EPiC Series in Computing

Volume 81, 2022, Pages 362–372

Proceedings of 11th International Congress on Advanced Applied Informatics

Realization of Discovery for Burst Topic Transition Using the Topic Change Point Detection Method for Time-Series Text Data

Yuta Ishii[1], Aiha Ikegami[1] and Takafumi Nakanishi[1*]

[1] Musashino University, Tokyo, Japan

s2022052@stu.musashino-u.ac.jp, s2022003@musashino-u.ac.jp, takafumi.nakanishi@ds.musashino-u.ac.jp

Abstract

In this paper, we present a realization method of discovery for burst topic transition using the topic change point detection method for time-series text data. In our method, we focus on the topic change point detection method for time-series text data. By similarity measure using the topic change point detection method for time-series text data, we can discover for burst topic transition. In general, when we would like to understand the outline or main points of an event, we often read articles written by people who know information about the event or ask others who are aware of the event to tell us about it. However, the information obtained by these means is hearsay from others and subject to third-party bias, it is difficult to comprehend the events objectively. In our paper, we focus on the topic change and extract the topic change point detection It enables us to discover burst topic transitions. In this paper, we describe an evaluation experiment of a prototype system using our discovery for burst topic transition to verify the effectiveness of our method. We also implement an application by the user interface that provides some crews of a trendy word.

1 Introduction

With the spread of social media such as Twitter and Instagram in recent years, how we can obtain information such as news has become more diverse. With the diversification of the means of collecting information, the opportunities for us to obtain information have increased dramatically. However, with the increasing availability of information, it has become very difficult for us to properly understand these large amounts of information. Today where we are overflowing with information, what is important is to comprehend the time-series trend of a large amount of information that is updated daily from a broad perspective.

* Corresponding author

In general, when we would like to understand the outline or main points of an event, we often read articles written by people who know information about the event or ask others who are aware of the event to tell us about it. However, the information obtained by these means is hearsay from others and subject to third-party bias, it is difficult to comprehend the events objectively. Therefore, we consider that by drawing from the large amount of information in the world, we would not be influenced by the bias of third parties. We need tools to capture changes in topic trends from large amounts of time-series text data due to the difficulty in grasping a large amount of data.

In this paper, we focus on the topic change and extract the topic change point detection, which enables us to discover burst topic transitions. We propose a new method for globally capturing time-series trends in a large amount of information that is updated and increased every day. In our method, we realize a method discovery for burst topic transition for time-series text data, focusing on the topic change point detection. We use the topic modeling method and the change point detection method to discover topic transition. In this paper. We propose a method of the topic change point detection, which is a method of finding the reason why an event has become a burst topic. In our method, the point where there is a large change in the topic about the event is regarded as the origin of the excitement of the event. In our method, the topic change points extracted by the change point detection method can be visualized as the origin of the hot topic.

The main features of this paper are as follows.

- We present a realization method of discovery for burst the topic transition using the topic change point detection method for time-series text data.
- We develop our prototype system for our method by similarity measure using the topic change point detection method for time-series text data. We also conduct some evaluation experiments on our prototype system.
- We implement an application for the user interface which provides some crews of the trendy word.

This paper is organized as follows. In Section 2, the related works are discussed. In Section 3, we present our proposed method of discovery for burst topic transition using the topic modeling method and change point detection method for time-series text data. In section 4, we construct an experimental system to verify the effectiveness of our method. In section 5, we describe the implementation of an application that uses our method. Finally, in Section 6, we summarize this paper.

2 RELATED WORKS

In this section, we discuss studies related to our method, topic modeling method and change point detection method.

In our method, we focus on the following two methods.

