the previous section. In addition, some interesting results obtained using our system are presented in the subsequent paragraphs.

As shown in Fig. 7A, a group of related topics was selected from the topic projection view shown as a highlight. Using the timeline view, we can determine that this topic group primarily appeared on April 7 and around April 16. The posts in these periods contained similar topics. The keyphrase view displayed some information by highlighting the words that occurred more frequently in these periods. For instance, the "state of emergency declaration (緊急事態宣言)," was displayed in a bigger size than other words. The line in the timeline view indicates that the largest peak period of the word's appearance occurred around April 7, and a smaller one occurred on April 16. According to the news, the Japanese government issued an emergency declaration for seven prefectures on April 7 and a decision to extend the emergency declaration to all prefectures on April 16.

Figure 7B demonstrates that one of the peaks in the number of posts is located in the middle-left of the topic projection view. By selecting the cluster, we can determine that this peak occurred on March 30, 2020. The highlighted words in the keyphrase view include important words such as "Ken Shimura (志村けん)," a famous Japanese comedian. On March 30, it was revealed that he passed away owing to complications from COVID-19.

The word "cash benefits (現金給付)" in the lower-left corner of the keyphrase view is selected (Fig. 7C). By clicking on the word, a line of occurrences is displayed in the timeline view for this word, with peaks around March 18, 2020, and April 3, 2020. The news articles on these dates were titled "Government considers cash benefits as a measure against COVID-19" on March 18 and "The government has decided on the cash benefit framework due to the spread of COVID-19 infection" on April 3.

In the second application example, between February 28, 2011, and March 31, 2013, 2,183,237 Japanese tweets containing keywords connected to the Fukushima nuclear disaster were collected. We set the group and window sizes for topic points to 1500 and 5, respectively. Because the number of posts and period is comparatively greater than in the initial COVID-19 dataset, we increased the group size to adjust the number of resultant topic points. The window size value has also been modified based on prior knowledge to accommodate the new dataset. In addition, we extracted five keyphrases for each group, resulting in a list of the 300 frequently occurring keyphrases. The interesting results of our system are presented in subsequent paragraphs.

We selected a glimpse of tweets from the timeline view. Then, to identify a complete cluster, we selected the cluster that contained the topics during the peak (Fig. 8A). As indicated by the keyphrase view, the word "Setagaya (世田谷)" appeared frequently. High amounts of radiation were detected in this region during this period and reported by the media. People frequently tweeted about this news during this time, causing this peak.

Then, we selected a cluster in the lower-left corner of Fig. 8B. The keyphrase view displays some words that appeared in October 2011. We selected the small word "Sonoda (園田)," a person's name. The small size of this word implies that it does not appear frequently. The word "Parliamentary Secretary (政務官)" next to it shows the positions of these individuals. We discovered that Sonoda, the parliamentary secretary, had drunk water from two reactor buildings at the Fukushima Daiichi nuclear power plant at this time by searching the news with these words and dates.

Another example is the word "total darkness (真っ暗)," located at the bottom-left of Fig. 8C. The word is small in the keyphrase view, indicating that its occurrence was uncommon compared with other

