Vietnam J Comput Sci (2017) 4:147–160
DOI 10.1007/s40595-016-0084-y

REGULAR PAPER

Extracting user behavior-related words and phrases using temporal patterns of sequential pattern evaluation indices

Hidenao Abe[1]

Received: 6 November 2015 / Accepted: 19 September 2016 / Published online: 13 October 2016

Abstract The growth of social media sites, such as Twitter, which can provide a visual record of the daily interests and concerns of people in the form of their tweets and tweeting behaviors, has led to an increasing demand among enterprise users, to be able to identify those users who are interested in the services and products that these enterprises offer. However, accurately determining whether people who receive information, such as tweets, from enterprise users have a genuine interest in it can be difficult. In this study, a method for extracting feature words and phrases from the past users' tweets using temporal patterns of sequential pattern evaluation indices and phrase importance evaluation indices is developed. In this method, a variety of the followers interests are first analyzed using the feature words and phrases retweeted by the followers. Next, the temporal patterns of each evaluation index that are created based on the usage frequencies of feature words and phrases obtained from the historical followers' tweeting behaviors are extracted. An experimental result has shown that this method successfully extracted the sets of words and phrases based on the followers' tweeting behaviors as the temporal patterns for each evaluation index and the following retailer's account. These sets of words and phrases lead to understand the variety of the followers' interests with more clues.

Keywords Temporal text mining · Sequential pattern evaluation index · User behavior prediction

✉ Hidenao Abe
hidenao@shonan.bunkyo.ac.jp

[1] Department of Information Systems, Bunkyo University, 1100 Namegaya, Chigasaki, Kanagawa 2538550, Japan

1 Introduction

The recent growth of social media sites, such as Twitter, provides a media-based visual records of the talk and behavior of people, simply called users in this paper, which reflect their interests and concerns. Consequently, users, such as corporations and politicians, have begun looking for more efficient ways to communicate ideas with a large number of other users who have similar interests or concerns to theirs. As a result, many approaches have been developed to distinguish users having similar interests, using network analysis methods or/and text mining-based analysis methods [1].

However, because accurate identification of future user behavior without considering the user's speech and behavior history is difficult, there is a severe need to develop other methods that more accurately describe user's interests as feature words for predicting a targeted behavior such as retweeting. To this end, some studies attempted to predict users' information diffusions using frequencies of appearances of words and phrases, users' actions count, and other features related to the users' behavior in the past tweets of the users [2–4]. In [3], the method needs the emotional categories of the words that were constructed using an expensive tool. This prediction method cannot work with the Linguistic Inquiry and Word Count (LIWC) dictionary. On the other hand, in [4], they introduced the time-dependent features that obtained from the history of tweeting actions and not the usage history of the words. Regarding these previous studies, the problem of connecting the users' behavior with the content of the users' tweets seems unsolved.

Considering the above-mentioned issue, this study focuses on the temporal behaviors of the Twitter service known as "retweeting," in which users disseminate information by resending a previous tweet, and develops a method for extracting feature words and phrases that can predict retweet-

ing behavior using the content of the users' tweeting history. To tackle the problem of absence of the words in a prepared dictionary, we use the values of pattern evaluation indices that assume words and phrases as patterns. The use of values of the pattern evaluation indices separates the surface form of the words and phrases from their statistical nature. In addition, we should also solve the problem of preparing the categories of the words and phrases for characterizing users' tweeting behaviors. As for grouping words and phrases, we can use the temporal values of the pattern evaluation indices. By clustering the temporal values as the temporal patterns, the words and the phrases are categorized based on their statistical nature. On a set of temporal clusters, the centroids of the clusters represent the temporal patterns of anonymized words and phrases that are represented by the averaged temporal values and their ranges on each timestamp of the established pattern evaluation index.

In this paper, the proposed method first extracts the differences between groups of feature words and phrases contained in retweeted text and groups of feature words contained in the tweeting history of users believed to be interested in information from particular Twitter account holders, who are known as followers. Then, the temporal patterns of the words and phrases evaluation indices, calculated using the usage frequencies of feature words and phrases contained in the followers' past tweets, are obtained. These results are used to discuss the development of the method for constructing a model that predicts information-retweeting behavior using the temporal patterns of evaluation indices in the tweeting history instead of using the feature words and phrases appearances directly.

Although the original idea of this study and a part of the experimental results are described in our previous work [5], the purpose of this paper is to describe availability and efficiency of this method more methodologically, and to explain the availability and efficiency of the approach using additional results of the experiment. Adding to the experimental results, we also introduce a temporal trend analysis for utilizing the obtained temporal patterns as the features for constructing information-retweeting predictive models.

This paper is organized as follows. Section 2 describes the proposed method. Section 3 presents evaluation index group definitions used for feature word selection. In Sect. 4, the feature words and phrases extraction is performed to three well-known online retailers' Twitter accounts that had a substantial number of followers in Japan. The method extracts feature word groups contained in retweeted text and in the tweet histories of retweeting followers. It also generates temporal patterns of the indices that are used for choosing keywords and feature sequential patterns. For evaluating the stability of the proposed method, the method is applied to another set of tweets from the different period in Sect. 5. Finally, a conclusion is offered in Sect. 6.

2 Extraction of difference of retweeting users using temporal patterns in tweeting history

In this method, we assume that users' targeted retweeting behavior is affected not only by the content of received tweets but also their history of tweets. To construct a model for predicting such targeted tweeting behavior of followers, we should set up more proper features for considering the history of their tweets, which are obtained from their past tweeted content and actions.

In the text analysis, feature word extraction from a text corpus is a well-known method for obtaining the features from text in previously posted content. Then, a huge number of the feature words and phrases are often selected by the conventional methods, which are dependent on one particular evaluation index. However, it is very difficult to develop universal evaluation index on various context. In various situation, there is no trivial answer for evaluating usefulness of the feature words. In addition, feature word groups that are obtained using feature word extraction do not indicate when the information was obtained or their temporal trends. Therefore, we focus on patterns of change over time (temporal patterns), and developed for constructing a model that predicts the appearance of phrases using the temporal patterns of the evaluation indices of multiple phrases [6].

From the point of view that more various features can enable more explicit descriptions of hidden dependent variable relationships, it is not trivial that conventional features based on the appearance of feature sequential patterns may or may not be better predictors than temporal patterns [7]. Therefore, an improved method should use both the appearance of feature words and the phrases' temporal patterns, which were obtained from the user tweet history, as features to more accurately characterize the content history. Moreover, this method could also identify behaviors by similarly linking temporal patterns of the tweet counts and intervals.

The whole process of this method is described as the followings:

1. Finding features of users who are interested in an account.
   (a) Finding words and phrases used in the main concerned tweets by extracting as feature them from retweeted tweets of the account.
2. Extracting feature words and phrases in the past users' tweets as temporal patterns of the evaluation indices.
   (b) Extracting candidates of users' feature words and phrases.
   (c) Calculating evaluation index values of the candidates in the temporal corpus of the users' tweets.
   (d) Constructing temporal patterns of each evaluation index.

(e) Extracting sets of words and phrases included in each temporal pattern.
(f) Constructing the features for predictive models based on the temporal pattern of evaluation indices.

