Hindawi Publishing Corporation
Applied Computational Intelligence and Soft Computing
Volume 2014, Article ID 871412, 11 pages
http://dx.doi.org/10.1155/2014/871412

Research Article

An Activation Method of Topic Dictionary to Expand Training Data for Trend Rule Discovery

Shigeaki Sakurai, Kyoko Makino, and Shigeru Matsumoto

IT Research and Development Center, Toshiba Solutions Corporation, 3-22 Katamachi, Fuchu, Tokyo 183-8512, Japan

Correspondence should be addressed to Shigeaki Sakurai; sakurai.shigeaki@toshiba-sol.co.jp

Received 23 August 2013; Revised 28 December 2013; Accepted 13 January 2014; Published 26 February 2014

Academic Editor: Ying-Tung Hsiao

This paper improves a method which predicts whether evaluation objects such as companies and products are to be attractive in near future. The attractiveness is evaluated by trend rules. The trend rules represent relationships among evaluation objects, keywords, and numerical changes related to the evaluation objects. They are inductively acquired from text sequential data and numerical sequential data. The method assigns evaluation objects to the text sequential data by activating a topic dictionary. The dictionary describes keywords representing the numerical change. It can expand the amount of the training data. It is anticipated that the expansion leads to the acquisition of more valid trend rules. This paper applies the method to a task which predicts attractive stock brands based on both news headlines and stock price sequences. It shows that the method can improve the detection performance of evaluation objects through numerical experiments.

1. Introduction

Recently, various kinds of sequential data are easily and cheaply collected from real world and virtual world. It is anticipated that the data includes the knowledge that brings smart life to us. Therefore, many researches aggressively tackle on the knowledge discovery task from the data [1–5]. On the other hand, the knowledge discovery task depends on features of the data and types of the knowledge. It is impossible to deal with all features and all types by only a method. It is indispensable to develop a discovery method reflecting target features and types.

We try to develop a method which predicts whether evaluation objects such as companies and products are to be attractive in near future. This is because target data is easily collected from internet environments and it is easy for the prediction task to quantitatively evaluate the accuracy. The method deals with both text sequential data and numerical sequential data related to evaluation objects. It discovers trend rules from them. Each trend rule represents a relationship among evaluation objects, keywords, and numerical changes. The method applies the trend rules to text sequential data collected in the designated period and predicts attractive evaluation objects in the next period. It regards evaluation objects whose trends change as attractive evaluation objects. This paper aims at discovering more valid trend rules in order to improve detection performance in the prediction. It focuses on the expansion of the training data because many machine learning researches show that the expansion brings about better learning results. This paper activates a topic dictionary for the expansion. The dictionary describes relationships between evaluation objects and keywords related to numerical changes. This paper verifies the effect of the method through numerical experiments. That is, this paper contributes to the data mining research field with the following three viewpoints. Firstly, it incorporates a new function expanding learning data based on the topic dictionary into the previous discovery method of trend rules. Secondly, it incorporates a new function excluding frequent patterns including some evaluation objects as uncharacteristic ones into it. Thirdly, it applies the revised method to the prediction task in the financial field and verifies its effect.

Thus, the remaining parts of this paper are composed of the followings. The second section introduces some related

works in the financial field. The third section introduces a discovery method of trend rules and a prediction method based on them [6]. The fourth section proposes an expansion method of training data. The fifth section explains the experimental data, the experimental method, and the experimental results. The sixth section discusses the effect of the proposed method. Lastly, the seventh section describes the summary and future works.

2. Related Works

This section introduces some related works based on financial data composed of numerical data and text data.

Antweiler and Frank [7] investigate relationships between the stock data of 45 companies in Dow Jones Industrial Average (DJIA) and more than 150 million messages. The paper shows that the messages can explain the volatility which is one of evaluation criteria for stock prices but cannot help to gain the revenue in the trade operations.

Bollen et al. [8] propose a method that inductively learns relationships between DJIA and 6 kinds of emotions included in the messages described in Twitter. The method acquires the relationships by using fuzzy self-organization maps. It shows that the relationship in the case of the emotion "Calm" can explain the daily changes in DJIA with 87.6% accuracy.

Zhang et al. [9] propose a method predicting stock market indexes based on tweets. It analyzes correlations between the stock market indexes and two collective emotions such as "Fear" and "Hope." It finds that the emotional tweet percentage is negatively related to the indexes and is positively related to volatility indeX (VIX).

Fung et al. [10] propose a method predicting the changes of stock prices in the stock market. The method segments numerical stock price data into three trends: "Rise," "Steady," and "Drop." The method focuses on two trends of them: "Rise" and "Drop" and inductively learns classification models. The models are used for the prediction.

Mittermayer and Knolmayer [11] propose a method that automatically classifies news articles to predict the trends of stock prices. The method uses a thesaurus created by humans and improves a labeling method of the articles. It selects the appropriate training data to construct the prediction model. The paper shows that the method arrives at the high prediction performance.

Peramunetilleke and Wong [12] propose a method that acquires classification rules. Their antecedent part is weighted keywords and their result part is classes discretizing changes for currency exchange. The keywords are selected by stock traders or human experts and the weights are calculated by referring to frequencies of the keywords in the target period.

Choudhury et al. [13] develop a model analyzing communication dynamics in the blogosphere to decide correlations with movement in stock market. The model is acquired from the data in the blogosphere by support vector machine (SVM) regression. The model can predict the magnitude of the movement about 78% accuracy and its direction about 87% accuracy.

Seo et al. [14] develop the intelligent multiagent system for the portfolio management. The system acquires a model classifying news articles related to financial information of companies into 5 classes. It classifies news articles and evaluates information collected by agents. The portfolio management is performed by the classification and the evaluation.

Some existing methods [7–9, 12] analyze relationships between a specific numerical sequence and texts. The numerical sequence is the synthetic stock indexes such as DJIA or the change of specific currency exchange. The other existing methods [10, 11, 13, 14] analyze relationships between texts related to limited stock brands and their numerical sequences. Therefore, it is difficult for these existing methods to simultaneously and respectively deal with many evaluation objects such as all stock brands in a stock exchange and currency exchange among currency all over the world.