1. Topic modeling method
2. Change point detection method

We will introduce some studies that apply the topic modeling method to time series data. Kanungsukkasem et al. (N. Kanungsukkasem, FinLDA: Feature Extraction in Text and Data Mining for Financial Time Series Prediction, 2019) used the topic modeling method to build a new system from a combination of news articles and financial time-series. Sota Kato et al. (S. Kato, Time-series topic analysis using singular spectrum transformation for detecting political business cycles, 2021) analyzed a text dataset containing the Japanese Prime Minister's (PM's) detailed daily schedule for over 32 years using the topic modeling method. D. Koike et al. (D. Koike, Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter, 2013) used the topic modeling method to show the correlation between online news and Twitter. In this section, we introduce some studies that used the

change point detection method for time-series data. Shinnosuke Takeda et al. (S. Takeda, Irregular Trend Finder: Visualization tool for analyzing time-series big data, 2012) developed a tool to analyze time-series big data and find irregular trends. Yoshinobu Kawahara et al. (Y. Kawahara, Change-point detection in time-series data based on subspace identification, 2007) proposed a series of algorithms for detecting change points in time-series data based on subspace identification. These studies used a change point detection method for time-series data. In general, the change point detection method is a method for time-series numerical data such as the above.

In our method, we applied the change finder to time-series text data. We realize the discovery for burst topic transition using the topic change point detection method by combining two methods: LDA (David M. Blei, Latent Dirichlet Allocation, 2003), which is a topic modeling method, and the change finder (J. Takeuchi, A Unifying Framework for Detecting Outliers and Change Points from Time Series, 2006), which is a change point detection method.

3 Discovery for Burst Topic Transition Using the Topic Change Point Detection Method for Time-Series Text Data

Our method consists of a topic analysis function and a change point detection function. In this section, we present our proposed method for discovery for burst topic transition using the topic change point detection method for time-series text data.

In section 3.1, we describe the overview of our method. In section 3.2, we describe the topic analysis function. In section 3.3, we describe the change point detection function.

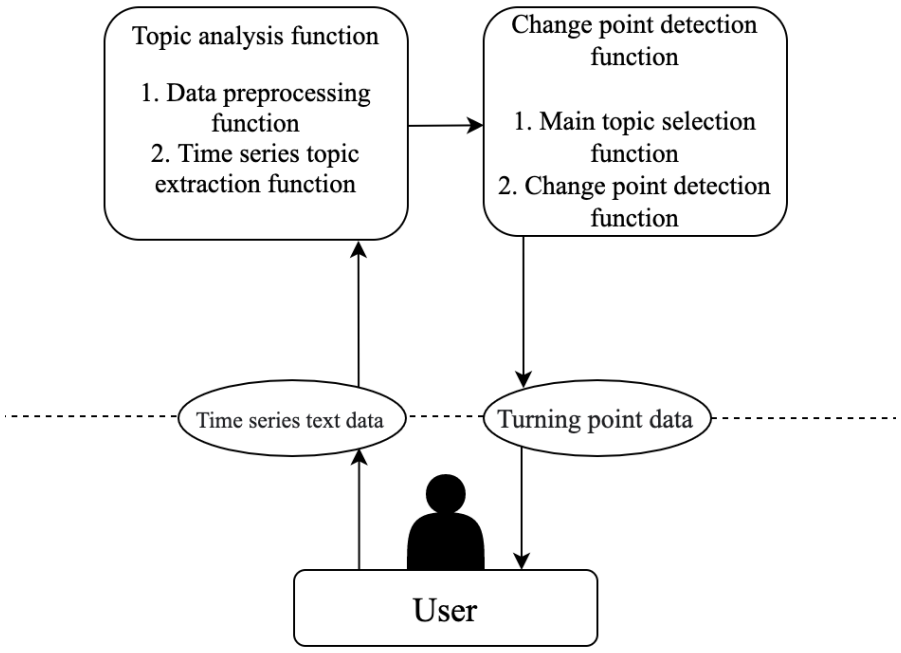

Figure 1: Overview of our method

3.1 Overview

In this section, we present an overview of our proposed method. The overview of our method is shown in Figure 1. This system consists of the topic analysis function and the change point detection function. The topic analysis function consists of the data preprocessing function and the time series topic extraction function. The change point detection consists of the main topic selection function and the change point detection function. The topic analysis function is a function to identify topics from time-series text data. The change point detection function is a function that detects change points in time-series topic data. This system enables the extraction of burst topic transition from time-series text data. To extract burst topic transitions from time-series text data, the topic analysis function is used to extract and cluster topics. We also use the change point detection function to extract burst topic transitions from the quantified time-series text data. This method targets Japanese text. however, it can be applied to other languages when the words can be extracted by some morphological analysis for the other language.