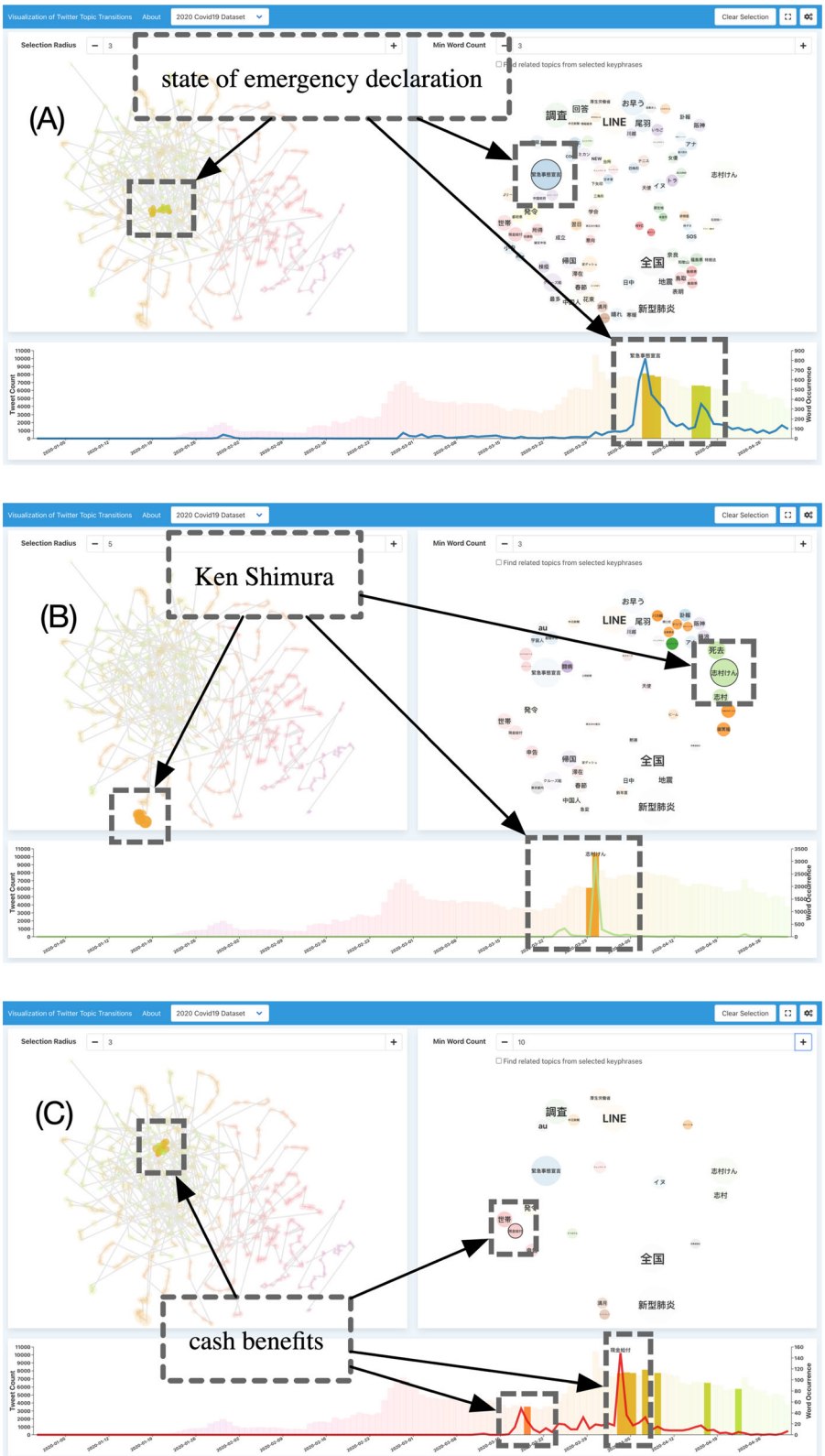

Fig. 7 Three visualization results of the proposed system with respect to the COVID-19 trend on Twitter in Japan. A keyphrase strongly related to the selected topic cluster was found for each result