Our previous research [15] proposes a ranking method of evaluation objects in order to overcome this problem. The method deals with numerical sequential data for each evaluation object and text sequential data including the contents related to evaluation objects. It constructs a ranking model by referring to the change of the numerical sequential data and the amount of the text sequential data. Many evaluation objects can be simultaneously and respectively analyzed. Attractive evaluation objects are extracted from them. However, the method cannot sufficiently explain the reason why the evaluation objects are selected as attractive ones. Thus, our another previous research [6] tackles the development of the prediction method which can represent the reason. It acquires trend rules from the sequential data. The rules can be displayed as the reason. Also, they are used in order to predict attractive evaluation objects. The method can predict attractive evaluation objects to some extent. However, their detection performance should be revised for the more valid prediction. Thus, this paper tries to improve the detection performance.

3. Analysis of Complex Sequential Data

This section explains the analysis method of complex sequential data [6]. The method has two phases: the learning phase and the prediction phase. The learning phase discovers trend rules from the text sequential data and the numerical sequential data. The prediction phase predicts attractive evaluation objects by using trend rules and text sequential data in the designated period. In the following, the format of the complex sequential data and these phases are explained.

3.1. Format of Complex Sequential Data. The complex sequential data is composed of text sequential data (D) and numerical sequential data (V). Each element ($d \in D$) in the text sequential data has the time stamp (t_d) and the text (m_d). The text describes the contents related to evaluation objects (E). Here, the evaluation objects are given by users in advance and are stored in an evaluation object list. Some elements can have the same time stamp. On the other hand, each evaluation object ($e \in E$) has respective numerical sequential data ($V[e] \in V$). The numerical sequential data is a sequence of numerical value ($v_{e,t} \in V[e]$) for each time stamp (t).

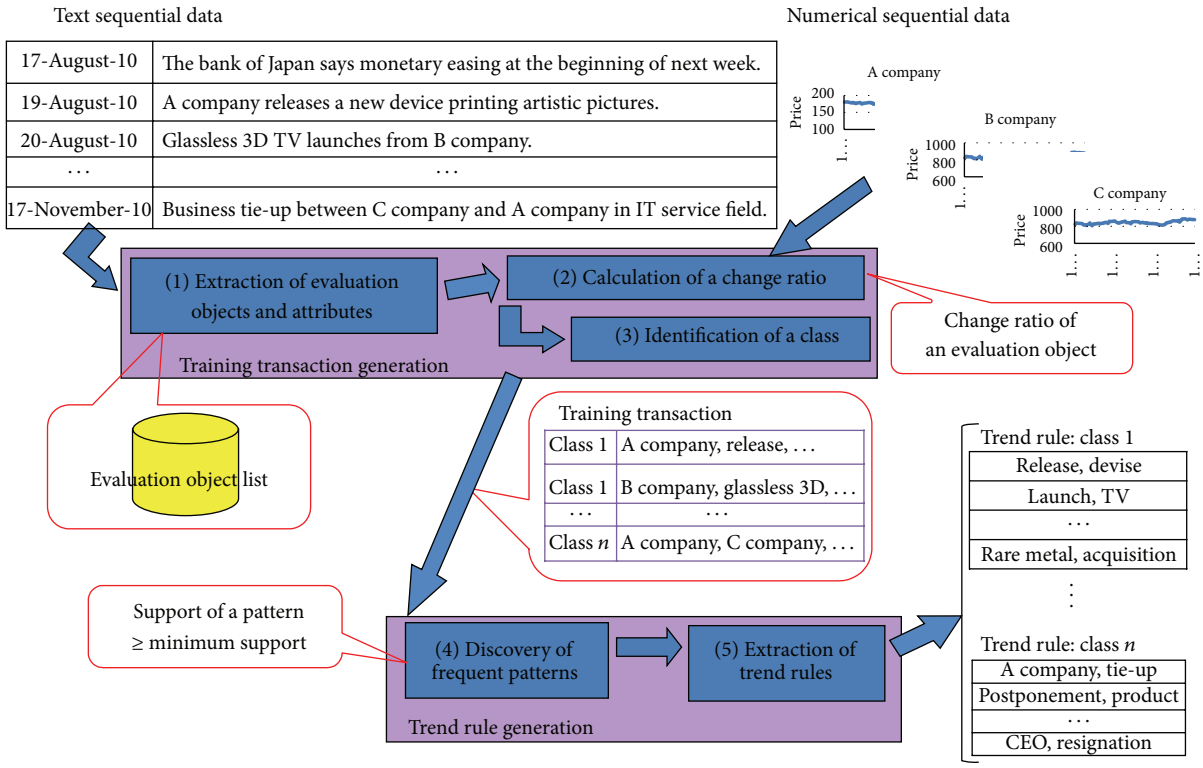


Figure 1: An outline of learning trend rules.

3.2. Learning of Trend Rules. The learning phase is composed of 5 subprocesses as shown in Figure 1. The subprocess 1 performs morphological analysis for the texts. Our experimental system deals with Japanese texts. It uses Chasen [16], which is one of representative Japanese morphological analysis engines. The engine separates a text into words and assigns their parts of speech to them. Noun words are extracted from the analyzed texts. Each noun word is evaluated whether it is one of proper nouns stored in the evaluation object list. If the word is one of the proper nouns, it is regarded as an evaluation object. Otherwise, it is regarded as an attribute. The subprocess 2 calculates a change ratio for the evaluation object described in the text by referring to its numerical sequence. That is, it calculates the difference between the numerical value in the time assigned to the text and the one in the next time and divides the difference by the former value. The subprocess 3 identifies a class by referring to the change ratio of each evaluation object included in the text. Here, each class has a respective range of change ratios. The combination of evaluation objects and attributes extracted from the text is a training transaction with a class. The training transaction is generated by the subprocesses 1, 2, and 3. We note that a training transaction is not generated when the text does not include an evaluation object.

The subprocess 4 applies the set of training transactions with the same class to the discovery method of frequent patterns [1, 17] and discovers frequent patterns for each class. Here, each frequent pattern is composed of evaluation objects and attributes. Its support is larger than or equal to the minimum support. Lastly, in the subprocess 5, a class assigned to the set is combined with discovered frequent patterns. The combination is a trend rule. The subprocess 4 and the subprocess 5 generate trend rules. The rules can represent the relationship among evaluation objects, attributes, and changes of the numerical value.