3.2 Topic analysis function

The topic analysis function is a function to identify topics from time-series text data.
The topic analysis function performs topic analysis on the time-series text data input by the user. The topic analysis function consists of two parts: the data preprocessing function and the time series topic extraction function. The purpose of this function is to classify topics and quantify time-series text data.

3.2.1 Data preprocessing function

The data preprocessing function in the topic analysis function involves extracting some words each time from the input data. As shown in Figure 1, the input data consist of by each time is text data separated by a fixed interval of time. An example of input data is time-series text data, in which the period is from the current time to one week ago, separated by one hour.

This function consists of 3 steps.

Step1: Morphological analysis and noun, adjective, and adjectival verb extraction

Morphological analysis is performed to extract words of a specific part of speech.

Part of speech is a classification of Japanese words according to their grammatical function or form. Among these parts of speech, nouns, adjectives, and adjectival verbs are the parts of speech that have meaning by themselves (independent words). The topic analysis considers a sentence as a set of multiple words, so nouns, adjectives, and adjectival verbs need to be extracted. In our method, we used MeCab (T. Kudo, MeCab: Yet Another Part-of-Speech and Morphological Analyzer, n.d.), to perform morphological analysis. As a morphological dictionary, we used mecab-ipadic-NEologd (T. Sato, Neologism dictionary for MeCab, n.d.). mecab-ipadic-NEologd is a morphological analysis dictionary for MeCab. This dictionary can properly morphological analyze new words and unique expressions because its update frequency is more than twice a week.

Step2: Elimination of words from the stop word list

Stop words are words with little information or words that are not related to the task, to reduce the number of words used. Examples of stop words are words that have no deep meaning, such as “do”, “be”, and “become”. These words are necessary for the sentence structure, but they don't have too many features for analysis, so they are removed.

Step3: Elimination of word with a frequency of one occurrence

Words with a frequency of very few occurrences need to be removed. This is because they become noise in the topic analysis.

3.2.2 Time series topic extraction function

The time series topic extraction function extracts some topics from time-series text data by topic model. The topic modeling method is a kind of clustering method with the extraction of important words as topics. This method has the feature that one data belongs to multiple clusters. In our method, we used the topic analysis, which can be assigned to multiple clusters, to capture topic transitions in time-series text data. In our method, we use LDA (Latent Dirichlet Allocation) as the topic analysis method. By using LDA, text data can be clustered into an arbitrary number of topics. In LDA, a sentence belongs to multiple clusters, and the percentage of topics in each cluster can be calculated. This feature allows us to determine the main topics of each time. The input to the time series topic extraction function is the entire text data. The output of the time series topic extraction function is a topic model. This topic model is applied to the text data for each time. This makes it possible to calculate the percentage of topics in topics each time. The numerical value calculated by the time series topic extraction function is called the LDA score. The LDA score is time-series numerical data. The time series topic extraction function enables the classification of topics and the quantification of time-series text data, making it possible to realize the next step, change point detection.

3.3 Change point detection function

The change point detection function detects change points in time-series topic data. This function enables us to extract the change point of the topic, which is the origin of the excitement of the event. In this function, change point detection is performed on the time-series data for each topic. The change point detection function consists of the main topic selection function and the change point detection function. The change point detection function enables the change point detection of topics and realizes the discovery of burst topic transitions.

3.3.1 Main topic selection function

The main topic selection function is a function that selects the main topic based on the LDA score. In our method, we focus on the variance of the LDA score. The topic with the largest LDA score variance among the topics is the main topic. The main topic selection function enables the discovery of the most rapidly changing topics.

3.3.2 Change point detection function

The change point detection function finds the points where the LDA scores have changed. The change point detection function targets the time-series LDA score of the main topic. The change point detection function uses the change finder. Change finder is an improved version of the conventional change detection method that can handle non-stationary data. This also allows for outliers to be considered. It calculates the change score for each time for time-series data and considers the part with a high change score value as “changed”. The input of the change finder is the time-series LDA score of the main topic. The change finder basically uses some data close to the origin as a reference. However, since the time-series data, in this case is a point in the past from the present, the origin can be said to be the present. Therefore, the change point is detected retroactively from the present to the past, unlike most methods. The change point detection enables to find the point where the topic has changed, thus realizing the topic change point detection.