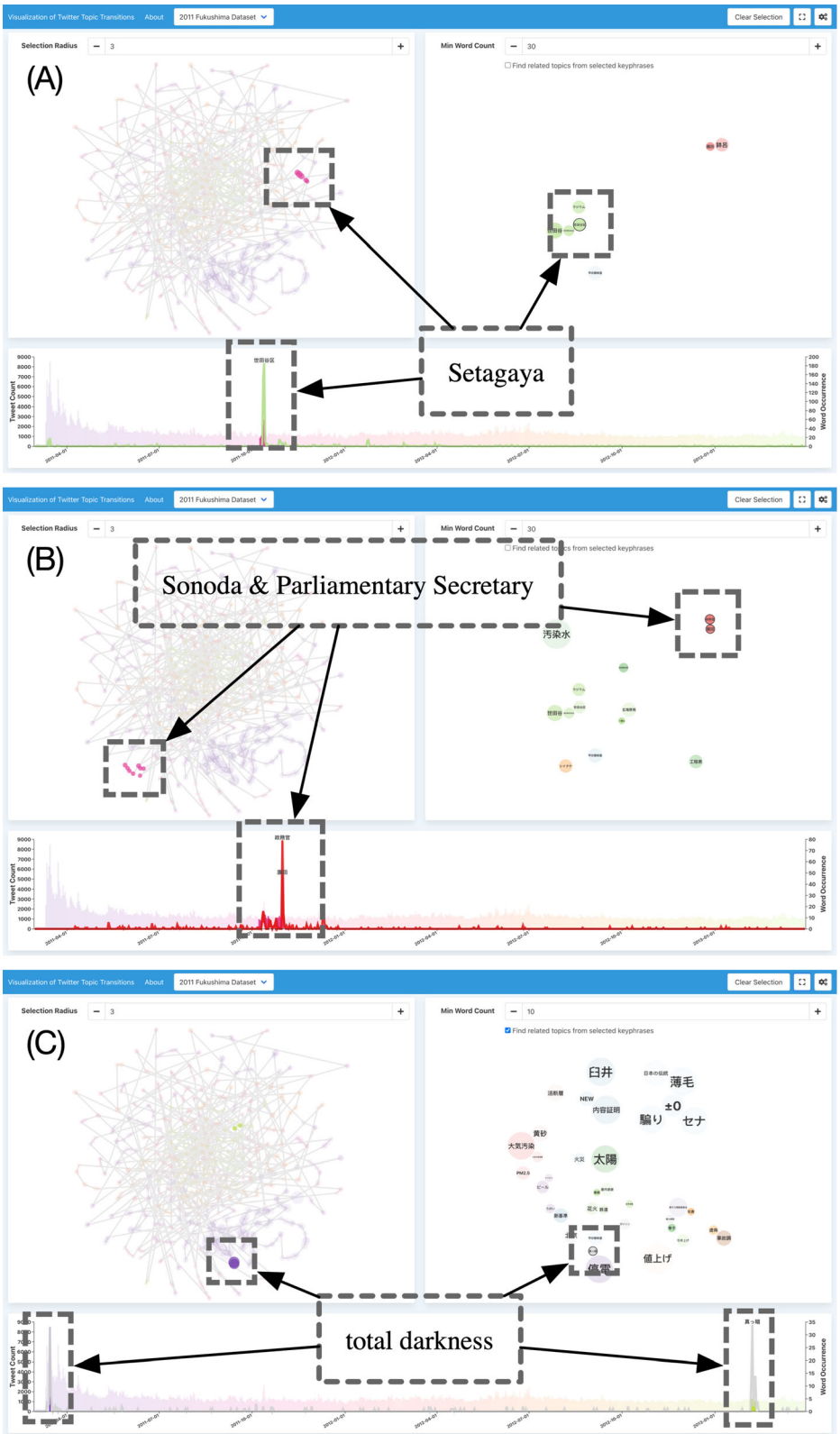

Fig. 8 Three visualization results of the proposed system with respect to the Fukushima nuclear disaster's topic trend on Twitter in Japan

keyphrases. Using the "Find-related topics from selected keyphrases" option and selecting this word, the topic projection view revealed that topics with frequent occurrences of this word have similar statuses. In addition, the timeline view revealed two peaks around March 15, 2011, and February 8, 2013. The news from these dates revealed that after Tokyo Electric Power Company said on March 11, 2011, that the nuclear power station building was in complete darkness, a report on February 7, 2013, stated that this was not true and that the facility's lights were on at the time.

Using our proposed approach, we quickly identified the characteristics of topical trends in Japanese tweets based on these two application cases. The characteristics well reflected real-life events and people's responses to them. Therefore, our system successfully visualized the topic transitions of these events as intended.