3. Finding concrete understandings of the difference of the interested users as predictive models.

In this paper, we implemented above step 1 and 2 using the tools for extracting the feature words and phrases. First, an automatic meaningful words and phrases extraction method obtained the candidates of the feature words and phrases. Then, the method calculates some sequential pattern evaluation indices and importance evaluation indices of words and phrases on the temporal corpus of the followers for obtaining temporal patterns of each evaluation index. After obtaining the temporal clusters, to utilize the clusters as the features for predictive model construction, we should know the meanings of the temporal patterns. For this issue, we calculate a temporal correlation between the overall averaged values on every time-point and each temporal pattern as the variance ratio (F-statistics) of the differences. According to the F-statistics, we determine whether we should use the levels of an index averaged value or the shapes of the index for each period using the F test. The variance ration is calculated as the following:

$$F\,(X, A) = \frac{\sum \left(\Delta x - \overline{\Delta x}\right)^2}{\sum \left(\Delta a - \overline{\Delta a}\right)^2},$$

where a centroid values of each temporal cluster consists of $\mathrm{X} = \{x_1, \ldots, x_n\}$, and the overall average of the evaluation index data set consists of $\mathrm{A} = \{a_1, \ldots, a_n\}$ within a period with n time-points. Δx denotes the differential between x_t and x_{t+1}, $1 \leq t \leq n-1$. Δa denotes the differential of the overall averages. $\overline{\Delta x}$ and $\overline{\Delta a}$ denote the average of the differentials of each temporal pattern and the differentials of the overall temporal values of the evaluation index, respectively. The F-statistics is test with the F-distribution under $n-1$ degree of freedom.

As for step 3, some actual extraction results show that this method can find the different set of words and phrases as the features of the interests of the followers.

3 Evaluation indices based on appearance frequency of words and sequential patterns

In the text, words and phrases[1] are represented by a series of one or more words. Given two words w_a and w_b in a sequential relationship, the order relation $a < b$ is always true, and a term term_i that is formed from these two words is expressed as $\mathrm{term}_i = \langle w_a, w_b \rangle$. Because of this phrase property, all sentences can be considered sequential data, having an ordered sequential relationship, and terms can be considered to be subsequences. With considering the common ordered relations of items in each data of the data set, both of a natural language processing-based phrase importance indices and sequential pattern evaluation indices can be used to evaluate phrases. The definitions of these indices are shown in Table 1.

3.1 Importance evaluation indices for words and phrases

Multiple importance evaluation indices have been developed for natural language processing and text mining to measure the importance of words and phrases for extracting features. The primary standard that these indices use is the appearance frequency of the word or phrases. Two appearance measurement references are commonly used for term appearance frequency: the term frequency (TF), which counts the number of times a term is repeated in one or more documents, and the document frequency (DF), which counts the number of documents in which the term appears.

Table 1 shows the typical evaluation indices for a term consisting of L words ($L \geq 1$), i.e., $\mathrm{term}_i = \langle w_1, \ldots, w_L \rangle$. The term frequency and inverse document frequency (TFIDF) method is most commonly used to evaluate the importance of the words and phrases for keyword extraction. It considers both the TF and DF and uses the ratio between the entire target document $|D|$ and the DF as a weight. A simple ratio that compares every pair of appearance frequency measurement standards can be used to index measuring properties based on appearance frequency.

3.2 Phrase evaluation indices using sequential pattern evaluation indices

Sequential pattern evaluation indices are indices that quantify multiple properties of sequential patterns using the appearance frequency $\mathrm{freq}(\alpha, D)$ of partial sequence α in sequential data set $D = \{s_i\}$, containing sequential data $s_i = \langle i_1, \ldots, i_m \rangle$, which are strings of items $i \in I$ that belong to item set I. Similar to the method used to determine keyword evaluation indices, two appearance frequency standards are typically used to count the appearance frequency of partial sequence $a = \langle i_1, \ldots, i_j \rangle$ $(j \leq m)$ in sequential data set D.

Applying the TF frequency standard, which considers repetitions in each document, and the DF frequency standard, which does not consider repetitions, a confidence-based index is defined using an evaluation index group for the non-

[1] Hereafter, words and phrases are called 'term.' Each term consists of one or more words.

Table 1 Importance evaluation indices and sequential pattern indices for each appearance frequency measurement standard for a term from [5]

	Frequency measurement standard	
	Document frequency $\mathrm{DF} = \lvert \mathrm{D}_{\in \mathrm{term}_i} \rvert$	Term frequency $\mathrm{TF} = \sum_j \mathrm{freq}(\mathrm{term}_i, d_j)$
Support	$\mathrm{DF}/\lvert \mathrm{D} \rvert$	$\mathrm{TF}/\sum \mathrm{TF}_{\mathrm{term}_i}$
Odds	$\mathrm{DF}/(\lvert \mathrm{D} \rvert - \mathrm{DF})$	$\mathrm{TF}/(\sum \mathrm{TF}_{\mathrm{term}_i} - \mathrm{TF})$
Self-information	$(\mathrm{DF}/\lvert \mathrm{D} \rvert) \log_2 (\mathrm{DF}/\lvert \mathrm{D} \rvert)$	$\mathrm{TF}/(\sum \mathrm{TF}_{\mathrm{term}_i}) \log_2 \mathrm{TF}/(\sum \mathrm{TF}_{\mathrm{term}_i})$
Jaccard coefficient	$\frac{\mathrm{DF}}{\mathrm{DF}(w_1 \cup \ldots \cup w_L)}$	$\frac{\mathrm{TF}}{\mathrm{TF}(w_1 \cup \ldots \cup w_L)}$
TFIDF	$\mathrm{TF} * \log(\lvert \mathrm{D} \rvert / \mathrm{DF})$	
Head confidence (H-Conf)	$\frac{\mathrm{DF}}{\mathrm{DF}(w_1, \mathrm{D})}$	$\frac{\mathrm{TF}}{\mathrm{TF}(w_1, \mathrm{D})}$
Max confidence (MaxConf)	$\max\left(\frac{\mathrm{DF}}{\mathrm{DF}(w_x, \mathrm{D})}\right)$	$\max\left(\frac{\mathrm{TF}}{\mathrm{TF}(w_x, \mathrm{D})}\right)$
All confidence (AllConf)	$\frac{DF}{\max(DF(w_x, D))}$	$\frac{TF}{\max(TF(w_x, D))}$
Sequential all confidence (SeqAIIConf)	$\frac{DF}{\max(DF(\beta \subseteq_{\mathrm{term}_i}, D))}$	$\frac{TF}{\max(TF(\beta \subseteq_{\mathrm{term}_i}, D))}$

sequential item set [8] and an evaluation index group that considers the items in a sequential pattern [9]. We consider a sequence α to be the term term_i and each item to be word w_x, where $1 \leq x \leq L$, within term_i.

As also shown in Table 1, when the sequential relationships between the items of a phrase's sequential pattern are considered, more than eight indices can be defined for the various confidences, which are the combined ratio of the appearance frequency of α and the appearance frequency of β, a subsequence of α.

4 Tweeting behavior analysis of online retail twitter accounts and their followers

This section examines user-sent texts (tweets), which contain 140 characters or less, obtained from a Twitter application programming interface (API) [10]. By gathering tweets both from some prominent account holders and from their followers tweets over the time, the relationships between the users' interests and concerns are analyzed as the features of resent tweets (retweets) that originate from the well-known Twitter accounts and of tweets sent by the retweeting users during a previous time period. To provide a broader analysis of general user interests, we define retweeting as the action of tweeting a feature word contained in a retweeted tweet.

