International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:2, 2014

A Prediction of Attractive Evaluation Objects Based On Complex Sequential Data

Shigeaki Sakurai, Makino Kyoko, Shigeru Matsumoto

Abstract—This paper proposes a method that predicts attractive evaluation objects. In the learning phase, the method inductively acquires trend rules from complex sequential data. The data is composed of two types of data. One is numerical sequential data. Each evaluation object has respective numerical sequential data. The other is text sequential data. Each evaluation object is described in texts. The trend rules represent changes of numerical values related to evaluation objects. In the prediction phase, the method applies new text sequential data to the trend rules and evaluates which evaluation objects are attractive. This paper verifies the effect of the proposed method by using stock price sequences and news headline sequences. In these sequences, each stock brand corresponds to an evaluation object. This paper discusses validity of predicted attractive evaluation objects, the process time of each phase, and the possibility of application tasks.

Keywords—Trend rule, frequent pattern, numerical sequential data, text sequential data, evaluation object.

I. Introduction

As many kinds of sensors are smaller and cheaper, they are more easily buried in the real world environment. Also, social media represented by Twitter and Facebook is to be recognized as social sensors. In near future, we think that they are tied to each other and they compose sensor networks. Large amount of sequential data will be collected through the networks. We can anticipate that the analysis of the data leads to the improvement of our daily life.

According to this background, many analysis methods of sequential data have been proposed, [9] proposes a method that discovers sequential patterns from text sequential data. The method extracts events representing texts from the data and generates event sequential data. It can pick up sequential patterns satisfying constrains based on the interests of users. Reference [10] proposes a new evaluation criterion measuring the interestingness of sequential patterns. The criterion can evaluate future relationships between sequential sub-patterns included in a sequential pattern. Also, the paper proposes a discovery method of patterns based on the criterion. Reference [11] proposes a method that discovers sequential patterns from sequential data with the tabular structure. In the case of the data, each item composing of the patterns is composed of an attribute and its attribute value. The method can efficiently discover the patterns by referring to relationships between attributes and attribute values. The method is applied to the medical examination data of employees in a company and its effect is verified.

S. Sakurai, K. Makino, and S. Matsumoto are IT Research and Development Center, Toshiba Solutions Corporation, Tokyo, 183-8512 Japan, (e-mail: {Sakurai.Shigeaki, Makino.Kyoko, Matsumoto.Shigeru}@toshiba-sol.co.jp).

Even if these methods can deal with large amount of sequential data, their time constraint for the data process is not so strict. Also, they cannot simultaneously deal with various data formats. That is, their target data is composed of only categorical data or only text data. Therefore, the methods cannot real-timely process complex sequential data occurring from the real world environment and the network environment. It is necessary to develop a method dealing with the complex sequential data.

In this paper, we focus on complex sequential data related to evaluation objects. The data is composed of numerical sequential data and text sequential data. For example, stock price sequences of each stock brand are an example of the former one. News headline sequences are an example of the latter one where texts in the sequences include company names corresponding to stock brands. Then, the evaluation objects are companies. This paper proposes a method that discovers trend rules from the data. The trend rules are used to predict attractive evaluation objects. This paper applies the method to the stock price sequences and the news headline sequences collected from Web sites. It verifies the effect of the proposed method through experiments.

In the remaining parts of this paper, Section II introduces related works dealing with complex sequential data in financial field. Section III proposes a method discovering trend rules. Also, it proposes a method predicting attractive evaluation objects based on the trend rules. Section IV explains the experimental data, the evaluation criteria, and the evaluation method. It shows experimental results and discuses the effect of the proposed method. Lastly, Section V summarizes this paper and shows the future research direction.

II. Related works

This section introduces some related works based on financial data composed of numerical data and text data. Reference [1] investigates relationships between the stock data of 45 companies in the Dow Jones Industrial Average (DJIA) and more than 150 million messages. The messages are collected from Yahoo! Finance and Rating Bull which is one of financial communities. The paper measures the bullishness included in the messages by using the power of the computational linguistics. The paper shows that the messages can explain the volatility which is one of evaluation criteria for stock prices to some extent. On the other hand, it shows that the messages cannot help to gain the revenue in the trade operations, because inconsistent messages are included in them. Reference [2] proposes a method that inductively

