TOPIC TRACKING AND VISUALIZATION METHOD USING INDEPENDENT TOPIC ANALYSIS

Takahiro Nishigaki[1], Kenta Yamamoto[2] and Takashi Onoda[1]

[1]Aoyama Gakuin University, Kanagawa, Japan
[2]Graduate School of Science and Engineering,
Aoyama Gakuin University, Kanagawa, Japan

ABSTRACT

In this paper propose a topic tracking and visualization method using Independent Topic Analysis. Independent Topic Analysis is a method for extracting mutually independent topics from the documents data by using the Independent Component Analysis. In recent years, as the amount of information increases, there is often a desire to analyse topic transitions in time-series documents and track topics. For example, it is possible to analyse the causes of trend and hoaxes by SNS and predict future changes. However, there is no topic tracking method in Independent Topic Analysis. There is also no way to visualize topic tracking. So, topics in each periodwas extracted, and topic transition was analysed based on the similarity of topics. And, a method for tracking these four topics was proposed. In addition, this paper developed an interface that visualizes time-series changes of the tracked topics and obtained effective results through user experiments.

KEYWORDS

Data Mining, Independent Topic Analysis, Text Mining, Topic Tracking

1. INTRODUCTION

In recent years, the provision of information and distribution services on the Internet has progressed, and the distribution of information via networks has been actively carried out. This accumulated data is called big data. Big data has the characteristics of large data size, various formats, and high incidence. The targets of big data are stock prices, sales of companies, images, videos, sounds, sensors, GPS, and document data. Among them, document data has a high percentage in big data. The development of search engines, it has become possible to obtain a large amount of document data with little input. However, it is difficult to confirm all document data. Therefore, text mining for finding and extracting useful information from a large amount of document data is an important research topic.

In this paper focuses on topic analysis, which is one of the themes of text mining. The topic of topic analysis is information expressed by multiple words among a large number of documents. Currently, there are many studies of topic analysis. First, there are topic models such as PLSA (Probabilistic Latent Semantic Analysis) proposed by Hoftman [1] and LDA (Latent Dirichlet Allocation) proposed by Brei [2]. Topic models are a method to extract topics by focusing on probabilistic generative models. Many topic models treat documents as a set of words (bag-of-words). Bag-of-words ignores the word order and often uses tf (term frequency) [3] or tf-idf

Dhinaharan Nagamalai et al. (Eds): AIAP, ICCIoT, CNSA, SIGML, IT, ICBB, DMDB - 2020
pp. 01-18, 2020. © CS & IT-CSCP 2020 DOI: 10.5121/csit.2020.100301

(term frequency - inverse document frequency) values. Topic models [4] can probabilistically represent the relationship between topics.

Second, there are LSA (Latent Semantic Analysis) [5] and ITA (independent topic analysis) [6, 14, 16, 17] as methods to extract topics by focusing on the relationship between topics. LSA is a method using singular value decomposition for document data and can extract topics with the largest variance of words and documents. By this method, LSA can extract uncorrelated topics. Uncorrelated topics are topics that changes in a certain topic does not affect the other topic. Also, ITA is a method using independent component analysis [7] and can extract independent Topics. Independent component analysis is a method that is often formulated as a signal processing problem of separating signals with different characteristics using statistical properties of input signals. Independent topics are topics that have no correlation and a small amount of mutual information. Also, there is a relation that independent topics include uncorrelated topics. However, uncorrelated topics and independent topics are equivalent only when topics in document data are normally distributed. Therefore, it is necessary to seek topics with high independence since topics of document data are not always distributed normally. In this paperuses ITA, which is one of the topic analysis methods focusing on the relationship between topics.

As the volume of information has grown in recent years, there has often been demands to analyze the transition of topics in time series documents and track the topics. Using topic analysis makes it possible to significantly reduce the effort of the user compared to directly viewing the document. However, ITA is not always effective for the user's request to analyze topics of time-series documents. ITA can analyze the topic of one analysis data. That is, ITA can analyze topics at a specific period but cannot analyze the entire time series documents. Also, it is not easy for users to track topics from the results at each point in time. So, in this paper proposes topic tracking method of time-series documents using ITA.

The composition of this paper is as follows. In section 2 is described related research on topic tracking and ITA. In section 3 is described the proposed topic tracking method. In section 4 is evaluated the proposed method and shows its effectiveness. Finally, in section 5 is concluded this paper.

2. RELATED WORK

In this section describes related research on topic tracking and ITA. Related research on topic tracking includes research on analysis of major topics in time series document data [8][9][10] and research on time series clustering [11][13].

2.1. Topic Tracking

First, research on analysis of main topics of time-series document data is described. There is Themeriver proposed by Havre [8][9]. Themeriver is a visualization method that analyzes time series documents and displays the flow of topics like a river flow. Where the transition of time is represented from left to right. Several colored flow are displayed on the screen, each flow represents a topic. In the paper, the topic is called the theme. In addition, keywords for expressing the contents of each theme are also displayed on the screen. In Themeriver, the width of the theme changes with the change of time and shows the transition of the amount of information at each time point. Where, the width of the theme represents the amounts of documents corresponding to the theme at each point in time. However, the method of selecting the theme from the document data is not emphasized in Themeriver. In fact, in the experiment in the paper, 64 keywords are extracted from the target document group by prior analysis.

Themeriver counts the number of documents that contain each keyword and this result is used to create the theme width. In other words, the Themeriver expresses the transition of the amount of information using the frequency of occurrence of keywords. However, Themeriver just counts the number of documents that contain each keyword. So, the degree of keyword relevance to the document is not considered. In other words, Themeriver is a way to understand rough trends, and not necessarily a powerful tool for analyze time-series documents.

Next, research of Time series clustering is described. First, clustering [11] is to divide a set of classification targets into subsets such that internal cohesion and external isolation are achieved. Also called cluster analysis in the field of statistical analysis and multivariate analysis [12]. And it is frequently used in data mining as a basic data analysis method and many topic analyses using clustering have also been proposed.

In topic analysis using clustering, each cluster extracted at a certain point in time corresponds to one topic each. Several time series clustering approaches have been proposed for tracking and analyzing temporally changing clusters. Mei proposed a statistical approach to discovering major themes from a set of time series documents [13]. The topic is regarded as a probability distribution of words in a certain period and expressed as a kind of cluster. However, there are some clusters that are not really related to one topic. In this statistical approach, the relevance of topics at successive times based on probabilistic indicators is determined. And the analysis result is a graph that shows the transition of the theme. Then, a graph is obtained that shows the transition of the theme in which the main topics are placed on the time axis as a result of analysis. Where, the time axis represents the transition of time from left to right and the unit of time axis is year. In addition, in the graph, links are established between topics based on relevance. Therefore, it is possible to grasp how the main topic changed with the passage of time. However, since independent topic analysis is different in concept from clustering, time-series clustering cannot be applied as it is to independent topic analysis.

2.2. Independent Topic Analysis

In this section, ITA (independent topic analysis) proposed by Shinohara [6] as a study seeking independent topics is described. ITA is a method of extracting topics from the document data using independent component analysis. Independent component analysis is a method that is often formulated as a signal processing problem of separating signals with different characteristics using statistical properties of input signals. In the followings that a small letter expresses scalar, a bold small letter expresses vector, and a bold capital letter expresses matrices. As common variables, $t \in (1, \ldots, k)$express topic variables, $d \in (1, \ldots, n)$ express document variables, and $w \in (1,\ \ldots, m)$ express word variables.