The analysis procedure is as follows.

1. The feature words contained in tweets that were retweeted by some followers from a well-known Twitter account during a given period are extracted.
2. The followers that retweeted tweets including the feature words in (1) are listed.
3. The tweets sent by the followers in (2) during a time period prior to that of (1) are gathered.
4. The feature words contained in the tweets gathered in (3) and the temporal patterns of evaluation index groups based on their appearance frequencies are extracted.

The goal of this study is to develop a method for constructing a model that predicts information-based retweeting behavior using the temporal patterns of the tweeting history. Thus, the following analysis was performed as an application of this method using the procedure described above.

To obtain candidate phrases, or characteristic phrases, from the gathered text, an automatic terminology extraction method, used in natural language processing, is applied to each document set. In addition, to extract the phrases serving as feature word candidates, we used the FLR score-based automatic terminology extraction method developed by Nakagawa [11], which is defined as follows:

$$\mathrm{FLR}\,(\mathrm{CN}) = F\,(\mathrm{CN}) \times \left\{ \prod_{i=1}^{L} (\mathrm{FL}\,(N_i)+1)\,(\mathrm{FR}\,(N_i)+1) \right\}^{\frac{1}{2L}},$$

where $F(\mathrm{CN})$ denotes the frequency of the candidate (composed) noun CN. Each CN consists of one or more nouns N_i. $\mathrm{FL}(N_i)$ denotes the counts of the different nouns on the left-hand side in each bigram of N_i. Similar to the FL function, $\mathrm{FR}(N_i)$ denotes the counts of the different nouns on the right-hand side in the bigrams of N_i. The basic idea of the FLR score comes from the HITS algorithm [12], which measures the degree of the hubness of each node in a linked network structure. The calculated values for each CN are the geometric average of the differences in each sequenced noun, which corresponds to each node in the continued network structure. This indicates meaningfulness on the statistical linguistic property of each CN.

The nouns N_i used in the FLR score calculation were identified from morphological analysis results using MeCab [13]

Table 2 Number of retweeted tweets and number of FLR score-based candidate phrases in tweets sent from three major retailTwitter accounts between January 15 and 20, 2015 from [5]

Retailer	$\|D\|$	FLR score-based candidate phrases
7 Net shopping	76	369
Amazon.co.jp	193	857
Rakuten Ichiba	127	540

and the IPA[2] dictionary (mecab-ipadic-2.7.0-20070801) distributed with MeCab. Applying the FLR score calculation results, the candidate words and phrases were selected from phrases having FLR(CN, D) > 1.0 in the experiments mentioned below.

4.1 Extraction of feature words contained in retweeted text

Our test extracted feature words from sets of retweeted text (retweets) that were sent from the official Twitter accounts of 7 Net Shopping (7_netshopping), Amazon.co.jp (AmazonJP), and Rakuten Ichiba (RakutenJP).

The covered tweets were sent from these Twitter accounts between January 15 and 20, 2015. Table 2 shows the number of retweeted tweets and the number of FLR score-based feature word candidate phrases in the retweeted tweets.

Table 3 shows the top ten phrases for the Twitter accounts based on their TFIDF values, along with the support and the head confidence (H-Conf) measures for these phrases. The support and the head confidence were calculated using a standard for counting frequencies based on document frequency (DF).

The results in Table 3 provide characteristic phrase groups for the tweets that the followers retweeted. These feature word groups are phrases contained in the tweets sent by the Twitter accounts and can be considered to align with some of the followers' interests.

However, these feature words and phrases do not reflect the followers' interests directly, because the followers do not tweet these terms in their tweets. The issue is to capture more implicit interests and concerns of the followers from their behavioral history. Therefore, to obtain term groups that corresponded to groups of follower interests, the historic tweet content of followers who retweeted the tweets containing the original phrases must be examined. This will result in changes in the usage frequency of feature words and phrases.

[2] This IPA dictionary is a Japanese morpheme dictionary made by the project run by the Information-Technology Promoting Agency in Japan.

4.2 Feature words and temporal patterns of text retweeted by users

Twitter accounts can attract followers, who will receive all the information sent from the account. Tweets sent from followers contain various phrases, which are determined by their interests and most likely reflect those interests. Thus, this method hypothesized that when users retweeted tweets sent from the three well-known Twitter accounts, these retweets would contain feature words that relate to their previously sent tweets.

This section describes our process for testing this hypothesis by examining the content of tweets sent from the followers of our three well-known online retail Twitter accounts before the retweets were sent. To obtain the temporal patterns of the evaluation indices, we extracted feature words and the patterns of temporal change from these previous tweets. For followers who retweeted tweets sent from the three well-known Twitter accounts between January 15 and 20, 2015, we gathered the tweets sent between January 1 and 20, 2015 by the followers[3] of each of the well-known accounts. Then, the tweets gathered between January 1 and 14 are used to obtain the following temporal patterns of each evaluation index on each well-known retailer followers' tweet. For all of the terms, the 18 evaluation indices values were calcluated in each timestamped data set. Then, the values of each term consists of each data of the term.

Table 4 shows the number of gathered tweets between January 1 and 14, 2015[4] of the followers who sent tweets containing one of the phrases listed in Table 3 between January 15 and 20, 2015 for each account.

After applying the FLR scores to extract feature terms from these user tweets, the candidate terms were extracted, as shown in Table 5. For each top 1000 terms with the FLR score, the importance evaluation indices and sequential pattern evaluation indices are calculated in each daily set of documents. Then, for each evaluation index, the values for each timepoint, every daily set, of each term was converted into one temporal data. Therefore, the data set of one particular evaluation index for temporal pattern extraction contains up to 1000 instances with the values on each timepoint in this experiment.

Subsequently, a clustering method was then applied to the converted temporal data sets to obtain the temporal patterns of each index. The instances in this data set consist of the index values in each timepoint that represent the followers' activities before retweeting the tweets from the well-known

[3] Due to restrictions in the Twitter API, these users were the users who met the criteria from the randomly acquired 5000 users.

[4] Considering more realistic situation, the gathered tweets are not re-retrieved in the prior period after listing the followers who tweeted the tweets containing the feature words and phrases listed in Table 3.

Table 3 Top ten phrases for Twitter accounts based on TFIDF values and the support and the head confidence levels for these phrases (document frequency standard) from [5]

7 Net Shopping				Amazon.co.jp				Rakuten Ichiba			
Terms	TFIDF	Support(DF)	H-Conf(DF)	Term	TFIDF	Support(DF)	H-Conf(DF)	Term	TFIDF	Support(DF)	H-Conf(DF)
限定 (Limited)	41.39	0.26	1.00	タイム セール (Time Sales)	73.57	0.22	1.00	楽天 ポイント ("Rakuten" Points)	84.18	0.23	0.58
セブン(Seven)	33.47	0.29	1.00	OFF	63.38	0.20	1.00	楽天 ("Rakuten")	74.57	0.39	1.00
特典 (Special-gift)	32.95	0.22	1.00	人気(Popular)	53.11	0.14	1.00	応募 (Application)	51.21	0.26	1.00
予約 (Reservation)	32.53	0.43	1.00	PC	52.78	0.06	1.00	フォロワー (Follower)	51.21	0.26	1.00
予約 受付 (Reservation Accepting)	31.95	0.37	0.85	限定(Limited)	52.14	0.10	1.00	♪	46.70	0.38	1.00
月 (Month)	30.02	0.17	1.00	受付 (Accepting)	52.12	0.12	1.00	フォロー 解除 (Unfollow)	44.47	0.26	1.00
発売(For-sale)	29.21	0.20	1.00	チェック (Check-it)	51.10	0.13	1.00	当選 確率 (Winning Probability)	44.47	0.26	1.00
アカチャンホンポ ("Akachan-honpo")	27.06	0.14	1.00	予約 受付 (Reservation Accepting)	50.03	0.12	0.89	完了 (Finished)	44.47	0.26	1.00
セブン ネット ("Seven Net")	26.96	0.22	0.77	% OFF	48.80	0.11	0.75	下(Under)	44.31	0.23	1.00
DVD	26.49	0.21	1.00	最大 (Maximum)	45.34	0.10	1.00	RT	43.29	0.24	1.00