6 User evaluation

6.1 Procedure

We conducted a user evaluation of the proposed system with the assistance of 19 university students interested in information visualization from the Department of Information Science at Nihon University. We asked the participants 10 questions corresponding to the design requirements listed in Table 1 and gathered their feedback for dataset 1 (COVID-19) and dataset 2 (Fukushima). Table 1 demonstrates the relationships between design requirements and questions. Each question is measured on a 7-point Likert scale, with 1 indicating "not applicable" and 7 indicating "applicable." The evaluation was completed remotely via voice calls and chats using Internet communication technology. The approach for evaluation was as follows:

1. Video-based explanation of the proposed system
2. Exploration of dataset 1 with the proposed system
3. Responding to questions for dataset 1
4. Using the proposed approach to explore dataset 2
5. Answering questions for dataset 2

In steps 2 and 4, the participant's screen was captured while they used the proposed system, allowing us to monitor their reactions to the system's operation.

6.2 Results

Figure 9 depicts the box plot of each question's responses in each dataset. The majority of questions were answered in the affirmative, leading to the "applicable" side from which it was concluded that the proposed system could meet the design specifications.

Except for Questions 6 and 8, the average value of the evaluation was greater when using dataset 2 than when using dataset 1. This is because the questions regarding dataset 1 were answered prior to those regarding dataset 2; hence, the participants already grasped the meanings of the questions and the system's working mechanism. Questions 6 through 8 pertain to the frequency of occurrences of key phrases, and it

Table 1 Questions for user evaluation

No.	Related design goals	Question
Q1	G1, G2	Utilizing the proposed system, users can search for and compare the difference in topical contents between several time periods
Q2	G1, G2	Users can discover when the SNS posts shared have similar topics (In a similar status)
Q3	G1	Users can find the topics that have the same appearance pattern
Q4	G3	Users can find a group of topics based on a date
Q5	G3	Users can find the SNS topic groups closest to the publication time of specific topic groups based on their contents
Q6	G4	The popularity of SNS topics is simple to determine and understand
Q7	G4	The fluctuating popularity of SNS topics are simple to discover and comprehend
Q8	G4	It is simple to compare the changes in popularity between different topics or distinct times
Q9	G5	The system functions are easy to comprehend
Q10	G5	The system's output is simple to understand

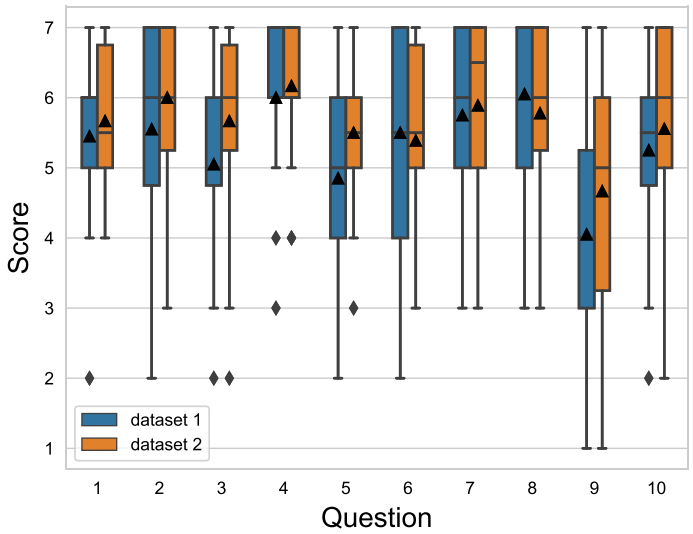

Fig. 9 Box plot of the user evaluation results

can be seen that the average ratings for dataset 2 were lower than those for dataset 1. Consequently, it may take a while to comprehend the system-specific term "keyphrase." In addition, it was discovered from the user's actions that it is difficult to comprehend the number of occurrences of keyphrases in dataset 2 due to the increased total amount of sentences.