First, the concepts of ITA [14] is described. Matrices $\boldsymbol{V}$have *m* rows and *k* columns, which called "importance of the word *w* in the topic t" .Also, the vector v_t represent the *t* columns of vector of matrices $\boldsymbol{V}$. The vector $\boldsymbol{v}_t$ is $(v_{1,t}, \ldots, v_{m,t})^T$. Inaddition, the vector $\boldsymbol{v}_w^{\mathrm{T}}$ represent the transposition of the *w*-th rows of vector of matrices $\boldsymbol{V}$. The Vector $\boldsymbol{v}_w$ is $(v_{w,1}, \ldots, v_{w,k})^T$.Matrices $\boldsymbol{U}$ have *n* rows and *k* columns, which called "importance of the document *d* in the topic *t*". Also, the vector u_t represent the *t* columns of vector of matrices$\boldsymbol{U}$. The vector $\boldsymbol{u}_t$ is $(u_{1,t}, \ldots, u_{n,t})^T$. In addition, the vector $\boldsymbol{u}_d^{\mathrm{T}}$ represent the transposition of the d -th rows of vector of matrices$\boldsymbol{U}$. The Vector $\boldsymbol{u}_d$ is $(u_{d,1}, \ldots, u_{d,k})^T$. Matrices $\boldsymbol{A}$ have *n* rows and *m* columns, which called "frequency of word *w* in a document *d*". Also, the vector at represent the *t* columns of vector of matrices $\boldsymbol{A}$. The vector at is $(a_{1,d}, \ldots, a_{n,d})^T$. In addition, the vector $\boldsymbol{a}_d^{\mathrm{T}}$ represent the transposition of the d-th rows of vector of matrices $\boldsymbol{A}$. The Vector ad is $(a_{d,1}, \ldots, a_{d,m})^T$.

Next, Degree of concentration of words is described. ITA uses the kurtosis of higher order statistics as a measure to assess the independence between topics. The kurtosis of higher order statistics is the difference of the fourth moment from the normal distribution with the same mean and variance. Degree of concentration of words of a topic using kurtosis is defined as follows.

Degree of Concentration of Words

$$\sum_{w}^{m}\left(v_{w,t}^{4}P(w)\right)-3\left(\sum_{w}^{m}v_{w,t}^{2}P(w)\right)^{2}$$

Where $v_{w,t}$ represents the elements w rows and t columns of matrix $\boldsymbol{V}$. Also, $P(w)$ represents the probability of occurrence of the word w in all documents. $P(w)$ is defined as follows.

The Probability of Occurrence of the Word *w*

$$P(w)=\frac{\sum_{d}^{n}a_{d,w}}{\sum_{d,w}^{n,m}a_{d,w}}$$

Where, $a_{d,w}$ represents the elements d rows and w columns of matrix $\boldsymbol{A}$. If the word concentration is high, the importance of many words and documents will approach 0. Therefore, the topic is expressed by a small number of high importance words and documents. ITA can extract independent topics because it finds $\boldsymbol{V}$ that maximizes the degree of concentration of words in topics from document data. Also, each topic satisfies orthonormality.

Then, the algorithm of ITA is described. ITA is formulated as an optimization problem as follows.

$$\max_{\boldsymbol{R}}\left|\sum_{t}^{k}\left\{\sum_{w}^{m}\left((\mathbf{VR})_{\cdot}^{4}P(w)\right)-3\left(\sum_{w}^{m}(\mathbf{VR})_{\cdot}^{2}P(w)\right)^{2}\right\}\right|$$

Subject to $\mathbf{R}^{\mathrm{T}}\mathbf{R}=1, ||\mathbf{R}||=1$

Where, $(\boldsymbol{VR}).^{4}$ is fourth power of each component of the matrix $\boldsymbol{VR}$.

First, perform singular value decomposition on the matrix $\boldsymbol{A}$. From the result of singular value decomposition, a k-dimensional space with the importance matrix of each topic $\boldsymbol{V},\boldsymbol{U}$ as the coordinate axis is determined, and each document and word are arranged. Also, in this algorithm, the number of topics k to extract is determined by the user. Here, the words contained in the document are located near the document. Next, find the location that maximizes the independence of each topic using independent component analysis.

The algorithm of independent topic analysis as follows.

1) Make a matrix $\boldsymbol{A}$ which is frequency of words in the document. Then the matrix $\boldsymbol{A}$ in the same way as FPICA is normalized and $\widetilde{\boldsymbol{A}}$is calculated.

2) Perform singular value decomposition on the matrix $\widetilde{\boldsymbol{A}}$and decompose the matrix $\widetilde{\boldsymbol{A}}$ as follows.

$$\widehat{\boldsymbol{U}}^T\widetilde{\boldsymbol{A}}\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{S}}$$

Where, $\widehat{\boldsymbol{S}}$ is a diagonal matrix of singular values.

3) Make matrices $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$ from matrices $\widehat{\boldsymbol{U}}$, $\widehat{\boldsymbol{S}}$, $\widehat{\boldsymbol{V}}$. each matrix is created by extracting *k* components in descending order of the value of matrix $\widehat{\boldsymbol{S}}$.

4) Define a matrix $\boldsymbol{X}$ that represents a topic in the k-dimensional space as follows. Where, the matrix $\boldsymbol{X}$ have k rows and m columns.

$$\mathbf{X} = \mathbf{S}^{-1/2}\mathbf{U}^T\widetilde{\mathbf{A}}$$

5) Maximize independence between each topic.

Calculate the rotation matrix $\boldsymbol{R}$ to maximize the independence of the matrix $\boldsymbol{X}$. This rotation matrix $\boldsymbol{R}$ is calculated based on FPICA:

A) Initialize the rotation matrix $\boldsymbol{R}$ as a zero matrix. Where the rotation matrix $\boldsymbol{R}$ have *k* rows and *m* columns.

$$\boldsymbol{R} = \boldsymbol{0}$$

Assign $\boldsymbol{e}_t$ to the *t* column of the rotation matrix $\boldsymbol{R}$. Where, the vector $\boldsymbol{e}_t$ is a vector of the *t* column ($t \in \{1, \dots, k\}$) of the identity matrix $\boldsymbol{I}$

$$\boldsymbol{I} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_k), \qquad \boldsymbol{r}_t = \boldsymbol{e}_t$$

Also, $\boldsymbol{e}_1 = (1, 0, \dots, 0)^T, \boldsymbol{e}_2 = (0, 1, \dots, 0)^T$. And vector $\boldsymbol{e}_t$ have *k* elements.

B) Initialize a vector $\boldsymbol{r}^{(old)}$. Assign the $k \times 1$ zero vector to the vector $\boldsymbol{r}^{(old)}$

$$\boldsymbol{r}^{(old)} = (0, 0, \dots, 0)^T$$

C) Update the $\boldsymbol{r}_{\mathrm{t}}$ as follows.

$$\boldsymbol{r}^{(old)} = \boldsymbol{r}_t \;, \; \boldsymbol{r}_t = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{r}_{\mathrm{t}}).^3 - 3\,\boldsymbol{r}_{\mathrm{t}}$$

Where $(\boldsymbol{X}^T \boldsymbol{r}_t).^3$is cube of components of the $\boldsymbol{X}^T \boldsymbol{r}_t$.