4 Experiment

Four experiments were conducted to verify the effectiveness of this method. We focused on the following points in each experiment.

1. The validity of topic classification (Experiment 1)
2. The changes in numerical values for each topic (Experiment 2)
3. The change score of each topic (Experiment 3)
4. The text data before and after the change point (Experiment 4)

4.1 Experimental Environment

In this experiment, we used the tweets as the experimental. Twitter is characterized by its short sentences and easy posting; thus, users tend to post their daily events and thoughts. For this reason, Twitter users often react sensitively to events in their daily lives, and it can be said to be a medium with an extremely small-time lag between Twitter and daily life. In addition, since Twitter has a very large number of users, it tends to generate exciting topics, and the topics change quickly.
For this reason, we chose tweet as the subject of our experiment. We used the Twitter API to search for and aggregate tweets containing the word “corona”. The word “corona” is intended to collect tweets related to COVID-19. The COVID-19 is a global and long-term topical word from 2019 to the present. In addition, the topics “COVID-19” are related to a wide range of fields such as medicine, economy, entertainment, and sports. The period of collecting tweets was set to one week, and a maximum of 100 tweets per hour was obtained for 168 hours.

The following three pre-processing steps are necessary for handling tweets.

1. Remove any words
2. Remove spam tweets
3. Remove URLs

In addition, since tweets are collected that contain arbitrary words, arbitrary words are always present in tweets. Therefore, the arbitrary words need to be removed since the words selected as arbitrary words have a higher frequency of occurrence. Also, some tweets are called spam tweets. Spam tweets in this paper are tweets posted in large numbers by the same user with the same content or the same text in a short period. In our paper, we consider all posts by users who post many spam tweets as spam tweets. We eliminated the spam tweets by excluding the relevant usernames. Tweets may contain not only the body text but also the URLs of links and images. Elements such as these that are not the body text need to be removed.

4.2 Experimental result

4.2.1 Experiment 1

In this section, we present the results of the verification of the method of discovery for burst topic transition using the topic change point detection method for tweet data. In this experiment, we focus on the tweet.

Table 1 shows the result of the validity of topic classification in the tweets. The words and weightings for each topic are shown here. In this experiment, the number of topics was set to three. As shown in Table 1, the main word in Topic 1 is “vaccine”. As shown in Table 1, the main word in Topic

2 is “corona disaster”. As shown in Table 1, the main word for Topic 3 is “vaccine”. By comparing each topic, we can discover that the content of Topic 1 and Topic 3 are similar.

Topic1	0.018 * "Vaccine" + 0.013 * "Corona sickness" + 0.007 * "People" + 0.007 * "Countermeasures" + 0.006 * "Before" + 0.006 * "New Corona" + 0.005 * "Infection" + 0.004 * "Inoculation" + 0.004 * "mask" + 0.004 * "now"
Topic2	0.013 * "Corona Mask" + 0.012 * "Vaccine" + 0.008 * "New Corona" + 0.008 * "Inoculation" + 0.006 * "People" + 0.006 * "Infection" + 0.006 * "Now" + 0.005 * "Countermeasures" + 0.004 * "Mask" + 0.004 * "Today"
Topic3	0.014 * "Vaccine" + 0.010 * "Corona Sickness" + 0.007 * "Inoculation" + 0.007 * "New Corona" + 0.007 * "People" + 0.006 * "Countermeasures" + 0.005 * "Infection" + 0.004 * "Now" + 0.004 * "Before" + 0.004 * "New Coronavirus"

Table 1: The result of the validity of topic classification in tweet data

4.2.2 Experiment 2

The result of changes in numerical values for each topic in the tweets is shown in Figure 2. In Figure 2, the vertical axis shows the LDA score, and the horizontal axis shows the time. In addition, each line indicates each topic: blue for Topic 1, red for Topic 2, and green for Topic 3. The value of Topic 2 was basically high, becoming especially high in the middle, and at the end, it was the same scale as at the beginning. The value of Topic 1 was also higher than Topic 2 and maintained cyclical stability, however, it dropped sharply at the timing of the spike in Topic 2. Topic 3 showed some change at first, however, after the surge of Topic 1, the change became smaller. The result of the variance of the LDA score is shown in Table 2. In Table 2, the vertical axis is the variance of the LDA score, and the horizontal axis is time. Based on this result, topic 2 with the maximum value of 0.06 among these topics is chosen as the main topic.