Table 4 Number of tweets between January 1 and 14 sent by the retweeting followers

	Following account		
	7 Net shopping	Amazon.co.jp	Rakuten Ichiba
1 st Januray	19,577	130	0
2nd January	20,906	55	15
3rd January	21,571	37	67
4th January	22,103	2,303	27
5th January	22,773	10,899	0
6th January	25,776	18,012	0
7th January	24,636	19,734	109
8th January	23,771	24,189	96
9th January	25,467	22,746	16
10th January	26,956	23,358	13
11th January	27,542	24,815	100
12th January	27,902	25,817	107
13th January	36,810	33,992	5073
14th January	35,835	30,273	22,651
Total	361,625	236,360	28,274

Table 5 Number of candidate feature words and phrases based on the FLR score in the entire data set of tweets counted in Table 4 for each well-known account from [5]

Retailer	|D|	FLR score-based candidate phrases
7 Net shopping	361,625	271,234
Amazon.co.jp	236,360	211,730
Rakuten Ichiba	28,274	40,080

account. As for the clustering method, a simple k-means implementation in Weka [14] was applied to the data sets in this analysis. The value of k was set up 10, which is the upper limit for obtaining clusters, since null clusters were allowed in this execution. For calculating the similarity between pairs of instances, the Euclidean distance with normalization on each variable was employed.

Table 6 shows the numbers of temporal clusters obtained and their sum of squared errors (s.s.e) values within the clusters on each data set from the five evaluation indices. The s.s.e is calculated using the following definition.

$$\text{s.s.e}(D_{\text{index}}) = \sum_{k}\sum_{i}\sum_{j}(v_{ij} - c_{kj})^2,$$

where c_{kj} is the jth time stamped value of the centroid of the temporal cluster k, and v_{ij} is the value of each evaluation index.

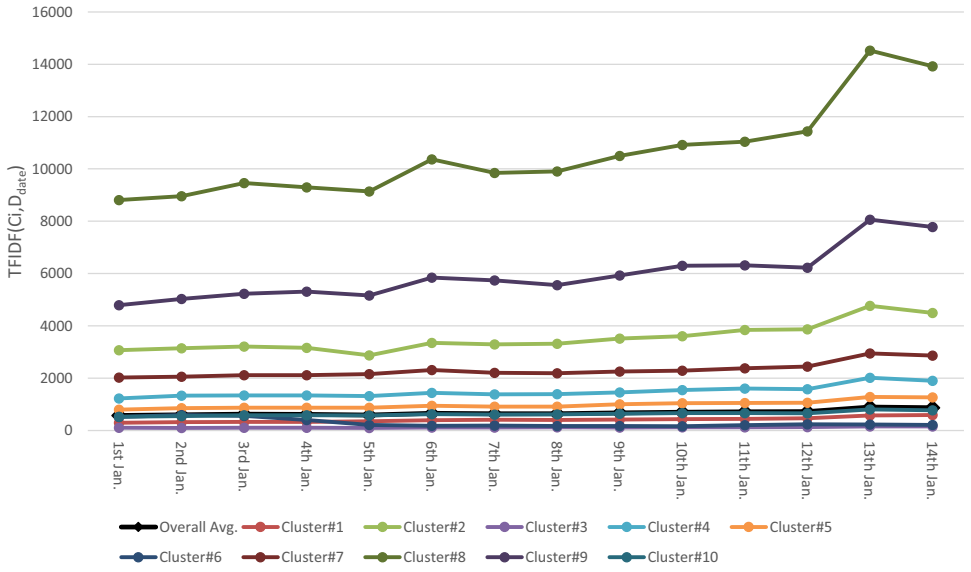
As shown in Table 6, TFIDF and support achieved smaller s.s.e values. This means the clusters obtained by these indices are cohering to the centroids of the clusters. In general, the s.s.e value increases when the number of clusters becomes smaller. Therefore, the clusters for MaxConf (DF) of 7 Net Shopping achieved greater cohesion as the temporal clustering.

4.3 Results for the temporal patterns with the evaluation indices from the different viewpoints

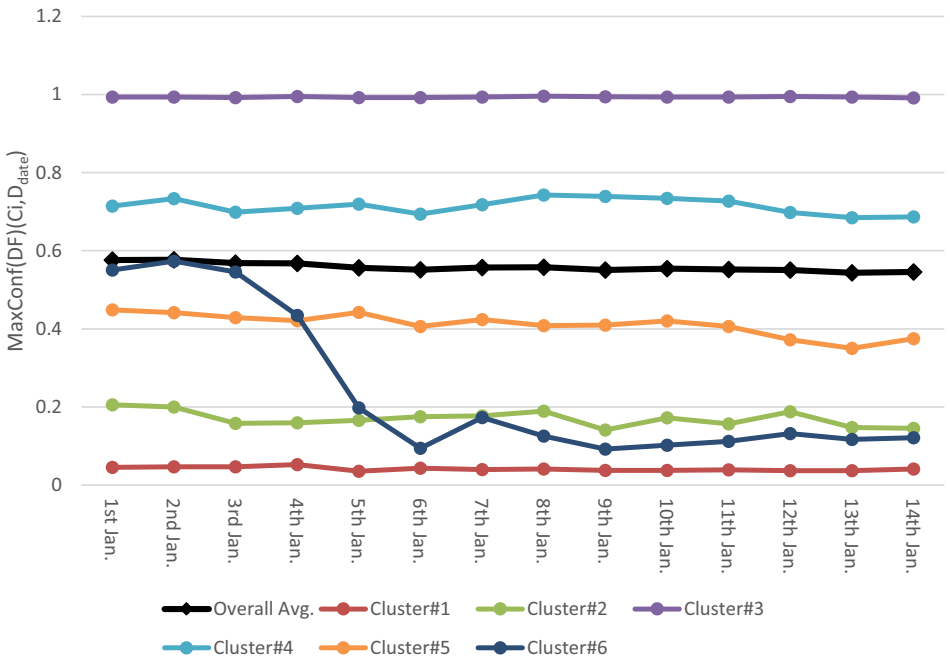
Figures 1 and 2 show the temporal patterns of the indices of 7 Net Shopping and Amazon.co.jp followers who retweeted

Table 6 Description of the clusters obtained for each evaluation index on each account data set from [5]