D) Rotate the vector $\boldsymbol{r}_t$ as follows

$$\boldsymbol{r}_t = \boldsymbol{r}_t - \boldsymbol{R}\boldsymbol{R}^T \boldsymbol{r}_t, \qquad \boldsymbol{r}_t = \boldsymbol{r}_t/||\boldsymbol{r}_t||$$

E) If it converges under the same condition as FPICA, go to step (6). If it does not converge, go to step (5C).

6) Calculating the matrix $\boldsymbol{V}^*$ and $\boldsymbol{U}^*$ as follows.

$${}^*\boldsymbol{V} = \boldsymbol{V}\boldsymbol{R}, \; {}^*\boldsymbol{U} = \boldsymbol{U}\boldsymbol{R}$$

Thus, it is possible to extract highly independent topics, such as is shown in the Fig.1.

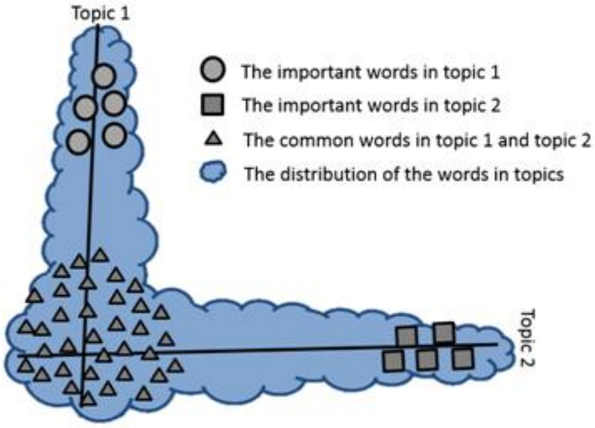
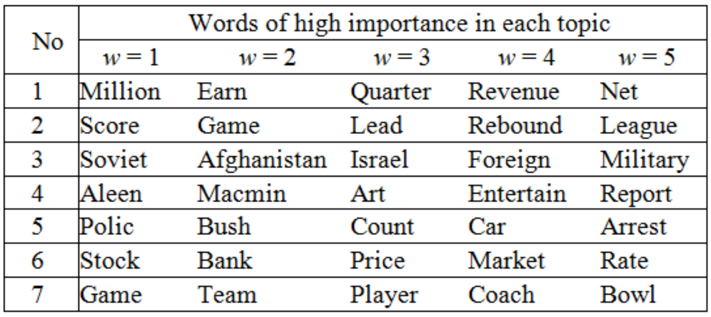

Topic 1
○ The important words in topic 1
■ The important words in topic 2
△ The common words in topic 1 and topic 2
The distribution of the words in topics
Topic 2

Fig 1. Image of ITA: the distribution of the words in Topic1 and Topic2.

Next, the topics obtained by independent topic analysis is described. For example, ITA is applied to Los Angeles Times (LA Times) data [15] as benchmark data. And Table 1 show that extracted 7 topics by ITA. In Table 1 show that the topic 1 indicates “Revenue”, the topic 2 indicates “Soccer”, the topic 3 indicates “Foreign”, the topic 4 indicates “Entertainment”, the topic 5 indicates “Affairs”, the topic 6 indicates “Stock”, and the topic 7 indicates “Team, Player”. The words in Table 1 is indicated the big absolute values of the column component of the $\boldsymbol{V}$. The topic of the independent topic analysis is represented by multiple high importance words.

Table 1. Examples of the Results of ITA (k = 7)



No	Words of high importance in each topic				
	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
1	Million	Earn	Quarter	Revenue	Net
2	Score	Game	Lead	Rebound	League
3	Soviet	Afghanistan	Israel	Foreign	Military
4	Aleen	Macmin	Art	Entertain	Report
5	Polic	Bush	Count	Car	Arrest
6	Stock	Bank	Price	Market	Rate
7	Game	Team	Player	Coach	Bowl

At present, ITA is a topic analysis method that extracts independent topics from one analysis data. ITA allows the user to grasp the topic included in the document data. However, although the topic of each document data can be grasped, the time series information is not expressed. Because time series documents are represented by a lot of time, the topic from one analysis data is only one of them.

So, from the results of ITA, it is not easy to understand and track the topic. Therefore, there is a problem that it does not correspond to time series documents such as news data. In order to perform topic tracking, it is necessary to analyze the topic of each period and the change of the topic with time. As described in the section 2.1, there is related research on topic tracking of time series documents. However, methods of related research cannot be used directly on topics extracted by ITA. In addition, there are many studies on ITA[14][16][17], but there is no research on topic tracking using independent topic analysis. Therefore, there is a need for a way to analyze time-series documents and track topics using ITA.

3. TRACKING METHOD FOR INDEPENDENT TOPIC ANALYSIS

In this section describes the topic tracking method. In this paper a method to analyze and track the topic transition of time series documents using ITA is proposed. At present, ITA is a method to analyze topics from one analysis data. Therefore, it is not easy to analyze the transition of topics in time series documents. However, ITA can analyze topics for each period with time series information. For example, if the news data of January 1, 2017 and January 2, 2017 are each one analysis data, ITA can be extracted each topic of each analysis data. Authors considered that the transition of topics could be analyzed from the similarity between topics in each period. So, executing the ITA multiple times analyze the transition of the topic and calculating the similarity between the extracted topics. Based on topic transition of time series documents, a method to track topic content and appearance timing was proposed. In the proposed methods, topic tracking is performed at the following four points.

- Similarity determination between the topic
- Change of topic with time change
- Visualization of topic transition
- Topic tracking method

First, in section 3.1 is described similarity determination between the topic. Second, in section 3.2 is described change of topic with time change. Next, in section 3.3 is described visualization of topic transition in section 3.3. Finally, in section 3.4 is described the topic tracking algorithm.

3.1. Similarity Determination between the Topic

In this section describes the similarity judgment between topics. In the proposed method, the similarity between topics is calculated in the adjacent period to visualize the topic transition. First, analyze the topics of each period and extract the Matrix $\boldsymbol{V}$using independent topic analysis. Create a vector (topic vector) representing each topic from Matrix $\boldsymbol{V}$. Where, the topic vector has N words with large absolute values of Matrix $\boldsymbol{V}$ as elements. That is, the topic vector excludes the influence of general words not directly related to the content of the topic or words with a low frequency of occurrence. Therefore, the similarity between topics becomes clearer.