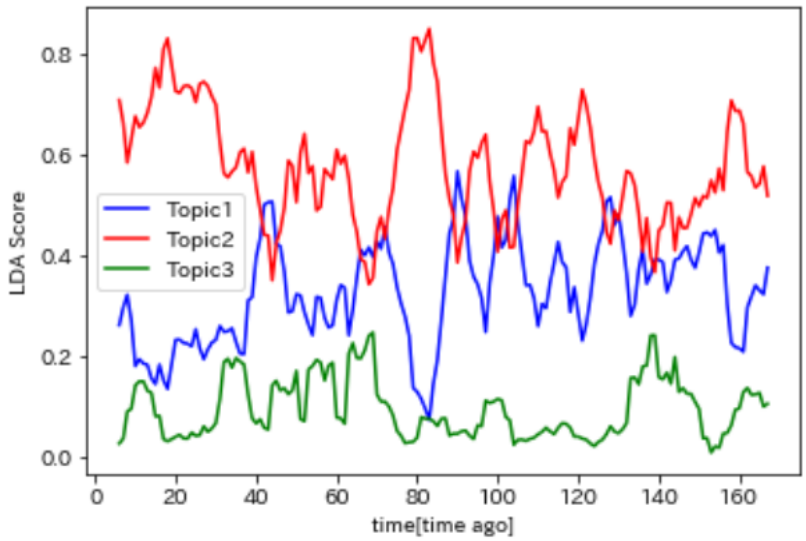

Figure 2: The result of changes in numerical values for each topic in tweet data

Topic	LDA Score variance
Topic1	0.0525
Topic2	0.0620
Topic3	0.0209

Table 2: The result of the validity of topic classification in tweet data

4.2.3 Experiment 3

The result of changes in numerical values and scores in the tweets (Topic 2) is shown in Figure 3. This figure shows the graphs of change score and LDA score for Topic 2, which is the main topic. The horizontal axis is time in both figures. The range of time is 0~168 hours ago. The values of the change score varied in the range of 0-75. The data for the first few hours is set to 0 because it is used as the basis for calculating the change score. The maximum value of the change score was 73.5, 79 hours ago. The value of the LDA score changed in the range of approximately 0 to 0.9. When the change score was at its maximum value, the LDA score also increased rapidly.

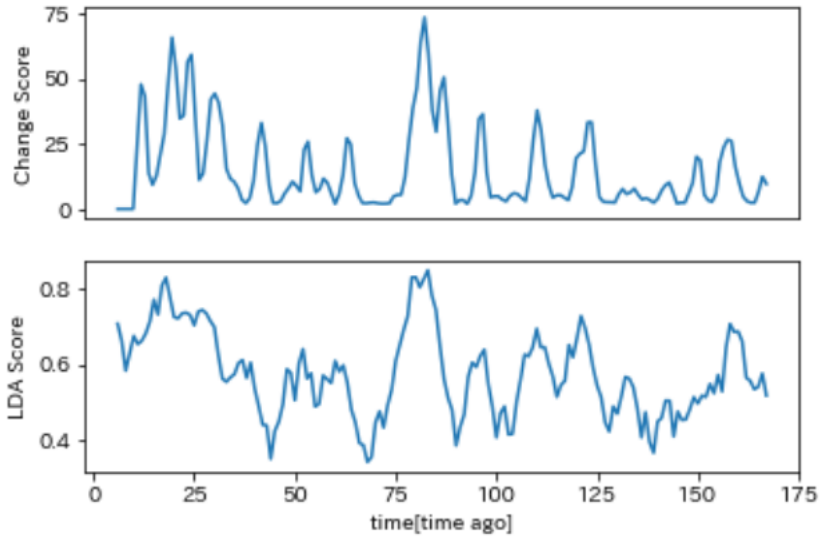

Figure 3: The result of changes in numerical values and score in tweet data (Topic 2)

4.2.4 Experiment 4

The result of text data before the change point in the tweet is shown in Figure 4. The result of text data after the change point in the tweet is shown in Figure 5. Before the change, the word cloud was dominated by words related to daily life, such as people and year. After the change, the word cloud showed many words related to politics, such as Democratic and Opposition. Also, the time of this change point was exactly 2:00 a.m. on October 29th. The previous day, October 28, was the day that the Tokyo Metropolitan Government lowered its infectious disease alert level to the lowest level.