	Following account					
	7 Net shopping		Amazon.co.jp		Rakuten Ichiba	
	# Clusters	s.s.e	# Clusters	s.s.e	# Clusters	s.s.e
TFIDF	10	4.07	10	9.95	10	15.62
Support (DF)	10	3.62	10	5.02	10	6.38
Support (TF)	10	1.34	10	6.77	10	9.14
MaxConf (DF)	6	105.31	7	126.30	7	17.87
MaxConf (TF)	7	115.95	7	148.77	7	20.50



(a) Temporal Cluster Centroids from the TFIDF Dataset



(b) Temporal Cluster Centroids from the MaxConf(DF) Dataset

Fig. 1 Temporal patterns of TFIDF (**a**) and MaxConf (DF) (**b**) in 7 Net Shopping's retweeting followers' tweets from [5]

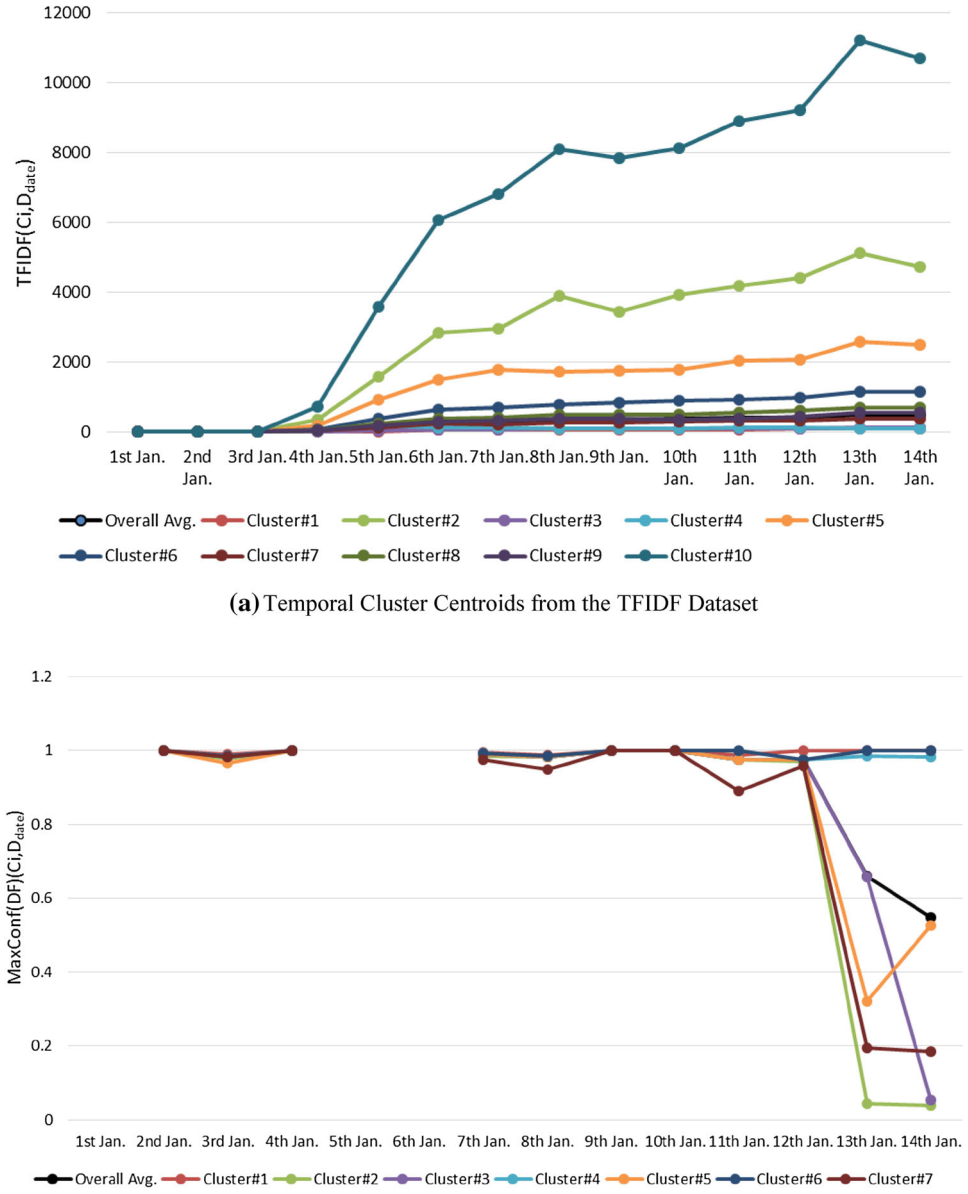


Fig. 2 Temporal patterns of TFIDF (**a**) and MaxConf (DF) (**b**) in AmazonJP's retweeting followers' tweets

(a) Temporal Cluster Centroids from the TFIDF Dataset

(b) Temporal Cluster Centroids from the MaxConf(DF) Dataset

their tweets between January 15 and 20, respectively. Since the clusters are obtained using the Euclidean distance, the lines, the centroids of each cluster, are calculated as the averages of each cluster.

Most of the patterns represents the differences of the averages of each evaluation index based on their temporal values in Fig. 1. As shown in Fig. 1a, the temporal patterns represent the levels of the averaged values of the member of each cluster. This means that the averaged value of this evaluation index is suitable for describing the features using TFIDF. However, as shown in Fig. 1b, Cluster#6 of MaxConf (DF) shows significant difference from the other patterns. This indicates that the use of both of the averaged value and the membership value to each cluster is suitable for MaxConf (DF) index.

To determine whether we should use the average of the obtained temporal clusters' centroid or the shape (membership) of them, the variance ratio between the differentials of the overall average and the differentials of each temporal cluster centroid is statistically tested. Table 7 shows the variance ratio of the differentials between the overall average and each temporal cluster and their F test results.

On the other hand, the temporal patterns made for tweeted terms of retweeting followers of Amazon.co.jp, as shown in Fig. 2, show the different patterns, compared with that of 7 Net Shopping. Since there are smaller numbers of tweets in

Table 7 Variance ratios of the differentials and F test results (**$p < 0.05$) of the TFIDF and MaxConf (DF) temporal clusters of 7 Net Shopping

	TFIDF			MaxConf (DF)		
	Variance	Var. ratio	F test	Variance	Var. ratio	F test
Overall avg.	2575.72			2.63E−05		
Cluster#1	827.51	0.32		3.44E−05	1.31	
Cluster#2	96, 603.54	37.51	**	6.64E−04	25.30	**
Cluster#3	96.24	0.04		3.23E−06	0.12	
Cluster#4	17, 849.92	6.93	**	3.91E−04	14.90	**
Cluster#5	4471.01	1.74		3.94E−04	14.99	**
Cluster#6	5417.91	2.10		6.39E−03	243.27	**
Cluster#7	21, 785.78	8.46	**			
Cluster#8	891, 739.52	346.21	**			
Cluster#9	305, 310.90	118.53	**			
Cluster#10	2074.96	0.81				

Table 8 Variance ratios of the differentials and F test results (**$p < 0.05$) of the TFIDF and MaxConf (DF) temporal clusters of AmazonJP

	TFIDF			MaxConK (DF)		
	Variance	Var. ratio	F test	Variance	Var. ratio	F test
Overall avg.	2016.04			0.12		
Cluster#1	416.88	0.21		0.10	1.00	
Cluster#2	295,954.64	146.80	**	0.20	2.06	
Cluster#3	290.38	0.14		0.16	1.63	
Cluster#4	2046.45	1.02		0.10	1.00	
Cluster#5	75,021.05	37.21	**	0.16	1.61	
Cluster#6	9735.64	4.83		0.10	1.00	
Cluster#7	1600.72	0.79		0.17	1.75	
Cluster#8	3813.31	1.89				
Cluster#9	2550.98	1.27				
Cluster#10	1,117,910.31	554.51	**			

the begining of the period, the values of these two evaluation indices are influenced by the size of these data sets. However, the centroids of TFIDF temporal patterns show that the levels of the average values through the period are more inportant than the trend of the index or the movement of the index. As shown in Fig. 2b, MaxConf shows the different tempral patterns that are both the levels of their averages and the other usage of the terms, because the index reflects combinations of words included in each term.