learns relationships between the 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 the DJIA with 87.6% accuracy. Reference [14] proposes a method predicting stock market indicators such as DJIA and National Association of Securities Dealers Automated Quotations (NASDAQ), and Standard & Poor's 500 Stocks Average (S & P 500) based on Twitter posts. It analyzes correlations between the stock market indicators and two collective emotions such as "fear" and "hope". It finds that the emotional tweet percentage is negatively related to the indicators and is positively related to Volatility Index (VIX). Reference [5] proposes a method predicting the changes of the trend in the stock market. The method segments numerical stock price data into three trends such as "Rise", "Steady", and "Drop". Also, it assigns news articles to the two trends "Rise" and "Drop" by referring to the distribution time of the articles. The articles assigned to each trend are segmented to two clusters, respectively. The similar articles in the respective trends are regarded as news articles unrelated to the trends. They are removed from original news articles. The method inductively learns respective classification models of the two trends from remaining news articles. The classification models are used in order to predict the change of stock prices in the stock market. Reference [7] proposes 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. The appropriate training data is selected in order to construct the prediction model. The paper shows that the method arrives at the high prediction performance. Reference [8] proposes a method that acquires classification rules. Their antecedent part is weighted keywords and their result part is classes discretizing changes of currency exchange. The method uses keywords selected by stock traders or human experts. The weights are calculated by referring to frequencies of the keywords in the target period. The classes are decided by referring to sequences of changes of currency exchange in the next period. Lastly, the method selects classification rules based on the weights. Reference [4] develops a model analyzing communication dynamics in the blogosphere in order to decide correlations with movement in stock market. The model is acquired by Support Vector Machine (SVM) regression and is characterized by some simple features such as the number of posts, the number of comments, the length and response time of comments, and so on. The paper shows that the models for selected companies can predict the magnitude of the movement about 78% accuracy and its direction about 87% accuracy. Reference [13] develops the intelligent multi-agent system for the portfolio management. The system acquires a classification model classifying news articles related to financial information of companies into 5 classes: "Good", "Good or Uncertain", "Neutral", "Bad or Uncertain", and "Bad". The model is inductively learned by using the semi-supervised technique. Each agent collects the stock price information, the financial information, and the news articles. Also, it evaluates the news articles based on the model. The system manages the portfolio by totally evaluating the collected information.

Some existing methods 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 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. This paper tries to consider an analysis method overcoming to this problem.

III. Analysis of complex sequential data

A. Problem Setting

The complex sequential data is collected from various information sources. It is composed of various types of data. This paper focuses on one of the types composed of both numerical sequential data and text sequential data related to evaluation objects. Also, it focuses on the prediction of attractive evaluation objects in the next period. Here, evaluation objects related to changes of trends are regarded as the attractive evaluation objects. This is because the detection of the changes is one of important tasks and the prediction can help our decision making in various application fields.

We focus on the change of the data. Then, we can think two types of changes in the case of the complex sequential data. One is the change of numerical one and the other is the change of text one. The analysis of the latter one is more difficult than the former one. This is because the text data has various meanings and is unstructured. It is difficult to easily and automatically interpret it. In our first step, we tackle on the change of the numerical data.

In order to understand the change, it is important to seek the cause of the change, when we observe it. This is because it is necessary to perform the countermeasure depending on the cause. We may be able to seek the cause to the other numerical data. However, many previous researches have still proposed methods discovering relationships between numerical sequential sequences. On the other hand, as far as we know, there are not so many methods simultaneously analyzing both text sequential data and numerical sequential data. Therefore, we can anticipate that we acquire a new type of knowledge by using a new type of information source. Thus, this paper assumes that the change of the numerical data can be interpreted by the text data. It tackles on the development of the method analyzing the complex sequential data. In the following subsections, this paper proposes the method composed of trend rule learning and prediction based on trend rules.

B. Trend Rule Learning

This subsection explains the learning method of trend rules. The method inductively acquires trend rules according to the outline as shown in Fig. 1. That is, it is composed of five subprocesses: 1) the extraction of evaluation objects and attributes, 2) the calculation of a change ratio, 3) the