Algorithm 1 shows a pseudocode for this phase. In Algorithm 1, text sequential data (D), numerical sequential data (V), evaluation object list (E), class set (C), minimum support (M_s), and threshold set for change ratios (Th) are inputs. But, the threshold for the last class is set to the infinity. Also, the sets of trend rules ($R[c]$) for each class (c) are outputs. $Tr[c]$ is training transactions for each class. W_d is evaluation objects and attributes extracted from an element (d) of the text sequential data. $N[c]$ is the number of each class identified from an element. $P[c]$ is frequent patterns for each class. In this pseudocode, step 3 extracts evaluation objects and attributes from an element (d). Step 8 and step 9 calculate a change ratio (ch_w) for an evaluation object (w). Step 10 identifies its class (c_w) and step 12 identifies a class (c_d) for an element. In addition, step 16 discovers frequent patterns for each class. Step 17 extracts trend rules from the generated training transactions for each class.

3.3. Prediction of Attractive Evaluation Objects. The prediction phase has 4 subprocesses as shown in Figure 2. The prediction phase deals with the acquired trend rules and the text sequential data collected in the designated period. The subprocess 1 extracts evaluation objects and attributes from texts in the data. The extraction is equal to the one of the learning phase. The subprocess generates an evaluation

```
(01) initializeTransaction(Tr)
(02) For d ∈ D
(03)    W_d = extractEvaluationObjectAndAttribute(d);
(04)    t_d = getTime(d);
(05)    initializeClassCounter(N);
(06)    For w ∈ W_d
(07)       if w ∈ E
(08)          (v_{w,t_d}, v_{w,t_d+1}) = getNumericalValue(V[w], t_d);
(09)          ch_w = calculateChangeRatio(v_{w,t_d}, v_{w,t_d+1});
(10)          c_w = identifyClassForEvaluationObject(ch_w, Th);
(11)          N[c_w] + = 1;
(12)    c_d = identifyClassForElement(N);
(13)    if c_d ≠ φ
(14)       Tr[c_d] = addTransaction(W_d, Tr[c_d]);
(15) For c ∈ C
(16)    P[c] = discoverFrequentPattern(Tr[c], M_s);
(17)    R[c] = extractTrendRule(P[c], c);
```

Algorithm 1: Pseudocode for learning trend rules.

transaction composed of the evaluation objects and the attributes.

Next, the subprocess 2 evaluates whether a trend rule matches the evaluation transaction. That is, it judges whether evaluation objects and attributes in the trend rule are completely included to the evaluation transaction. The evaluation transaction is assigned to a class of the trend rule when the trend rule matches it. The evaluation is performed for each trend rule whose number of items is larger than or equal to the minimum number of items. The class of the evaluation transaction is decided by referring to the number of assigned classes. We note that the evaluation transaction is not processed in the following subprocesses when any trend rules do not match it. The subprocess 3 accumulates 1 to class counters of evaluation objects included in the evaluation transaction where the class of the class counters is equal to the one of the evaluation object and the evaluation objects have class counters of respective classes. These subprocesses are repeated for all collected text sequential data. Lastly, the subprocess 4 identifies whether each evaluation object is attractive by referring to its class counter. That is, if the maximum value of the class counters assigned to the evaluation object is larger than or equal to the minimum number of transactions, the evaluation object is attractive ones. The minimum number is given by users in advance. The subprocesses 2, 3, and 4 predict attractive evaluation objects based on the trend rules.

This paper deals with a prediction task of attractive stock brands. The task defines three classes "Drop," "Steady," and "Rise." It is anticipated that stock traders are interested in changes of stock prices. That is, this paper focuses on two classes "Drop" and "Rise." On the other hand, it does not care about the difference of the classes. This is because it is difficult to predict directions where stock prices change by using only limited text information and the notification of stock brands changing the prices brings useful information to the stock traders even if the directions are not identified. Thus, attractive evaluation objects are identified by the total value of the class counters corresponding to "Drop" and "Rise." This task deals with the total value for two classes. However, the method can deal with respective values of the class counters for three or more classes in other tasks.

Algorithm 2 shows a pseudocode for the prediction phase. In Algorithm 2, text sequential data (D), sets of trend rules ($R[c]$) for each class (c), minimum number of transactions (M_t), minimum number of items (M_i) are input. Also, attractive evaluation objects (e), their classes (c_e), and values ($M[e, c_e]$) of their class counters are output. In this pseudocode, step 2 picks up trend rules used by this prediction phase. Step 5 extracts evaluation objects and attributes from an element (d) of text sequential data. Step 9 evaluates an evaluation transaction (W_d) based on a trend rule (r) and step 12 identifies a class (c_d) for the element. If the class is not a null value, steps 14, 15, and 16 accumulate values ($M[w, c_d]$) of the class counters for evaluation objects (w) included in the evaluation transaction. Lastly, step 18 identifies attractive evaluation objects.

4. Expansion of Training Data

The previous research [6] shows that there are many texts such that evaluation objects are not extracted. On the other hand, it is anticipated that the increase of training data realizes more valid inductive learning. It is important for the learning of trend rules to increase training data. This section proposes a method that activates a text not directly describing evaluation objects.

In the prediction task of attractive stock brands, text sequential data, numerical sequential data, and evaluation objects are news headlines, stock price sequences, and stock brands, respectively. We know that the corporate performance of export companies gets worse when their currency exchange rises. The rise tends to lead to the drop of their stock