As the same as Table 7, Table 8 shows the F test results of the differentials between the overall averaged sequence and the other temporal cluster centroid sequences. Based on the result in Table 8, the features constructed for the AmazonJP retweeting followers' tweets can be distinguished whether we should use the levels of averages of TFIDF and the membership to Cluster #2, #5, and #10 of TFIDF.

Based on these results, analysts can understand that the followers who are interested in the targeted account have not only different usages of words and phrases, but also they make some similar patterns by focusing on their temporal changes of their usages as reflected by the evaluation indices.

Subsequently, according to the F test results of the variance ratio of the differentials, we can construct the features consisting of the average values of the period and the temporal pattern memberships for predicting followers' retweeting behavior.

Detailed results of the obtained temporal patterns

To consider more detailed differences of the contents of the extracted temporal patterns, the followings are the details of the results shown in Figs. 1 and 2. Those contained words and phrases included in the temporal patterns as the temporal clusters have calculated their similarity to the cluster centroid in each cluster using Euclidean distance.

Figure 3 shows that the top ten similar words and phrases included in Cluster#6 of Fig. 1b. By examining the temporal values of these words and phrases, most of these phrases appeared after January 5 for promoting new popular games on smartphones such as Android terminals. The pattern obtained using MaxConf (DF), that is Max Confidence based on Document Frequency, and its contents describe the changing of

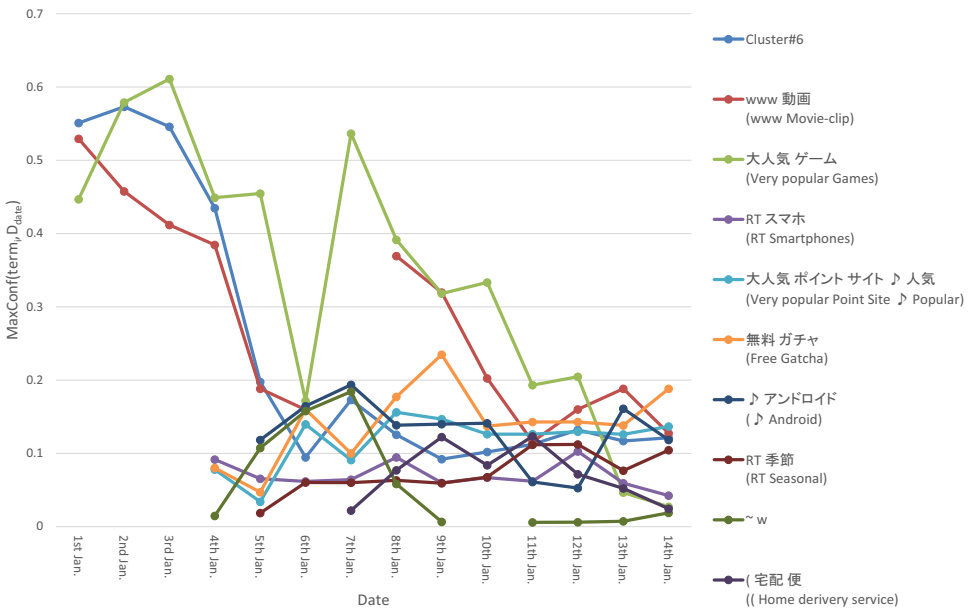


Fig. 3 Representative words and phrases of Cluster#6 of MaxConf (DF) in 7 Net Shopping retweeting followers' tweets from [5]

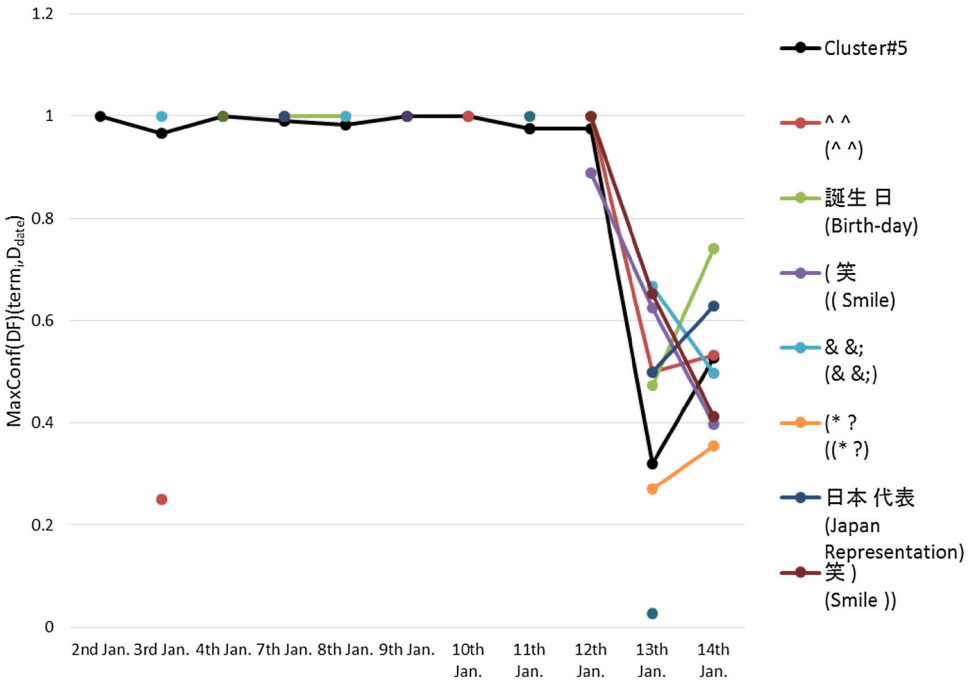


Fig. 4 Representative words and phrases of Cluster#5 of MaxConf (DF) in Amazon.co.jp retweeting followers' tweets

the promotion targets of the retweeting followers. They seem interested in smartphones and games on the smartphones. At the same time, they also tried to advertise their affiliated services such as online points and in-game coins.

Figure 4 also shows the top ten similar words and phrases on the Euclidean distance to the cluster centroid of Cluster#5 in Fig. 2b. These words and phrases are related to some face marks and emoticons in Japanese that are frequently used for make their tweets more friendly. In contrast to the words and phrases shown in Fig. 3, the interested followers to Amazon.co.jp often use less advertising expreessions in their tweets. In addition, they use more different combinations including these phrases in the last few days before retweeting the tweets of Amazon.co.jp or tweeting with containing the feature terms of Amazon.co.jp's tweets.