Comparing by time series (Q7, Q8) and searching for related topics (Q2) are likely to yield high scores, indicating that the system performs properly. In addition, the features for adjusting the clustering radius and the visualization results were intuitive. Nevertheless, the average evaluation score was low in terms of intuitively comprehending the operation (e.g., the knowledge of element selection) of the system prior to displaying the visualization results. It was concluded that this point should be improved in a future study.

We summarized some of the interesting feedback from participants as follows:

- I noticed many people's names are featured in the keyphrase view, and their occurrences frequently endure for a considerable amount of time.
- It was amusing to discover that words that I believed to be unrelated are connected according to the algorithm.
- The line chart in the timeline view made it simpler for me to comprehend the relationship between keyphrases and their occurrence counts.
- ven though I do not recall much about it, I felt that applying this technique would help me understand the flow and provide me with a clue to search for additional information.

This feedback demonstrates that our proposed system is functional and meets its design specifications.

7 Discussion

The effectiveness and utilization of the proposed system were validated by application examples and user evaluation. The application examples showed that our system could find complex topical transitions even when similar topics are temporally separated (e.g., Figs. 7A, C, 8C). The user evaluations showed that our system has the required functionalities.

Our system focuses on visualizing the status of changes in social media topics, the center group of the topic points in the projection view is often large, and the content points are close to each other. This indicates that most topic points have similar contents, and our system considers these topic points as having similar statuses. Although the density of the topic points in the projection view could be modified by adjusting the parameters during document embedding and dimensionality reduction, the visualization design for the central part of the topic points might be improved with data interpretation.

8 Conclusion and future work

This study proposes and develops a novel visual analytics system that facilitates exploring and analyzing topic transitions on social networking sites. Using document embedding and dimensionality reduction techniques, the proposed system summarizes successive SNS posts as topics and keyphrases, then visualizes them in three interconnected views. Using actual SNS data, we presented two application examples to demonstrate the usefulness of the proposed approach. Overall, the approach provided in this paper can be utilized to investigate the patterns of topical transitions and to visualize social media topics at certain periods, even for complex topic transitions. Our source code and demo are accessible at https://github.com/vdslab/twitter-topic-transitions.

Additional research can be undertaken in the future. First, the natural language processing and machine learning techniques used in this study are fundamental and can be improved in terms of accuracy and performance. For example, a suitable dimensional reduction algorithm depends on data and task features. In addition to the already employed dimensional reduction technique t-SNE, it could be beneficial to include a range of additional algorithms for interactive comparison. Moreover, sentiment analysis techniques can be incorporated into our proposed system to describe topic transitions more precisely. Second, existing topic visualization tools, such as EvoRiver and time curves, can be integrated into our system to visualize temporal information in greater detail. To evaluate and validate our methodology, we intend to apply it to another dataset, such as another Twitter topic, data gathered from websites other than Twitter, and topical transitions in more general text datasets (such as newspapers and short reviews on e-commerce websites).

References

Adä I, Thiel K, Berthold MR (2010) Distance aware tag clouds. In: 2010 IEEE international conference on systems, man and cybernetics, pp 2316–2322. https://doi.org/10.1109/ICSMC.2010.5641993

Bach B, Shi C, Heulot N, Madhyastha T, Grabowski T, Dragicevic P (2016) Time curves: folding time to visualize patterns of temporal evolution in data. IEEE Trans Visual Comput Graphics 22(1):559–568. https://doi.org/10.1109/TVCG.2015.2467851

Barth L, Fabrikant SI, Kobourov SG, Lubiw A, Nöllenburg M, Okamoto Y, Pupyrev S, Squarcella C, Ueckerdt T, Wolff A (2014) Semantic word cloud representations: Hardness and approximation algorithms. In: Latin American symposium on theoretical informatics. Springer, pp 514–525

Binucci C, Didimo W, Spataro E (2016) Fully dynamic semantic word clouds. In: 2016 7th international conference on information, intelligence, systems applications (IISA), pp 1–6 . https://doi.org/10.1109/IISA.2016.7785428