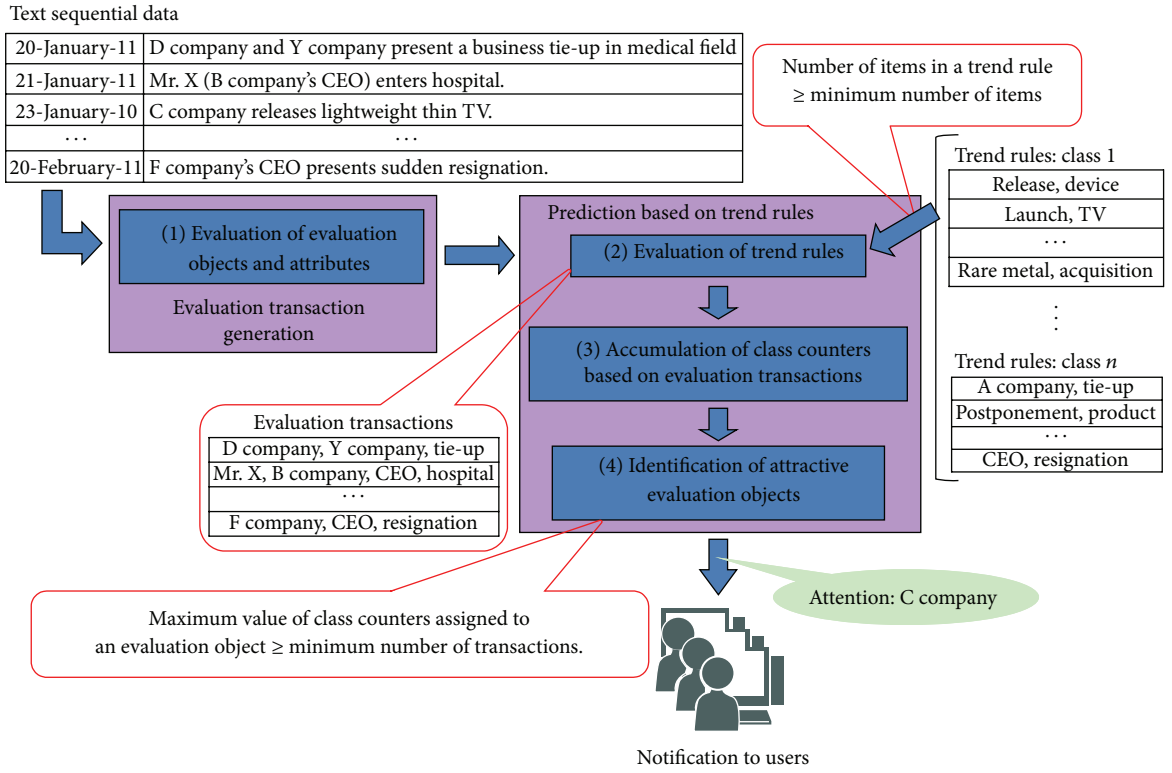


Figure 2: An outline of prediction of attractive evaluation objects.

```
(01)  For c ∈ C
(02)      R[c] = pickUpTrendRule(R[c], M_i);
(03)  initializeEvaluationObjectCounter(M);
(04)  For d ∈ D
(05)      W_d = extractEvaluationObjectAndAttribute(d);
(06)      initializeClassCounter(N);
(07)      For c ∈ C
(08)          For r ∈ R[c]
(09)              f = evaluateTrendRule(W_d, r);
(10)              if f == TRUE
(11)                  N[c] + = 1;
(12)      c_d = identifyClassFor Element(N);
(13)      if c_d ≠ φ
(14)          For w ∈ W_d
(15)              if w ∈ E
(16)                  M [w, c_d] + = 1;
(17)  For e ∈ E
(18)      f = identifyAttractiveEvaluationObject(M[e, ·], M_t);
(19)      if f == TRUE
(20)          output(e, c_e, M [e, c_e]);
```

Algorithm 2: Pseudocode for prediction of attractive evaluation objects.

prices. Even if a text does not directly describe a specific stock brand, the text can give a big impact on the stock price of the brand. That is, a specific keyword in the text can represent the change of the stock price. This keyword is regarded as a topic. This paper activates various topics in order to expand the learning data. The topics are managed by a topic dictionary with a three-layer structure. First layer is middle categories representing related topics, second layer is topics, and third layer is evaluation objects. Table 1 shows an example of the topic dictionary for the prediction of attractive stock brands. In this table, the topic dictionary has six or more relations between keywords and evaluation objects. The relation in the

Table 1: An example of topic dictionary.

First	Second	Third
Currency exchange	Strong yen	A company
Currency exchange	Strong yen	B company
—	—	—
Rare metal	Nickel	C company
Rare metal	Tungsten	D company
—	—	—
Eco-point	Refrigerator	E company
Eco-point	Air conditioner	F company

Table 2: Number of text sequential data.

Site	Period	
	D1	D2
Excite	132,878	38,761
Goo	143,062	30,593
Infoseek	240,141	62,407
Livedoor	233,773	66,740
Yahoo	253,619	70,184
Total	**1,003,473**	**268,685**

second line shows that the stock price of "A company" is related to "Strong yen" and "Currency exchange." Also, the relation in the fifth line shows that the one of "C company" is related to "Nickel" and "Rare metal." Similarly, the eighth line shows that "Eco-point" and "Refrigerator" relate to the one of "E company." Eco-point is the policy that encourages consumers to replace old home appliance with ecological new one and that was performed in Japan.

This paper tries to simply use the topic dictionary in order to confirm its effect for the detection performance of evaluation objects. If a text includes keywords included in the first layer and the second layer, they are replaced with evaluation objects stored in their third layer. Figure 3 shows an outline of the transformation. That is, in the case of the keyword "Strong yen" in the first original transaction, the method extracts "Strong yen" in the dictionary and extracts corresponding stock brands "A company" and "B company." The keyword "Strong yen" is replaced with the stock brands "A company" and "B company." Similarly, the keyword "Rare metal" in the second original transaction is replaced with the stock brands "C company" and "D company." Also, the keyword "refrigerator" in the third one is replaced with "E company." The original transactions are transformed into the transactions as shown in the right bottom of Figure 3. The transformation can be performed by incorporating additional steps as shown in Algorithm 3 between step 3 and step 4 in Algorithm 1. In the additional steps, *TD* is a topic dictionary. Step 4 in Algorithm 3 transforms a topic (w) into evaluation objects.

A topic usually represents some stock brands. The transformed transaction tends to include some evaluation objects. The discovery method of frequent patterns tends to discover many frequent patterns including some stock brands. However, the patterns are not always characteristic patterns for representing the change of stock prices because the patterns only reflect on the topic dictionary. Thus, the method deletes patterns including some evaluation objects as uncharacteristic patterns. Due to the deletion, the method can avoid such risk that important patterns are hidden in the large amount of patterns. In the case of Figure 4, the two former patterns are deleted because they include two companies which are evaluation objects, respectively. On the other hand, the two latter patterns are trend rules because they do not include two or more companies. The training data can be expanded without generating many meaningless patterns due to the improvement of the discovery method. The improvement can be performed by incorporating an additional step as shown in Algorithm 4 between step 16 and step 17 in Algorithm 1. But, if we try to more efficiently discover frequent patterns without including uncharacteristic ones, the additional step should be called by step 16 in Algorithm 1.