Figure 4: The result of text data before the change point in tweet data

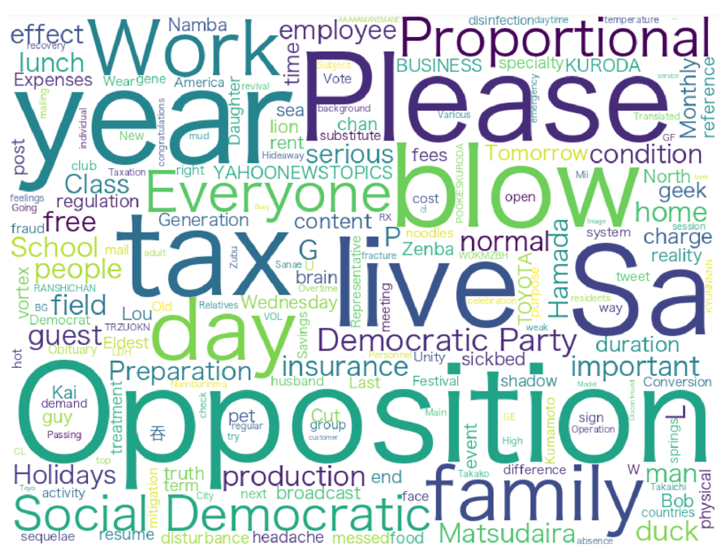

Figure 5: The result of text data after the change point in tweet data

4.3 Discussion

In the result of Experiment 1, we observe that the words with the highest weight are Topic1 and Topic3. Since there are three topics to classify, the contents of Topic 1 and Topic 3 are similar. We consider this to be due to the limited meaning of the search terms used in the experiment. This can be attributed to the fact that the search term was a pinpoint word.

In the results of Experiment 2, we consider that the main topics calculated by variance are consistent with the topics that humans can judge as main topics in Figure 2. Therefore, we conclude that we can extract the main topic correctly.

In the results of Experiment 3, we consider that the LDA score changed significantly in Figure 3, and the change score also increased. Therefore, we conclude that it enables to correctly capture the change in the topic.

In the results of Experiment 4, we consider that the main content is different before and after the change. Therefore, we concluded that the change in the topic was correctly captured.