Next, calculate the similarity of topic vectors in adjacent periods. In this paper uses cosine similarity, which is used to calculate vector similarity, as an evaluation index. The topic vectors for period τ and period $\tau - 1$are as follows.

$$\boldsymbol{V}_i^{(\tau)}(i \in (1, ..., k)), \quad \boldsymbol{V}_j^{(\tau-1)} \ (j \in (1, ..., k))$$

The cosine similarity $\cos(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)})$ of the topic vectors $\boldsymbol{V}_i^{(\boldsymbol{\tau})}$ and $\boldsymbol{V}_j^{(\tau-1)}$ is as follows.

$$\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) = \frac{\boldsymbol{V}_i^{(\tau)}\boldsymbol{V}_j^{(\tau-1)}}{|\boldsymbol{V}_i^{(\tau)}| \cdot |\boldsymbol{V}_j^{(\tau-1)}|}$$

The value of cosine similarity is the cosine value of the angle between vectors. Therefore, the cosine similarity is $0 \leq \cos(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}) \leq 1$. Also, as the cosine similarity value approaches 1, the two topic vectors are more similar.

In the proposed method, cosine similarity is calculated for all combinations of topic vectors in adjacent periods. And extract topic vector pair with similarities exceeding threshold ξ among calculation results. Where, this topic vector pairs calls similar topics. Also, the threshold S for similarity determination is determined by the user. If topic i in period τ and topic j in period $\tau - 1$ are similar topics, the following equation holds.

$$\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) > \xi$$
$$i, j \in \ (1, \dots, k)$$

3.2. Change of Topic with flow of the time

In section3.1 describes similarity determination between two topics. However, similar topics alone are not enough for topic tracking. In this section, we describe four more topic transitions in the similar topics. First, there are Topic integration and Topic division that appear in adjacent periods. In addition, there are Topic revival and New topic that appear in consideration of nonadjacent periods. These four topic transitions below is described.

(1) Topic Integration

Topic Integration that appears in adjacent time periods $(\tau, \tau - 1)$ is described. Topic Integration means that two or more topics of period $\tau - 1$ are similar the topics $\boldsymbol{V}_j^{(\tau)}$ of period τ in adjacent periods. If topic i in period τ and topic j, m in period $\tau - 1$ are Topic Integration, the following equation holds.

$$\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) > \xi$$
$$\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_m^{(\tau-1)}\right) > \xi$$
$$i, j, m \in \ (1, \dots, k), j \neq \ m$$

Also, Topic Integration means that the number of topics in the entire document is to decrease.

(2) Topic Division

Topic Division that appears in adjacent time periods $(\tau, \tau - 1)$ is described. Topic Division means that two or more topics of period τ are similar the topics $\boldsymbol{V}_j^{(\tau-1)}$ of period τ − 1 in adjacent periods. If topic i, m in period τ and topic j in period $\tau - 1$ are Topic Division, the following equation holds.

$$\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) > \ \xi$$
$$\cos\left(\boldsymbol{V}_m^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) > \xi$$
$$i, j, m \in \ (1, \dots, k), i \neq \ m$$

Also, Topic Division means that the number of topics in the entire document is to increase.

(3) Topic Revival

Topic Revival that appear in consideration of nonadjacent periods $(\tau, \tau - 1)$ is described. Topic Revival means that topics that have disappeared once will reappear. Therefore, there is no similar topics to the adjacent period. However, there are similar topics to some period in the past. If topic i in period τ and topic m in period $\mathrm{\tau} - l, \left(l \in (2, \dots, \mathrm{\tau} - 1)\right)$ are Topic Revival, the following equation holds. Where, period l represents the past period that is not adjacent to the period τ.

$$
\begin{gathered}
\cos\left(\mathbf{V}_i^{(\tau)}, \mathbf{V}_j^{(\tau-1)}\right) < \xi \\
\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_m^{(\tau-l)}\right) > \xi \\
i, j, m \in (1, \dots, k), l \in (2, \dots, \tau - l)
\end{gathered}
$$

In addition, Topic Revival is seasonal topics, regular events such as the Olympics, and irregular events such as elections are extracted.

(4) New Topic

New Topic that appear in consideration of nonadjacent periods $(\tau, \tau - 1)$ is described. New Topic means that there is no similar topic in the past. Therefore, there is no similar topic in the adjacent period $(\tau, \tau - 1)$and the past period $\tau - l, \left(l \in (2, \dots, \tau - 1)\right)$. If topic i in period τ is New Topic, the following equation holds.

$$
\begin{gathered}
\cos\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) < \xi \\
\cos\left(\mathbf{V}_{\mathrm{i}}^{(\tau)}, \mathbf{V}_{\mathrm{m}}^{(\tau-\mathrm{l})}\right) < \xi \\
i, j, m \in (1, \dots, k), l \in (2, \dots, \tau - l)
\end{gathered}
$$

Also, New Topics often include new characters, words, and information.

3.3. Visualization of Topic

In this section describes visualization of topic transition. Perform visualization as follows.

Place the topics of each period extracted by multiple times ITA on the time axis. Therefore, the topic of each period can be grasped. At the time of place, it is displayed together with a label indicating the contents of the topic. Where, the label is the word with the largest absolute value of the importance in each topic. Calculate the similarity between topics in a certain period and topics in adjacent periods. And link the pairs of topics whose similarity is higher than the threshold. Also, topics that link in adjacent periods set the same color. Therefore, the transition of the topic can be grasped in the adjacent period. If there is no topic to link in the adjacent period, topic have white. Also, new topics set new colors.

3.4. Topic Tracking Method

In this section describes topic tracking method using ITA. In this paper, a topic tracking method related to the keyword selected by the user was proposed. First, the degree of keyword relevance to a topic uses the importance of the topic extracted by ITA. A high degree of importance means that the relation is strong and the influence on the topic is large. Also, it is assumed that there are multiple topics related to the keyword. In related research, one keyword was treated as one topic (theme). However, one keyword may be related to multiple topics. Thus, proposed method is assumed that there are multiple topics related to the keyword.

From the above, in this paper considered topic tracking method that can track each topic associated with the keyword and the period each topic occurred. The algorithm of topic tracking method using ITA is as follows.

1) Select the words that the user wants to track.
2) Extract matrix $\boldsymbol{V}$ that is the importance of the word in the topic in each period using ITA

multiple times.

3) From the topic extracted in (2), extract the topic with the importance of keyword is 1 or more. This is a process for deleting topics not related to words.
4) Calculate the similarity of topics in adjacent periods. If there are no similar topics in adjacent periods, search until similar topics are found in the past period.
5) Extract the importance of keyword in each period. If there is no topic with the importance of keyword is 1 or more in the period, the importance is 0.
6) Place the importance of keyword for each topic link on the time axis.

By this method, it is possible to analyze each topic associated with the keyword and the period each topic occurred.

4. EXPERIMENTS

In this section describes an experiment of the proposed topic tracking method.

4.1. Dataset

In the experiment was used the dataset that CD-Mainichi Shimbun data collection. Where, the Mainichi Shimbun is one of the Japanese newspapers. CD-Mainichi Shimbun data collection is the data provided by Mainichi Shimbun and sold by Nichigai Associates. The dataset used four years data from 2014 to 2017 in CD-Mainichi Shimbun data collection.

Next, the size of the dataset is described. The number of articles is 150 to 200 per day. In addition, the number of articles is about 100,000 per year. Each article contains article ID, article keyword, title, text, time series information, and insertion surface code. In the experiment was used time series information and the text of each article for topic tracking.

4.2. Experimental Setup

In this section describes the Experimental conditions. Where, there are two experimental conditions. First is condition of making matrices $\boldsymbol{A}$ which is frequency of word in a document. Second is a threshold setting to determine the similarity between topics.