We note that the transformation can be applied to the prediction phase. That is, the additional steps as shown in Algorithm 3 are incorporated between step 5 and step 6 in Algorithm 2. More evaluation transactions are used in order to identify attractive evaluation objects. This is because values of the class counters for evaluation objects are calculated based on evaluation objects in the evaluation transactions assigned classes. It can be additionally anticipated that more valid identification is performed.

5. Experiment

This section explains experiments verifying the effect of the proposed method. That is, the experimental data, the topic dictionary, the evaluation criteria, the experimental method, and the experimental results are explained in order.

5.1. Experimental Data. This experiment uses news headlines distributed by five sites (Excite, Goo, Infoseek, Livedoor, and Yahoo) as text sequential data. The ones for the learning phase (D1) are collected from August 28, 2010, to January 31, 2011, and the ones for the prediction phase (D2) are collected from February 1, 2011, to March 10, 2011. The last day of the prediction is the day before the huge earthquake that occurred in East Japan. We think that trend rules have changed before and after the earthquake. In order to avoid this influence, the prediction period is decided.

Table 2 shows the number of news headlines for each site. About a million news headlines are used to acquire trend rules and about 0.27 million ones are used to predict attractive stock brands.

This experiment also uses daily stock price sequences for each stock brand in http://www.geocities.jp/sundaysoftware/csv/keiretu.html as numerical sequential data. The site stores the sequences for 250 business days with csv format. Collected business days include the periods of both D1 and D2. Each sequence is composed of stock brand code, date, opening price, lowest price, highest price, closing price,

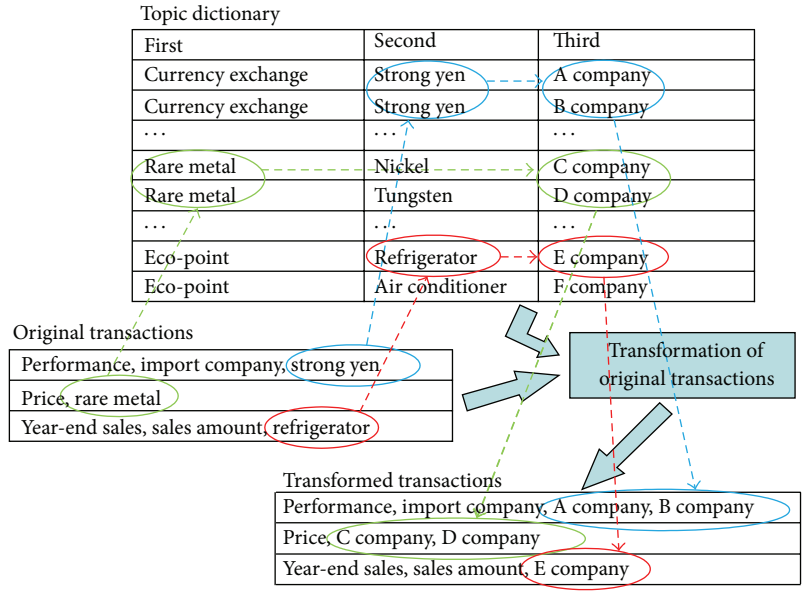


Figure 3: Transformation of transaction based on a topic dictionary.

```
(01) EW_d = ϕ;
(02) For w ∈ W_d
(03)     if w ∈ TD
(04)         EW_d += transformTopic(w);
(05)     else
(06)         EW_d += w;
(07) W_d = EW_d;
```

Algorithm 3: Pseudocode for the transformation of transaction based on a topic dictionary.

and turnover. This experiment focuses on the sequences of opening price.

On the other hand, evaluation objects are stock brands listed in Tokyo Stock Exchange, Sapporo Stock Exchange, Osaka Stock Exchange, Nagoya Stock Exchange, and Fukuoka Stock Exchange. This experiment collects their names from the home pages of each stock exchange. The number of evaluation objects arrives at 3,951.

5.2. Topic Dictionary. This experiment uses a keyword table stored in http://www.asset-alive.com/thema/ as a topic dictionary. The table is downloaded on October 25, 2011. It includes 134 keywords in the first layer and 867 keywords in the second layer. The keywords are related to the changes of stock price. Some keywords are compound nouns. They are separated into respective nouns. The third layer includes stock brand related to the keywords. The table has 4,842 relations between the keywords and the stock brands.

5.3. Evaluation Criteria. Stock traders cannot focus on all attractive stock brands because many stock brands are listed and many parts of them can be attractive. In addition, even if the recommendation system misses some attractive stock brands, the stock traders do not always care about the miss. On the other hand, if many recommended stock brands are not attractive, the stock traders cannot believe the recommendation of the system. It is important for the system to recommend attractive stock brands with high probability. Therefore, the precision (p) defined by (1) is very important:

$$p = \frac{\text{Number of extracted truly attractive stock brands}}{\text{Number of extracted stock brands}}. \tag{1}$$

In this experiment, the prediction phase extracts attractive stock brands for each day. The numbers accumulated for both the days and the stock brands are used to calculate the precision.

Although the precision is very important, the recall and the F-measure can be good criteria for the recommendation system. Thus, this experiment calculates the recall (r) and the F-measure (f). They are defined as shown as follows:

$$r = \frac{\text{Number of extracted truly attractive stock brands}}{\text{Number of truly attractive stock brands}},$$

$$f = \frac{2 \cdot p \cdot r}{p + r}. \tag{2}$$

```
(01)  P[c] = deleteUncharacteristicPattern(P[c]);
```

Algorithm 4: Pseudocode for the deletion of uncharacteristic patterns.

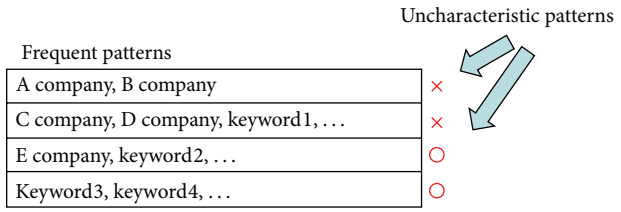


Figure 4: Deletion of uncharacteristic patterns.