These results demonstrate that we can capture different aspects of the historical behaviors of the users by obtaining the temporal clusters of the different types of evaluation indices. These patterns will reveal about the followers' concerns from the viewpoint of the enterprise users more concretely, because the temporal patterns contain concrete feature words and phrases of the followers. This will help

analysts for promoting their sales items to adequate followers by selecting the temporal patterns and the term groups included in the selected temporal pattern. By targeting the follower's action to promotions of false rumors, it will be able to detect demagogues more quickly based on the values of the evaluation indices of the temporal pattern, which are not needed any preliminary appearance of particular word or phrase itself.

In addition to the above-mentioned effect, these behaviors also include mechanical accounts (called "bots") as well as human accounts. Among the viral marketing and social media mining field, it is one of the important issue to distinguish such mechanical bots as spam. Therefore, the different behaviors reflected by the temporal patterns and their including words and phrases will also help in distinguishing the spam accounts.

5 Evaluating stability of the proposed retweeting user analysis method based on the evaluation index behavior as the temporal pattern

To evaluate our proposed user behavior analysis method, we applied this method to the sets of retweeted tweets and followers' tweets from the different period. In this experiment, we use the three retailers retweeted tweets and the followers' tweets in 2016.

5.1 Extracting feature words and phrases of retweeted tweets in 2016

As the same as the feature words and phrases extraction in Sect. 4.1, we gathered the retrweeted tweets of the three retailers from February 17 to 23, 2016. Table 9 shows the numbers of retweeted tweets and the extracted composed nouns using the automatic term extraction method based on the FLR score [11].

For the candidate words and phrases in Table 7, the 19 evaluation indices defined in Table 1 are calculated. Table 10 shows the top ten words and phrases, which are sorted by the TFIDF index.

As shown in Table 10, the retweeted words and phrases are not so different in the 7 Net Shopping and Amazon.co.jp.

Table 9 Number of retweeted tweets and number of FLR score-based candidate phrases in tweets sent from three major retail Twitter accounts between February 17 to 23, 2016

Retailer	|D|	FLR score-based candidate phrases
7 Net shopping	113	336
Amazon.co.jp	499	1492
Rakuten Ichiba	81	162

In addition, the retailer's tweets and its retweeted tweets are changed by that of Rakuten Ichiba. Although the rank of the feature words and phrases is effected by their appearances, the meaning of the rank is not effected.

5.2 Feature words and temporal patterns of text retweeted by users in 2016

As for the temporal set of tweets, we gathered the followers' tweets who were retweeted the tweets from the three retailer accounts. As the same as the setting in Sect. 4.2, the tweets from the followers were picked up from the former period of the retweeting actions. The period for observing the followers' behavior was from February 1 to 16, 2016. The numbers of gathered followers' tweets in this period are shown in Table 11.

From the entire set of follower's tweets for each retailer's account, we extracted the candidate words and phrases for extracting feature words and phrases as the temporal patterns. Table 12 shows the number of entire follows' tweets on the 2016 period and the extracted candidate words and phrases based on the FLR score using $\mathrm{FLR}(\mathrm{term}_i) > 1.0$.

After calculating the evaluation indices, the data sets for temporal clustering on each evaluation index are constructed. Then, we selected top 1000 words and phrases based on FLR score for each data set. This process is for selecting more meaningful words and phrases based on the statistical score.

By applying the k-means clustering algorithm to these data sets, the results are obtained, as shown in Table 13. For constructing clusters, the value of k was set up 10, which is the upper limit for obtaining clusters, since null clusters were allowed in this execution. For calculating the similarity between pairs of instances, the Euclidean distance with normalization on each variable was employed.

The values of the sum of squared errors (s.s.e.) show larger gap between the Support indices and the maximum confidence (MaxConf) indices, as shown in Table 13. This indicates that the clusters using MaxConf have more variance in each cluster. Although the s.s.e. values of the support are smaller than the values of TFIDF, the raw values of TFIDF are some thousands times bigger than the raw values of the Support.

Results for the temporal patterns with the evaluation indices on the 2016 data set

Figure 5 shows the temporal patterns of the indices of 7 Net Shopping followers who retweeted 7 Net's tweets between February 17 and 23 in 2016. Since the raw values of TFIDF are bigger than the values of support, we show the result of TFIDF as Fig. 5.

The lines in Fig. 5 show the centroids of each cluster, which are calculated as the averages of members included in

Table 10 Top ten retweeted words and phrases based on TFIDF values and the support and the head confidence levels for these phrases (document frequency standard) on 2016

7 Net Shopping				Amazon.co.jp				Rakuten Ichiba			
Terms	TFIDF	Support(DF)	H-Conf(DF)	Term	TFIDF	Support(DF)	H-Conf(DF)	Term	TFIDF	Support(DF)	H-Conf(DF)
セブン(Seven)	32.13	0.32	1.00	RT	100.49	0.18	1.00	アプリ (Application)	84.18	0.23	0.58
企業 (Enterprise)	27.40	0.05	1.00	セール (For-Sale)	97.73	0.13	1.00	楽天 市場 ("Rakuten" "Ichiba")	74.57	0.39	1.00
特典 (Special-gift)	27.07	0.22	1.00	%	92.49	0.13	1.00	お客様 (Customers)	51.21	0.26	1.00
DVD	25.57	0.22	1.00	#	91.74	0.11	1.00	楽天 市場 アプリ ("Rakuten" "Ichiba" Application)	51.21	0.26	1.00
限定 (Limited)	25.27	0.33	1.00	OFF	90.53	0.12	1.00	県 (Prefecture)	46.70	0.38	1.00
発売(For-sale)	23.80	0.29	1.00	% OFF	86.17	0.11	0.86	Ver	44.47	0.26	1.00
予約 受付 (Reservation Accepting)	23.72	0.43	0.82	タイム セール (Time Sales)	80.05	0.09	0.97	メルマガ (Mail-Magazine)	44.47	0.26	1.00
月 (Month)	20.94	0.17	1.00	チェック (Check-it)	70.35	0.09	1.00	利用 (Use-of)	44.47	0.26	1.00
セブン ネット ("Seven Net")	20.52	0.21	0.65	春 (Spring)	68.71	0.06	1.00	県民 (Residents)	44.31	0.23	1.00
付き (With a S/T)	19.90	0.19	1.00	登場 (First-Appearance)	66.93	0.08	1.00	問い合わせ (Contact)	43.29	0.24	1.00

Table 11 Number of tweets between February 1 and 16, 2016 sent by the retweeting followers

	Following account		
	7 Net shopping	Amazon.co.jp	Rakuten Ichiba
1st February	1205	1504	926
2nd February	1302	1760	924
3rd February	1520	1851	1005
4th February	1376	1771	1229
5th February	1643	2115	1250
6th February	2098	2374	1379
7th February	2223	2787	1683
8th February	2255	2987	1811
9th February	2369	3731	2192
10th February	2923	4081	2165
11th February	3419	4848	3309
12th February	4588	5978	4500
13th February	6247	7527	5186
14th February	9376	9729	6729
15th February	11,403	10,808	7190
16th February	13,799	16,882	11,006
Total	67,746	80,733	52,484

each cluster. Most of these centroids as the temporal pattterns show almost same trends compared to the overall average.

On the other hand, the temporal patterns of the cluster centroids show different trends excepting Cluster #3, #4, #9, and #10 based on th F test for the differentials of the sequences, as shown in Table 14. This indicates that the membership to each cluster is more informative than using the levels of the index values.