(1) Make Matrices A

As stated in section 2, in ITA, document data cannot used as it is. In ITA, makes matrices $\boldsymbol{A}$which is frequency of word in a document from the data set and uses it for analysis. Therefore, divide the daily document data into words and create a frequency matrix. Where, in make matrices $\boldsymbol{A}$, Japanese morphological analysis is used to divide a document into words. In this experiment was used Janome, which is one of the common Japanese morphological analysis software. In addition, it was used the dictionary called Nikkei Thesaurus. Where, the Nikkei Thesaurus is an IPA dictionary used in Japanese morphological analysis and is often used in newspaper data analysis. ITA is a method in which there is no default value setting, and even if analysis is performed multiple times on one analysis data, the topic extracted does not change. Therefore, the topic to be extracted changes greatly depending on the words included in matrices $\boldsymbol{A}$. In the experiment, matrices $\boldsymbol{A}$ used only nouns such as general nouns and proper nouns among morpheme analysis results.

(2) Similarity index between Topics

In this experiment, the similarity of topics is compared using the following three distance: Euclidean distance and Mahalanobis distance, Cosine distance.

Euclidean distance and Mahalanobis distance are defined as follows. Euclidean distance similarity was defined, which is used to calculate vector similarity, as an evaluation index. The topic vectors for period τ and period $\tau - 1$ are as follows.

$$\boldsymbol{V}_i^{(\tau)}(i \in (1, ..., k)), \boldsymbol{V}_j^{(\tau-1)} \ (j \in (1, ..., k))$$

The Euclidean distance similarity $Euc\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right)$ of the topic vectors $\boldsymbol{V}_i^{(\tau)}$ and $\boldsymbol{V}_j^{(\tau-1)}$ is as follows.

$$Euc\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) = \left\|\boldsymbol{V}_i^{(\tau)} - \boldsymbol{V}_j^{(\tau-1)}\right\|^2$$

Also,$\boldsymbol{V}_i^{(\tau)}$ and $\boldsymbol{V}_j^{(\tau-1)}$are similar when the Euclidean distance similarity is smaller.

Next, Mahalanobis distance similarity was defined, which is used to calculate vector similarity, as an evaluation index.

The Mahalanobis distance similarity $Maha\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right)$ of the topic vectors$\boldsymbol{V}_i^{(\tau)}$ and $\boldsymbol{V}_j^{(\tau-1)}$is as follows.

$$Maha\left(\boldsymbol{V}_i^{(\tau)}, \boldsymbol{V}_j^{(\tau-1)}\right) = \left|\frac{\boldsymbol{V}_i^{(\tau)} - \mu_i}{\sigma_i^2} - \frac{\boldsymbol{V}_j^{(\tau-1)} - \mu_j}{\sigma_j^2}\right|$$

Where μ_i is average of $\boldsymbol{V}_i^{(\tau)}$, σ_i^2 is variance of $\boldsymbol{V}_i^{(\tau)}$. And μ_i is average of $\boldsymbol{V}_j^{(\tau-1)}$, σ_j^2 is variance of $\boldsymbol{V}_j^{(\tau-1)}$. Also, $\boldsymbol{V}_i^{(\tau)}$and $\boldsymbol{V}_j^{(\tau-1)}$ are similar when the Mahalanobis distance similarity is smaller.

(3) Threshold Setting

In topic tracking method, determines the similarity between topics. And the threshold for similarity determination is determined by the user. If the threshold is low, even between slightly similar topics, it is determined that there is similarity. Also, if the threshold is high, it is determined that there is similarity only between similar topics. In other words, if the threshold is high, topic transition such as topic separation and integration may become difficult to detect. In addition, the appearance of new topics is increased by reducing similar topics. In the experiment, the user confirms topics from various angles and sets a threshold ξ based on the user's evaluation.

4.3. Visualization of Topic Transition

Visualization results of the topic transition using ITA is described. Fig 2 shows the topics transition in datasets. The topic extracted on each day is placed on the time axis. Where, the time axis represents the transition of time from left to right. At the time of arrangement, it is displayed together with a label indicating the contents of the topic. From Fig 2 can grasp Topic transition, Topic integration, Topic division and New topic. In addition, it is possible to grasp the transition of the topic by the link between the adjacent topics.

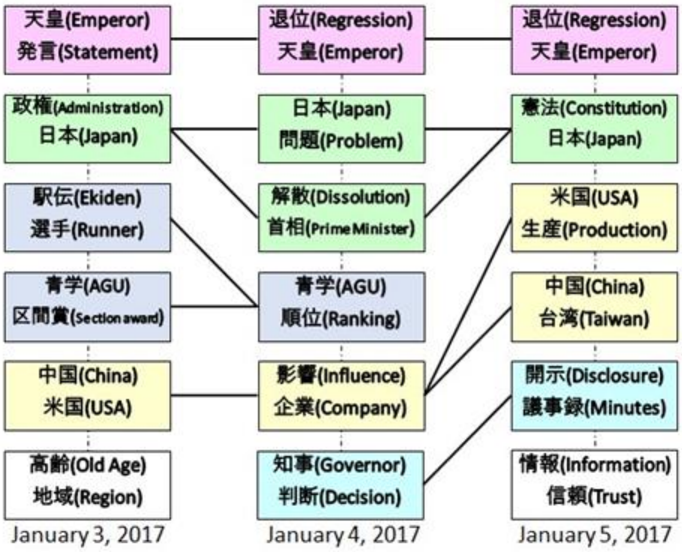
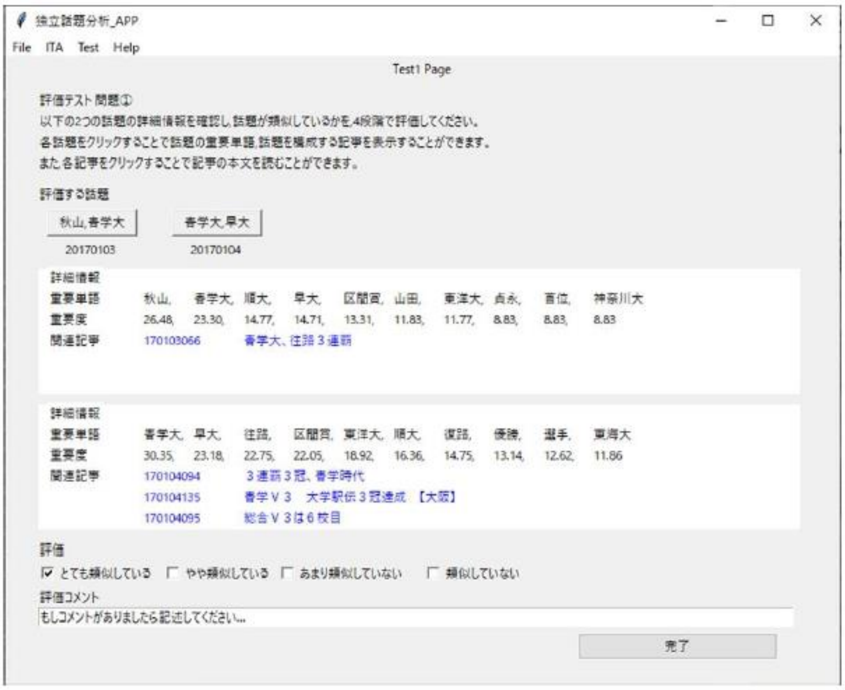


Fig 2. Visualization of Topic Transition

4.4. Similarity index between Topics by user experiments

In order to conduct user experiments, a user interface as shown in the Fig. 3 was created and experiments were conducted with six participants. In the experiment was investigated which index was the most effective using topic similarity index. The topics were arranged in descending order of cosine similarity and the participants see the topic. The participants choose whether the two topics are similar or not similar. The Fig. 3 shows the interface created in the experiment. In this figure, the participants can see important words and important value of two topics. In addition, this figure includes headlines for related articles. The participants click the related articles heading to see the articles as shown Fig. 4.