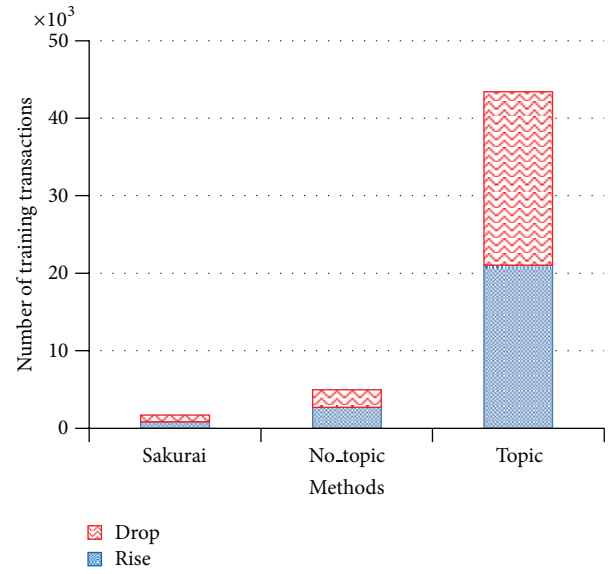


Figure 5: Number of training transactions.

This experiment regards stock brands whose change ratios are included in either the class "Drop" or the class "Rise" as truly attractive stock brands. Each stock brand is judged whether it is truly attractive in the designated day by referring to the change ratios for the next day. The judged results are used as the ground truth for this experiment. The judge is performed for each day in the prediction period. In addition, this experiment focuses on attractive stock brands which can be extracted from news headlines. That is, if some stock brands are not described in news headlines, the stock brands are excluded from the calculation of these evaluation criteria. This exclusion is performed for each day. This is because we cannot understand changes of their stock prices from the news headlines. The changes exceed the scope of the proposed prediction method. It is necessary for the recommendation system to use additional information in order to understand them. In near future, the system will be revised by referring to the knowledge in a financial engineering field.

5.4. Experimental Method. This experiment discovers trend rules from the data set D1 and predicts attractive stock brands by referring to the data set D2. It calculates evaluation criteria as shown in the previous subsection in order to compare results of the proposed method with previous results [6]. Then, the same parameter set is used. That is, the threshold of the change ratio in the learning phase is 0.05, and the one in the prediction phase is 0.01, 0.02, 0.025, 0.03, and 0.05. In the case of the threshold a, ranges "$\leqq -a$" and "$a<$" correspond to the class "Drop" and the class "Rise," respectively. We note that a higher change ratio is set in order to use news headlines with big impacts in the learning phase. The minimum support is 0.005, 0.01, 0.02, and 0.03. It is a criterion which evaluates whether a pattern is frequent. In addition, the minimum number of items is 2, and the minimum number of transactions is 1, 3, 5, and 10. The two latter parameters are used in order to judge whether a stock brand is attractive for each day. The first minimum number shows the minimum number of items included in trend rules. When a trend rule is composed of items whose number is smaller than the minimum number, the trend rule is not used. Also, the second minimum number shows the minimum number of transactions which is assigned to stock brands in order to be attractive. When a stock brand has the value which is smaller than the minimum number of transactions, it is not regarded as an attractive stock brand. In this experiment, the total value of the class "Drop" and the class "Rise" is referred to.

5.5. Experimental Result. This section shows parts of experimental results in Figures 5–10. In each figure, "Sakurai" shows the results in [6] and it does not use the topic dictionary. "No_topic" shows the results in the case the topic dictionary is not activated and "Topic" shows the one in the case the topic dictionary is activated. Also, "Sakurai" deals with only stock brands in the first section of Tokyo Stock Exchange. This section deals with 1,680 stock brands. "No_topic" and "Topic" deals with all stock brands. It is anticipated that the expansion of stock brands leads to the one of training transactions. However, our preliminary experiments show that the difference of stock brands does not give a big impact on the detection performance. We think that the reason is why the first section of Tokyo Stock Exchange is composed of major stock brands in Japan and the minor stock brands unincluded in the exchange rarely appear in news headlines. Therefore, discovered trend rules are limitedly related to the minor stock brands and their impact is small.

Figure 5 shows the number of training transactions extracted from the data set D1. In this figure, each bar graph is accumulated by the number of "Drop" and "Rise."

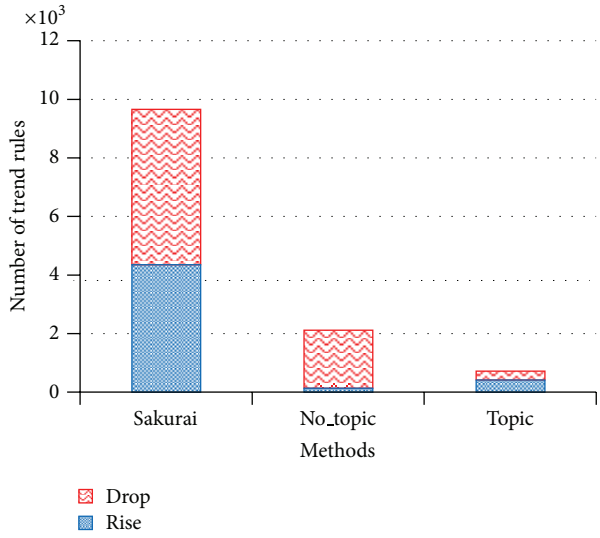


Figure 6: Number of trend rules; minimum support is 0.005.

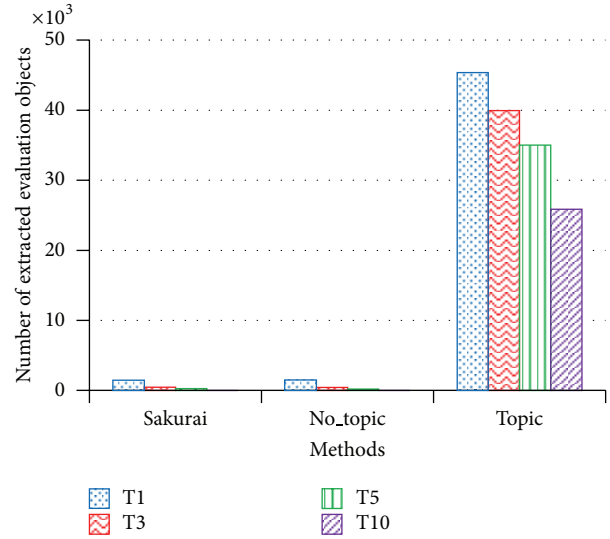


Figure 8: Number of extracted evaluation objects.