Table 12 Number of candidate feature words and phrases based on the FLR score in the entire data set of tweets counted in Table 9 for each well-known account

Retailer	$\|D\|$	FLR score-based candidate phrases
7 Net shopping	67,746	98,283
Amazon.co.jp	80,733	113,923
Rakuten Ichiba	52,484	85,204

As shown in this result, the proposed temporal pattern extraction method using the sequential pattern evaluation indices from the different viewpoints obtains both of the features with/without temporal trends and the groups of the words and phrases at the same time, according to the temporal values of each evaluation index property.

6 Conclusion

In this paper, we examined the Twitter behavior known as retweeting, which refers to the dissemination of information that occurs when users resend tweets. We examined the differences between the feature word groups that are contained in retweeted text and the feature word groups contained in the tweet history of the followers. We assume that these users have an interest in the information sent from specific Twitter accounts. In our assessment of three well-known retailers' Twitter accounts, the method discovered the significant terms

Table 13 Description of the clusters obtained for each evaluation index on each account data set in 2016 February

	Following account					
	7 Net shopping		Amazon.co.jp		Rakuten Ichiba	
	# Clusters	s.s.e	# Clusters	s.s.e	# Clusters	s.s.e
TFIDF	10	11.90	10	9.59	10	13.39
Support (DF)	10	0.64	10	0.52	10	0.68
Support (TF)	10	4.04	10	0.35	10	1.40
MaxConf (DF)	9	434.92	9	415.64	7	430.52
MaxConf (TF)	9	383.46	9	369.96	7	380.99

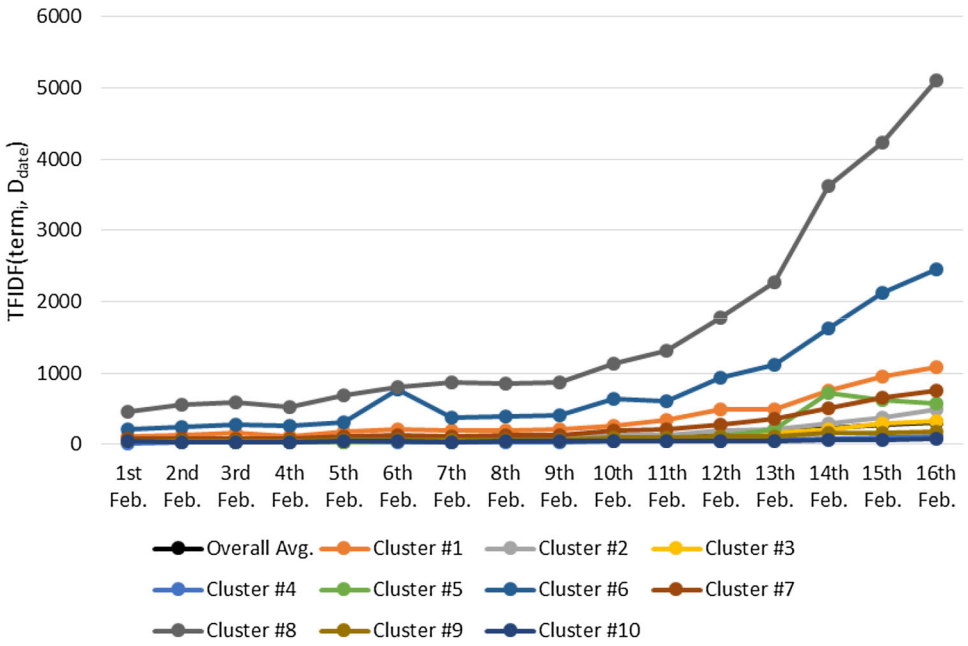


Fig. 5 Temporal patterns of TFIDF in 7 Net Shopping's retweeting followers' tweets between February 1 and 16, 2016

Table 14 Variance ratios of the differentials and F test results (**$p < 0.05$) of the TFIDF temporal clusters of 7 Net Shopping in 2016 February data set

	TFIDF		
	Variance	Var. ratio	F test
Overall Avg.	451.96		
Cluster#1	8684.86	19.22	**
Cluster#2	1710.86	3.79	**
Cluster#3	617.08	1.37	
Cluster#4	56.85	0.13	
Cluster#5	20,509.42	45.38	**
Cluster#6	64,502.22	142.72	**
Cluster#7	3397.68	7.52	**
Cluster#8	161,977.50	358.39	**
Cluster#9	161.18	0.36	
Cluster#10	66.64	0.15	

that the terms contained in retweets differed from those of the users' previous tweets. We further conclude that the obtained temporal patterns enable to describe the followers' characteristics not only as their tweeting behaviors but also as their tweeted words and phrases. By testing the trends of the temporal patterns, the results indicate whether the temporal patterns mean the characteristic trends or the levels of the evaluation index value.

Our future goal is to use this study for constructing a predictive model using the presence or absence of these temporal patterns as an explanatory variable for retweeting behavior. In addition to the patterns of phrase usage frequency changes, we will also acquire the temporal patterns of behavior changes, such as tweeting intervals, for the purpose of developing a method of constructing a predictive model that can be combined with the conventional feature word-based characterization.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers 24500175 and 26240036.

References

1. Goonetilleke, O., Sellis, T., Zhang, X., Sathe, S.: Twitter analytics: a big data management perspective. SIGKDD Explor. Newsl. **16**(1), 11–20 (2014)
2. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. SIGMOD Rec. **42**(2), 17–28 (2013)
3. Mahmud, J., Chen, J., Nichols, J.: Why are you more engaged? Predicting Social Engagement from Word Use, The Computer Research Repository (CoRR) (2014). arXiv:1402.6690
4. Bo, J., Sha, Y., Wang, L.: A multi-view retweeting behaviors prediction in social networks, Web Technologies and Applications, pp. 756–767. Springer International Publishing, New York (2015)
5. Abe, H.: Analyzing user behaviors based on temporal patterns of sequential pattern evaluation indices on twitter. Trends and Applications in Knowledge Discovery and Data Mining, pp. 177–188. Springer International Publishing, New York (2015)
6. Abe, H.: Analysis for finding innovative concepts based on temporal patterns of terms in documents, theory and applications for advanced text mining. In: Sakurai, S. (ed.), pp. 37–50 (2012)
7. Abe, H., Tsumoto, S.: Mining classification rules for detecting medication order changes by using characteristic CPOE subsequences, foundations of intelligent systems. In: Proceedings of ISMIS 2011, LNCS 6804, pp. 80–89 (2011)
8. Wu, T., Chen, Y., Han, J.: Association mining in large databases: a re-examination of its measures. In: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 621–628 (2007)
9. Lin C. X, Ji, M., Danilevsky, M., Han, J.: Efficient mining of correlated sequential patterns based on null hypothesis. In: Proceedings of the 2012 international workshop on Web-scale knowledge representation, retrieval and reasoning (Web-KR '12), pp. 17–24 (2012)
10. Twitter Web API 1.1. http://dev.twitter.com/docs/api/1.1
11. Nakagawa, H.: Automatic term recognition based on statistics of compound nouns. Terminology **6**(2), 195–210 (2000)
12. Kleinberg, M.J.: Authoritative Sources in a Hyperlinked Environment. J. ACM **46**(5), 604–632 (1999)
13. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab/
14. Witten, I.H., Frank, E.: Data Mining: PracticalMachine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, USA (2000)