Test1 Page

評価テスト 問題①

以下の2つの話題の詳細情報を確認し,話題が類似しているかを,4段階で評価してください。
各話題をクリックすることで話題の重要単語,話題を構成する記事を表示することができます。
また,各記事をクリックすることで記事の本文を読むことができます。

評価する話題

秋山,青学大	青学大,早大
20170103	20170104

詳細情報

重要単語	秋山,	青学大,	順大,	早大,	区間賞,	山田,	東洋大,	貞永,	首位,	神奈川大
重要度	26.48,	23.30,	14.77,	14.71,	13.31,	11.83,	11.77,	8.83,	8.83,	8.83
関連記事	170103066	青学大、往路3連覇								

詳細情報

重要単語	青学大,	早大,	往路,	区間賞,	東洋大,	順大,	復路,	優勝,	選手,	東海大
重要度	30.35,	23.18,	22.75,	22.05,	18.92,	16.36,	14.75,	13.14,	12.62,	11.86
関連記事	170104094	3連覇3冠、青学時代								
	170104135	青学V3 大学駅伝3冠達成 【大阪】								
	170104095	総合V3は6校目								

評価

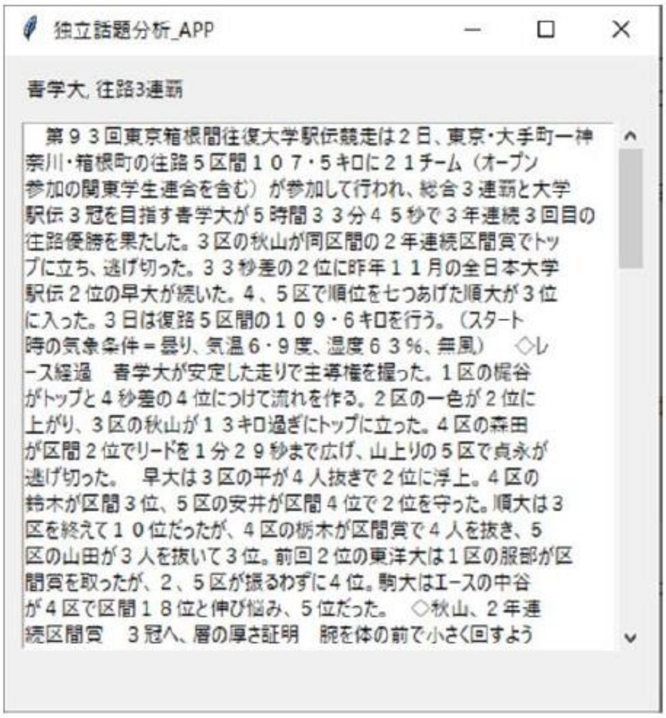
☑ とても類似している ☐ やや類似している ☐ あまり類似していない ☐ 類似していない

評価コメント

もしコメントがありましたら記述してください...

完了

Fig 3. Example user interface used in experiment: Topic browsing



青学大、往路3連覇

第９３回東京箱根間往復大学駅伝競走は２日、東京・大手町－神奈川・箱根町の往路５区間１０７・５キロに２１チーム（オープン参加の関東学生連合を含む）が参加して行われ、総合３連覇と大学駅伝３冠を目指す青学大が５時間３３分４５秒で３年連続３回目の往路優勝を果たした。３区の秋山が同区間の２年連続区間賞でトップに立ち、逃げ切った。３３秒差の２位に昨年１１月の全日本大学駅伝２位の早大が続いた。４、５区で順位を七つあげた順大が３位に入った。３日は復路５区間の１０９・６キロを行う。（スタート時の気象条件＝曇り、気温６・９度、湿度６３％、無風）　◇レース経過　青学大が安定した走りで主導権を握った。１区の梶谷がトップと４秒差の４位につけて流れを作る。２区の一色が２位に上がり、３区の秋山が１３キロ過ぎにトップに立った。４区の森田が区間２位でリードを１分２９秒まで広げ、山上りの５区で貞永が逃げ切った。　早大は３区の平が４人抜きで２位に浮上。４区の鈴木が区間３位、５区の安井が区間４位で２位を守った。順大は３区を終えて１０位だったが、４区の栃木が区間賞で４人を抜き、５区の山田が３人を抜いて３位。前回２位の東洋大は１区の服部が区間賞を取ったが、２、５区が振るわずに４位。駒大はエースの中谷が４区で区間１８位と伸び悩み、５位だった。　◇秋山、２年連続区間賞　３冠へ、層の厚さ証明　腕を体の前で小さく回すよう

Fig 4. Example user interface 2 used in experiment: Article browsing

The Table. 2 show the important words in two topics. The important words of the Table.2 left are high cosine similarity. It is two topics that the user felt in the experiment was similar. The important words of the Table.2 right are low cosine similarity. It is two topics that the user felt in the experiment was not similar. As a result of the experiment, it was found that even if the cosine similarity value is around 0.7, it is judged that it is similar depending on the topic.In the future, authors plan to create a user interface that allows users to learn thresholds and change them flexibly.

Table 2. The important word in two topics with similar case and dissimilar case

	Similar Topics (high cosine similarity)		Dissimilar Topics (low cosine similarity)	
No. word	Topic i	Topic j	Topic i	Topic j
$w = 1$	North Korea	North Korea	China	U.S.A.
$w = 2$	U.S.A.	U.S.A.	U.S.A.	China
$w = 3$	Remark	Correspondence	Trade war	Japan
$w = 4$	Emphasis	Remark	Specified	Foreign
$w = 5$	Intercept	Attack	Chinese products	Hozumi

4.5. Topic Tracking

Experiment of topic tracking method related to the keyword selected by the user is described. In the experiment, the keyword to track is “AGU: Aoyama Gakuin University”. The comparison method uses the analysis method of Themeriver. In Themeriver, the topic was visualized using the number of occurrences of keywords. Therefore, the transition of the number of occurrences of keywords is used as a comparison method.

(1) The Transition of the Number of Word Occurrences

Create a graph of the number of word occurrences. First, count the number of times the document data for each day contains the keyword. If the data for a certain day does not contain the keyword, the number of occurrences for that day shall be 0. Fig. 5 shows the transition of the number of occurrences AGU. From Fig. 5, it is possible to roughly grasp the number of times keyword has become a document. In addition, it can grasp period when the keyword appeared.