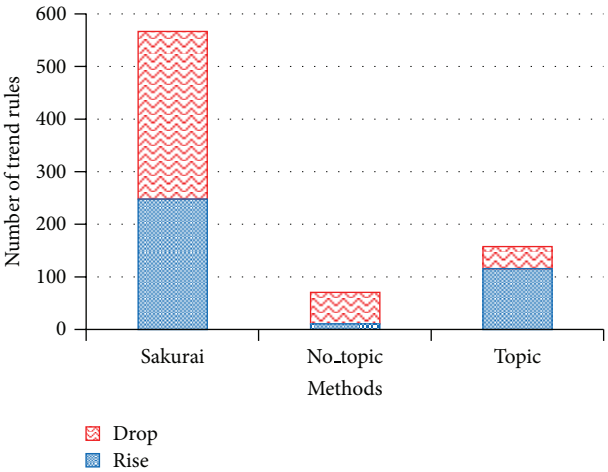


Figure 7: Number of trend rules; minimum support is 0.01.

Figures 6 and 7 show the number of trend rules in the case that the minimum support is 0.005 and 0.01, respectively. "Sakurai," "No_topic," and "Topic" show results of respective methods. Each bar graph is accumulated by the number of "Drop" and "Rise."

Figure 8 shows the number of extracted evaluation objects. Each method has 4 bar graphs. Each bar graph shows the number in the case that the minimum transaction is changed by 1, 3, 5, and 10.

Figures 9 and 10 show the detection performance by each evaluation criterion. The former figure is a result in the case the minimum transaction is 1 and the latter one is a result in the case of 3. The results in the cases of 5 and 10 are left out because only too fewer evaluation objects are extracted and the detection performance excessively depends on them. Each figure has 3 subfigures related to the change ratio in the prediction phase: 0.01, 0.02, and 0.025. Each method in each subfigure has 3 bar graphs corresponding to the precision, the recall, and the F-measure.

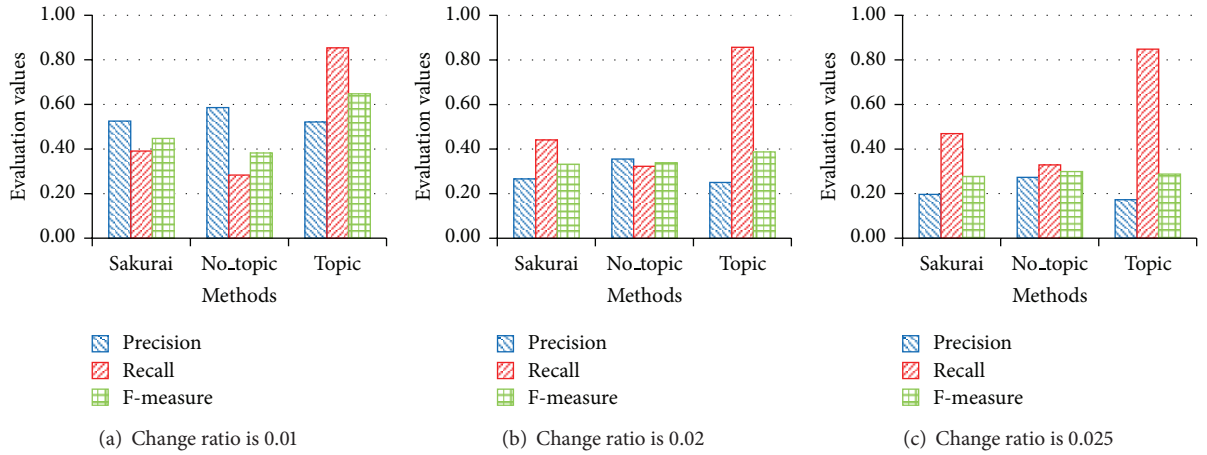
6. Discussions

This section discusses the effect of the proposed method with four viewpoints: the expansion of training transactions, the discovered trend rules, the extracted evaluation objects, and the detection performance.

6.1. Expansion of Training Transactions. Figure 5 shows that the number of training transactions increases. "No_topic" is about 2.9 times as large as "Sakurai." The expansion ratio is close to the expansion ratio of stock brands. "Topic" is about 25 times as large as "Sakurai" and is about 8.7 times as large as "No_topic." We can confirm that the topic dictionary expands the training transactions. The topic dictionary gets the effect of the expansion.

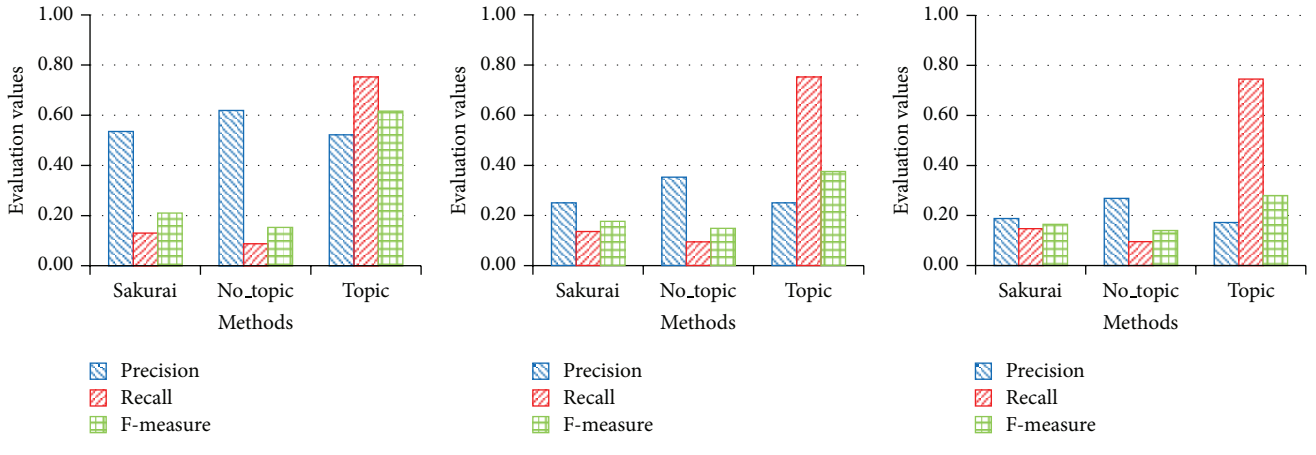
6.2. Discovered Trend Rules. Figures 6 and 7 show that the number of trend rules decreases in "Topic." We think that the results are caused by relatively small support of patterns. That is, "Topic" deals with additional 2,300 evaluation objects by referring to all stock brands. Also, it deals with more attributes than the previous research does because more training transactions are generated from more news headlines. Relative frequency of each item tends to decrease as the kind of items increases. The supports of the patterns tend to be small. On the other hand, some experimental results show that good detection performance is not given when very few trend rules are used in the prediction phase. In near future, it may be necessary to consider the method that decides the number of trend rules for the good detection performance.