Bostock M, Ogievetsky V, Heer J (2011) D^3 data-driven documents. IEEE Trans Visual Comput Graphics 17(12):2301–2309. https://doi.org/10.1109/TVCG.2011.185

Cao N, Lin YR, Sun X, Lazer D, Liu S, Qu H (2012) Whisper: tracing the spatiotemporal process of information diffusion in real time. IEEE Trans Visual Comput Graphics 18(12):2649–2658. https://doi.org/10.1109/TVCG.2012.291

Cui W, Wu Y, Liu S, Wei F, Zhou MX, Qu H (2010) Context preserving dynamic word cloud visualization. In: 2010 IEEE pacific visualization symposium (PacificVis), pp 121–128 . https://doi.org/10.1109/PACIFICVIS.2010.5429600

Daszykowski M, Walczak B (2009) Density-based clustering methods. Compr Chemom 2:635–654. https://doi.org/10.1016/B978-044452701-1.00067-3

Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21(11):1129–1164. https://doi.org/10.1002/spe.4380211102

Kruskal JB (1978) Multidimensional scaling, vol 11. Sage

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196

MacQueen J (1967) Others: some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297

Natsukawa H, Deyle ER, Pao GM, Koyamada K, Sugihara G (2020) A visual analytics approach for ecosystem dynamics based on empirical dynamic modeling. IEEE Trans Visual Comput Graphics 27(2):506–516

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523

Sun G, Wu Y, Liu S, Peng TQ, Zhu JJ, Liang R (2014) EvoRiver: visual analysis of topic coopetition on social media. IEEE Trans Visual Comput Graphics 20(12):1753–1762. https://doi.org/10.1109/TVCG.2014.2346919

Taku K (2005) MeCab: Yet another part-of-speech and morphological analyzer. http://chasen.org/taku/software/mecab/ (2013/04/15)

Toshinori S (2015) Neologism dictionary based on the language resources on the Web for Mecab. https://github.com/neologd/mecab-ipadic-neologd

Van Den Elzen S, Holten D, Blaas J, Van Wijk JJ (2016) Reducing snapshots to points: a visual analytics approach to dynamic network exploration. IEEE Trans Visual Comput Graphics 22(1):1–10. https://doi.org/10.1109/TVCG.2015.2468078

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605

Verlet L (1967) Computer "Experiments'' on classical fluids: I: thermodynamical properties of lennard-jones molecules. Phys Review 159(1):98–103. https://doi.org/10.1103/PhysRev.159.98

Viégas F, Wattenberg M, Hebert J, Borggaard G, Cichowlas A, Feinberg J, Orwant J, Wren C (2013) Google+ripples: a native visualization of information flow. In: Proceedings of the 22nd international conference on world wide web, WWW '13, pp 1389–1398. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2488388.2488504. https://doi.org/10.1145/2488388.2488504

Viegas FB, Wattenberg M, Feinberg J (2009) Participatory visualization with wordle. IEEE Trans Visual Comput Graphics 15 (6):1137–1144. https://doi.org/10.1109/TVCG.2009.171

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1–3):37–52

Wu Y, Liu S, Yan K, Liu M, Wu F (2014) OpinionFlow: visual analysis of opinion diffusion on social media. IEEE Trans Visual Comput Graphics 20(12):1763–1772. https://doi.org/10.1109/TVCG.2014.2346920

Wu Y, Provan T, Wei F, Liu S, Ma KL (2011) Semantic-preserving word clouds by seam carving. Comput Graphics Forum 30 (3):741–750. https://doi.org/10.1111/j.1467-8659.2011.01923.x

Zhao J, Cao N, Wen Z, Song Y, Lin YR, Collins C (2014) #FluxFlow: visual analysis of anomalous information spreading on social media. IEEE Trans Visual Comput Graphics 20(12):1773–1782. https://doi.org/10.1109/TVCG.2014.2346922