However, in the transition of the number of occurrences of the word, it is not possible to grasp what topic the word is used in. Therefore, in the transition of the number of word occurrences, only the occurrence period of the topic can be grasped.

(2) Topic Tracking using ITA

Create a graph of topic tracking results using ITA. First, extract the importance of the keyword of each topic of each day using ITA. Next, each topic based on the similarity between topics and graph the importance is classified. Fig 6 show the transition of the importance of the keyword in each topic. Similar to the number of word occurrences, Fig. 6 makes it possible to grasp at what timing the topic appears. In addition, it is possible to analyze what kind of topics are included by incorporating the similarity between topics using ITA. In fact, from Fig. 6, it is possible to grasp the inclusion of Athletics, Baseball, Election and other topics, and the timing at which each topic appeared. For example, the topic of January, which has a high degree of importance, is an athletic topic.

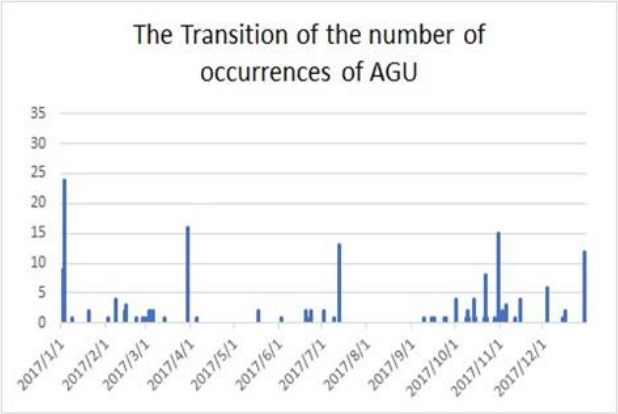


Fig 5. The Transition of the Number of Occurrences of AGU

(3) Results

Themeriver (The transition of the number of word occurrences) shows the transition of the number of keyword occurrences and period. In the transition of the number of word occurrences, one keyword was considered as one topic. Therefore, Fig. 5 shows when keyword occurred. However, From Fig. 5 cannot analyze the content of the topic related to the keyword. Also, the relevance of topics cannot be analyzed as well. And, it cannot analyze how much the keyword is affecting the topic.

On the other hand, in topic tracking result using ITA, it is assumed that there are multiple topics related to the keyword. Therefore, Fig. 6 shows each topic that is associated with the keyword and the period each topic occurred. From Fig 6 can analyze the content of the topic related to the keyword. In fact, Fig. 6 shows four topics related to keywords and each topic appearance periods. Since topic tracking using ITA can calculate the similarity between topics that are not consecutive periods, it is possible to confirm the resurgence of topics. Therefore, topic tracking using ITA can analyze topics that appear only during a specific period, such as regular events such as the Olympics and irregular events such as elections. The methods using Themeriver and clustering do not support topic tracking with non-consecutive appearance periods. Therefore, a topic tracking method using ITA that can analyze the resurgence of topics is effective. Table 3 shows the results of the topic tracking evaluation experiment.

Also, there are cases where even if the number of word occurrences is large, it cannot be taken up. For example, consider March 30 of the AGU. The number of word occurrences on March 30 in Fig. 5 is 17 times, the second highest number in 2017. On the other hand, the value on March 30 in Fig. 4 is not large. The articles where AGU on March 30 appears are extracted and shown in Table 3.

Table 3 shows that many of the articles are professional baseball player directories. Where, a professional baseball player directory is an article that introduces Japanese baseball players by team. AGU appeared in the article as a school from which professional baseball players came from. In other words, it was found that the topic of AGU did not appear on March 30. From the above results, it was shown that topic tracking using ITA can extract cases where words do not become topics even when they appear. Since AGU has few word occurrences, it is difficult to confirm the topic integration and the topic separation. By using words with a large number of occurrences for topic tracking, it is also possible to change topic transition.

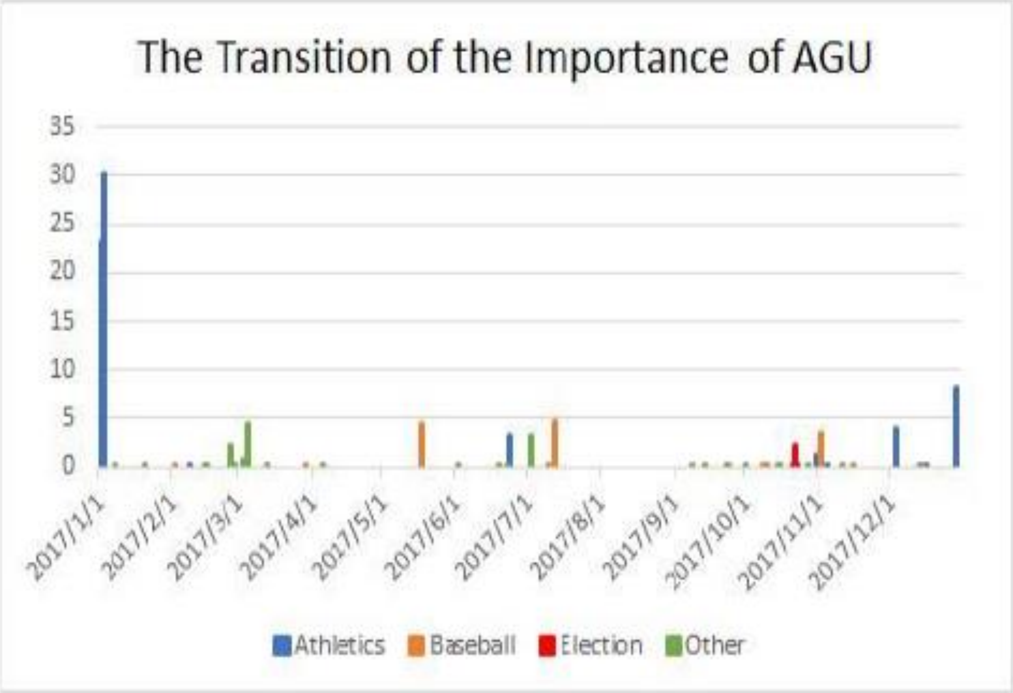
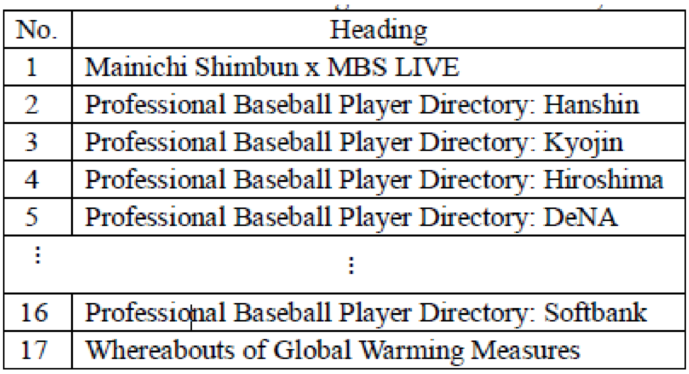