6.3. Extracted Evaluation Objects. Figure 8 shows that the number of extracted evaluation objects increases. Particularly, "Topic" extracts more evaluation transactions than the other methods do. This is because many evaluation objects are assigned to evaluation transactions by the topic dictionary.



(a) Change ratio is 0.01 (b) Change ratio is 0.02 (c) Change ratio is 0.025

Figure 9: Detection performance; minimum transaction is 1.



(a) Change ratio is 0.01 (b) Change ratio is 0.02 (c) Change ratio is 0.025

Figure 10: Detection performance; minimum transaction is 3.

It is anticipated that the recall increases as the number of extracted objects increases.

On the other hand, the proposed method cannot identify original evaluation objects directly described in news headlines with the ones based on the topic dictionary. The prediction phase may be able to extract more valid evaluation objects if different kinds of evaluation objects are identified and their difference is reflected on the prediction phase. In near future, we will try to reconsider the activation method of the topic dictionary.

6.4. Detection Performance. Figures 9 and 10 show that "Topic" gives precisions which are similar to "Sakurai" but "Topic" gives more or less smaller precisions than "No_topic" does. On the other hand, "Topic" drastically improves recalls. It gives the best F-measures in the most cases. For example, when we focus on the result in Figure 9(b), the result shows that "Topic" is 14.3% higher than "Sakurai" and "Topic" is 12.7% higher than "No_topic." The drastic improvement of recalls contributes to the improvement of F-measures. We think that the activation of the topic dictionary contributes to the improvement of the detection performance even if the precisions more or less deteriorate.

According to the above discussions, we believe that the proposed expansion method can acquire more valid trend rules.

7. Summary and Future Works

This paper proposed a new method for the expansion of training data. The expansion can lead to the discovery of more valid trend rules from complex sequential data. The method was applied to the prediction task of attractive stock brands in the next period. This paper verified the effect of the method through the comparison with the previous method.

In our future work, we will try to improve detection performance. For example, we are planning to reconsider the transformation method of evaluation objects based on the

topic dictionary. This is because the method cannot identify original evaluation objects included in texts with the ones added by the topic dictionary. Also, we are planning to activate certainty of topics in the topic dictionary. The certainty can be updated for each unit time. It can flexibly evaluate relationships between evaluation objects and topics. It is anticipated that these improvements lead to the acquisition of more valid trend rules. Also, they lead to the improvement of the detection performance.

On the other hand, we will try to apply the method to other application fields such as smart community field and healthcare field. For example, in the smart community field, Twitter messages, amount of electrical power consumption, and communities are text sequential data, numerical sequential data, and evaluation objects. In the healthcare field, nursing texts, test values of medical examination, and patients correspond to them. Various application fields will be considered. Through the application to the various fields, we will evaluate the effect of the proposed method in detail.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the 1995 International Conference on Knowledge Discovery and Data Mining*, pp. 3–14, March 1995.

[2] J. Pei, J. Han, B. Mortazavi-Asl et al., "PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth," in *Proceedings of the 17th International Conference on Data Engineering*, pp. 215–224, April 2001.

[3] S. Sakurai and R. Orihara, "Discovery of important threads from bulletin board sites," *International Journal of Information Technology and Intelligent Computing*, vol. 1, no. 1, pp. 217–228, 2006.

[4] S. Sakurai and K. Ueno, "Analysis of daily business reports based on sequential text mining method," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '04)*, vol. 4, pp. 3279–3284, October 2004.

[5] S. Yen, "Mining interesting sequential patterns for intelligent systems," *International Journal of Intelligent Systems*, vol. 20, no. 1, pp. 73–87, 2005.

[6] S. Sakurai, K. Makino, and S. Matsumoto, "A discovery method of trend rules from complex sequential data," in *Proceedings of the 26th IEEE International Conference on Advanced Information Networking and Applications Workshops (AINA '12)*, pp. 950–955, 2012.

[7] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of Internet stock message boards," *Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," October 2010, http://arxiv.org/PS_cache/arxiv/pdf/1010/1010.3003v1.pdf.

[9] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting Stock Market Indicators through Twitter, 'I hope it is not as bad as I fear'," *Procedia*, vol. 26, pp. 55–62, 2011.

[10] G. P. C. Fung, J. X. Yu, and W. Lam, "News sensitive stock trend prediction," in *Proceedngs of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 481–493, 2002.

[11] M. Mittermayer and G. F. Knolmayer, "NewsCATS: a news categorization and trading system," in *Proceedings of the 6th International Conference on Data Mining*, pp. 1002–1007, December 2006.

[12] D. Peramunetilleke and R. K. Wong, "Currency exchange rate forecasting from news headlines," in *Proceedings of the 13th Australasian Database Conference*, vol. 5, pp. 131–139, 2002.

[13] M. de Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Can blog communication dynamics be correlated with stock market activity?" in *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia (HT '08)*, pp. 55–60, June 2008.

[14] Y. Seo, J. A. Giampapa, and K. P. Sycaratech, "Financial news analysis for intelligent portfolio management," Report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, 2004.

[15] S. Sakurai, K. Makino, H. Suzuki, and Y. Masaoka, "Ranking of evaluation targets based on complex sequential data," in *Proceedings of the 25th Annual Conference of the Japanese Society for Artificial Intelligence*, 2G2-01, 2011, (Japanese).

[16] Chasen, 2010 (Japanese), http://chasen.naist.jp/hiki/ChaSen/.

[17] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, vol. 29, no. 2, pp. 1–12, 2000.