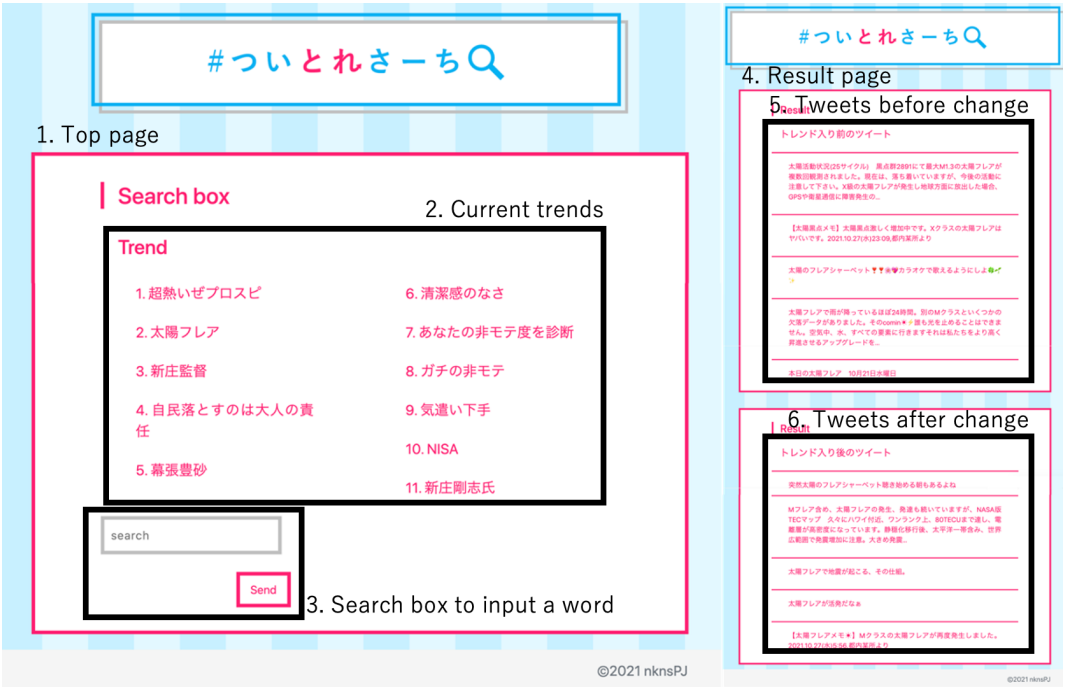

Figure 6: A user interface of the created application. 1. This application's top field, 2. A user can select a trend word and the system represents new results of the filed-4, 3. A user can search by keywords and the system can represent new results of the field-4, 4. This application's result page, 5. The system represents the tweet before the topic change, 6. The system represents the tweet after the topic change

5 Implementation

In this section, we present the implementation of for user interface application that provides some crews of a trendy word. We implemented an application with a user interface. The user interface of the application is shown in Figure 6. The idea behind this application is to find triggers for topics related to

trending words on Twitter. This application uses the “trending” feature of Twitter. A trend in Twitter is an algorithm that determines which topics are getting the most attention, such as those that are being tweeted about a lot. By looking at Twitter's “trends”, you can find the latest topics in real-time. The top page in Figure 6 has the current trend and a search field. The user selects a word that he or she wants to know why it is trending and enters it in the search field. In Figure 6, the tweets related to the word entered by the user are automatically retrieved, analyzed using our method, and the tweets before and after the change are presented. When a user enters a trending word, the search result screen will display 5 tweets before and after the topic changed.

6 Conclusion

In this section, we present the conclusion of our method. In this paper, we present a realization method of discovery for burst topic transition using the topic change point detection method for time-series text data. In our method, we focus on the topic change point detection method for time-series text data. By similarity measure using the topic change point detection method for time-series text data, we can discover for burst topic transition. In our paper, we focus on topic change and extract the topic change point detection, which enables us to discover burst topic transitions. We also describe an evaluation experiment of a prototype system using our discovery for burst topic transition to verify the effectiveness of our method. Furthermore, we also implement an application for the user interface which provides some crews of a trendy word.

In addition, we would like to improve the method of discovery for burst topic transition described in this paper and develop a system that can present users with the triggers of burst topics by identifying the source of the burst topics. We will realize the discovery method of a user who delivered the tweet that triggered the topic's excitement.

References

Kanungsukkasem, N. (2019). *FinLDA: Feature Extraction in Text and Data Mining for Financial Time Series Prediction*. In *Proceedings of the IEEE Access, vol. 7* (pp. 71645–71664).

Kato, S. (2021). *Time-series topic analysis using singular spectrum transformation for detecting political business cycles*. In *Proceedings of the Journal of Cloud Comp 10, 21*.

Koike D. (2013, October). *Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter.* In *Proceedings of the International Joint Conference on Natural Language Processing* (pp. 917-921)

Takeda S. (2012). *Irregular Trend Finder: Visualization tool for analyzing time-series big data.* In *Proceedings of the IEEE Computer Society.* (pp. 305–306)

Kawahara Y. (2007). *Change-point detection in time-series data based on subspace identification.* In *Proceedings of the 7th IEEE International.* (pp. 559-564)

David M. Blei (2003). *Latent Dirichlet Allocation*. In *Proceedings of the Journal of Machine Learning Research 3* (pp. 993-1022)

Takeuchi J. (2006). *A Unifying Framework for Detecting Outliers and Change Points from Time Series.* In *Proceedings of the IEEE Transaction on Knowledge and Data Engineering, Vol. 18, No. 4, pp482-492.* (pp. 482-492)

Kudo T. (n.d.). *MeCab: Yet Another Part-of-Speech and Morphological Analyzer.* Retrieved from MeCab: http://taku910.github.io/mecab/

Sato T. (n.d.). *Neologism dictionary for MeCab.* Retrieved from GitHub: https://github.com/neologd/mecab-ipadic-neologd