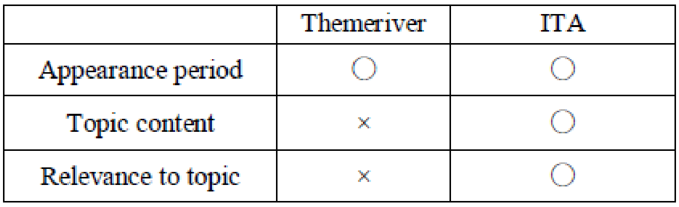
Fig 6. The Transition of the Importance of AGU

Table 3. Articles Containing “AGU” on March 30, 2017



No.	Heading
1	Mainichi Shimbun x MBS LIVE
2	Professional Baseball Player Directory: Hanshin
3	Professional Baseball Player Directory: Kyojin
4	Professional Baseball Player Directory: Hiroshima
5	Professional Baseball Player Directory: DeNA
⋮	⋮
16	Professional Baseball Player Directory: Softbank
17	Whereabouts of Global Warming Measures

Table 4. Features of each method



	Themeriver	ITA
Appearance period	○	○
Topic content	×	○
Relevance to topic	×	○

5. Conclusion

In this paper was proposed a topic tracking method using independent topic analysis. For problem where independent topic analysis did not adapt to time-series documents, topics of each period were extracted, and topic transition of time series documents was analyzed based on the similarity of topics. In addition, this paper was defined the following four topic transitions as Topic Integration, Topic Separation, Topic Revival and New Topic. Moreover, it was created a user interface to track the content and appearance period of the topic based on the topic transition of time-series documents and was obtained the effective result by the evaluation experiment. Table 4 shows the difference between the existing method and the proposed method. In Table 4, Themeriver can display the appearance period of a topic, but cannot know the topic content and relevance to topic. On the other hand, the proposed method can display the relationship between topic appearance period, topic content, and relevance to topic.

For our future work, authors plan the following. First, it is necessary to consider the method of setting the threshold. In this paper, the threshold for topic similarity determine set by the user. However, the threshold value significantly changes the extraction result of the topic transition. Therefore, a method that can be determined automatically by an algorithm is considered necessary. Second, we need to conduct user experiments for the evaluation of topic tracking by threshold change. Third, it is necessary to compare topic tracking method using ITA with methods other than Themeriver. Finally, the authors will compare the interface with other methods using many participants.

References

[1] Hofmann Tomas. (1999) "Probabilistic latent semantic analysis". Proceeding of the 15th Conference on Uncertainty in African Intelligence (UAI’99) , pp289–296.

[2] Blei M. David, Ng Y. Andrew, and Jordan I. Michael. (2003) "Latent dirichlet allocation". The journal of Machine Learning Research 3 , pp993–1022.

[3] Salton Gerard, Fox A. Edward and Wu Harry. (1983) "Extended boolean information retrieval". Common ACM 26, 11 , pp1022–1036.

[4] Blei M. David and Lafferty D. John. (2006) "Dynamic topic models". In Proceedings of the 23rd International Conference on Machine Learning (ICML '06). ACM, New York, NY, USA, pp113–120.https://doi.org/10.1145/1143844.1143859

[5] Deerwester Scott, Dumais T. Susan, Furnas W. George, Landauer K. Thomas and Harshman Richard. (1990) "Indexing by latent semantic analysis". the American Society of Information Science 40, 6 , pp391–407.

[6] Shinohara Satoshi. (2000) "Development of a document browsing support system for tracking key topics in document databases". Research report of Central Research Institute of Electric Power Industry Research report R 99036 , 1–4, 1-24. https://ci.nii.ac.jp/naid/40005003933/

[7] OjaErkki, HyvarinenAapo and KarhunenJuha. (2005) "Independent component analysis". Tokyo Denki University Press, Tokyo.

[8] Havre Susan, Hetzler Beth, and Nowell Lucy. (2000) "ThemeRiver: Visualizing theme changes over time". In Proceedings of the IEEE Symposium on Information Visualization, (INFOVIS '00). IEEE Computer Society, Washington, DC, USA, 115–. http://dl.acm.org/citation.cfm?id=857190.857680

[9] Havre Susan, Hetzler Elizabeth, Whitney Paul, and Nowell Lucy. (2002) "ThemeRiver: Visualizing thematic challenges in large document collection". IEEE Trans. Visualization and Computer Graphics 8, 1 , 9–20.

[10] Swan Russell and Allan James. (2000) "Automatic generation of overview timelines". In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00). ACM, New York, NY, USA, 49–56. https://doi.org/10.1145/345508.345546

[11] Everitt S. Braian, Landau Sabine, Leese Morven and Stahl Daniel. (2011) "Cluster Analysis (5th ed.)", Wiley Series in Probability and Statistics.

[12] James Allan (Ed.) (2002) "Topic Detection and Tracking: Event-Based Information Organization", Kluwer Academic Publishers, Nrwell.

[13] Mei Qiaozhuand ZhaiChengXiang. (2005) "Discovering evolutionary theme patterns from text: An exploration of temporal text mining". In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05). ACM, New York, NY, USA, pp198–207.

[14] Nishigaki Takahiro, Nitta Katsumi, Onoda Takashi. (2016) "Constrained independent topic analysis". Japanese Society for Artificial Intelligence Journal Vol.31, No. 4 , pp1–13. https://doi.org/10.1527/tjsai.D-FB1

[15] Zhong Shi and Ghosh Joydeep. (2003) "A comparative study of generative models for document clustering". Clustering High Dimensional Data and Its Application.

[16] Nishigaki Takahiro, Nitta Katsumi, Onoda Takashi. (2017) "Incremental Learning of Independent Topic Analysis", International Journal of Computer, Electrical Automation, Control and Information Engineering, Vol.11, No.2, pp.206-212.

[17] Nishigaki Takahiro, Nitta Katsumi, Onoda Takashi. (2018) "An Interactive Independent Topic Analysis for a Mass Document Review Service". The Review of Socionetwork Strategies, Vol. 12, No.1, pp. 47-69.

AUTHORS

Takahiro Nishigaki He was born in Japan 1987. He recieved the B. E. Degree from Kyoto Institute of Technology in 2011. He recieved the M. E. Degree from Tokyo Institute of Technology in 2013. He recieved the Dr.Eng Degree from Tokyo Institute of Technology in 2017. He is a member of the Japanese Society of Artficial Intelligence (JSAI). He has been an Assistant Professor with Aoyama Gakuin University since 2017.

Kenta Yamamoto He was born in Japan 1996.He recieved the B. S. Degree from Aoyama Gakuin Univercity in 2018. He has been in the master's program of Aoyama Gakuin University since 2018. His major field of study is topic analysis and text mining.

Takashi Onoda He was born in Japan in 1962. He received the B. S. Degree from International Christian University in 1986. He received the M. S. Degree from Tokyo Institute of Technology in 1988. He received the Dr. Eng. Degree from University of Tokyo in 2000.He was a Visiting Researcher in GMD FIRST (Fraunfoher FIRST) from 1997 to 1998. Since 2007, he has also been an Visiting Professor at Department of Computational Intelligence and System Science, Tokyo Institute of Technology, Japan. He is a member of JSAI (Japanese Society of Artficial Intelligence). He researched in Central Research Institute of Electric Power Industry from 1988 to 2015.He has been a Professor with Aoyama Gakuin University since 2016.