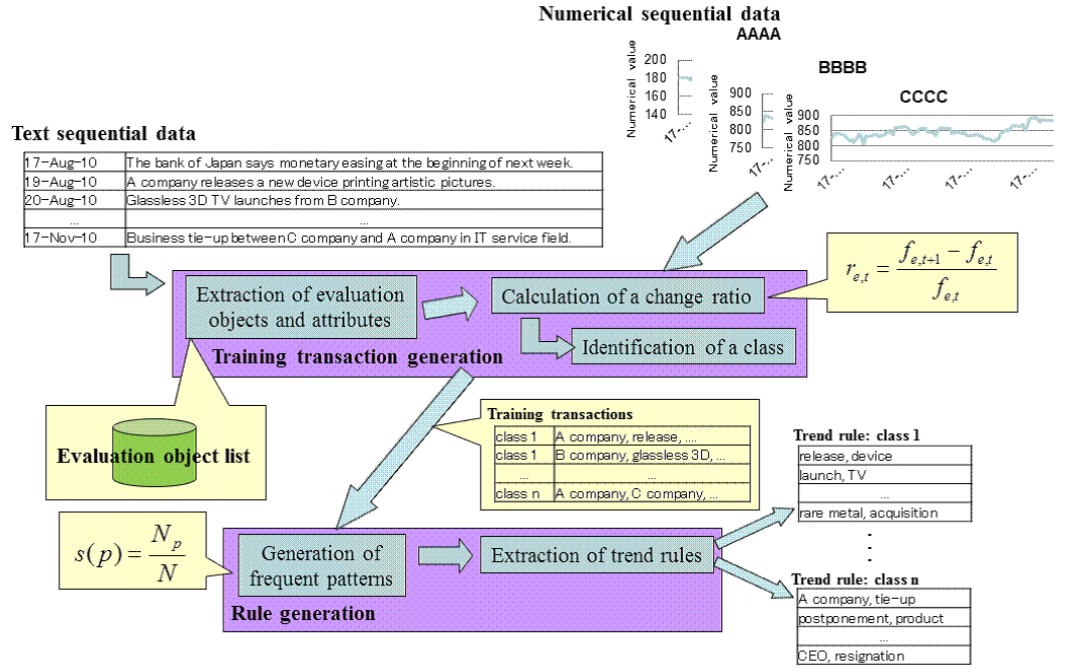


Fig. 1. An outline of trend rule learning

identification of a class, 4) the generation of frequent patterns, and 5) the extraction of trend rules.

The first subprocess applies a morphological analysis engine to a text in text sequential data. The engine separates the text to words and analyzes their parts of speech. This subprocess uses Chasen [3] which is one of the engines. This engine can deal with Japanese texts. Proper nouns related to organization and other nouns are extracted from the text. The proper nouns are evaluated whether they are evaluation objects by referring to the evaluation object list. The proper nouns except evaluation objects and the other nouns are regarded as attributes. A training transaction composed of the evaluation objects and the attributes is generated. But, if a text does not include an evaluation object, a training transaction cannot be generated from it. The second subprocess finds numerical sequential data of an extracted evaluation object. Two numerical values are picked up from the numerical sequential data by referring to the time stamp of the text. One is a value corresponding to the time stamp and the other is a value corresponding to the next time stamp. Their change ratio is calculated by referring to (1). In this equation, $f_{e,t}$ is a numerical value corresponding to both an evaluation object e and a target time t.

$$r_{e,t} = \frac{f_{e,t+1} - f_{e,t}}{f_{e,t}} \quad (1)$$

The third subprocess identifies a class of the text including the evaluation object. It uses disjoint ranges of change ratios representing classes. The ranges are given by a user in advance. This subprocess decides a range including the calculated change ratio and identifies a class. The training transaction is assigned a identified class. The generation of training transactions is repeated for all texts in the text sequential data. The fourth subprocess applies each training transaction set for each class to the discovery method of frequent patterns. Here, a frequent pattern is a frequent combination of evaluation objects or attributes. If the support of the combination is larger than or equal to the minimum support given by a user, the combination is regarded as a frequent pattern. The support is calculated by (2). In this equation, p is a pattern, N_p is the number of training transactions including p, and N is the number of training transactions.

$$s(p) = \frac{N_p}{N} \quad (2)$$

This subprocess uses the discovery method based on FP-trees [6] which represent training transactions with tree format. The method can efficiently discover all frequent patterns. Lastly, the fifth subprocess generates the combination of a frequent pattern and a class. The combination is regarded as a trend rule.

This paper discovers frequent patterns without taking consideration into the difference of classes. On the other hand, we can use the discovery method which directly deals with transactions with classes [12]. The method can avoid discovering patterns that are related to many classes or that are related to a class including many transactions. We can anticipate that the method discovers more valid patterns representing classes. However, in the case of collected experimental data sets, there is a great deal of disparity in the numbers of training transactions related to classes. The disparity leads to the increase of discovery time. Therefore, this paper does not use the method. In our future works, we

will consider easing the increase and try to apply the method to the data sets.

C. Prediction Based On Trend Rules

This subsection explains the prediction method based on trend rules. The method predicts attractive evaluation objects according to the outline as shown in Fig. 2. That is, it is composed of four subprocesses: 1) the extraction of evaluation objects and attributes, 2) the matching based on trend rules, 3) the accumulation of evaluation transactions, and 4) the extraction of attractive evaluation objects.

The first subprocess applies the morphological analysis engine to texts in distributed text sequential data. Proper nouns and other nouns are extracted. An evaluation transaction is generated from each text. It is composed of evaluation objects or attributes, but its class is not identified. The extraction corresponds to the one in the learning phase. The second subprocess compares a set of nouns in an evaluation transaction with a set of nouns in a trend rule. If the set of nouns in the trend rule is included in the one in the evaluation transaction, the evaluation transaction matches to the trend rule. The class of the trend rule is accumulated for the evaluation transaction. The matching is performed for all trend rules. This subprocess evaluates the number of accumulated classes for the evaluation transaction. The class with the maximum number is assigned to the evaluation transaction. If all trend rules do not match to the evaluation transaction, the evaluation transaction is not dealt with the following subprocesses. The third subprocess assigns the evaluation transaction to evaluation objects included in it. The number of evaluation transactions is counted up for each evaluation object and each class. These subprocesses are repeated until the next time stamp. Lastly, the fourth process evaluates whether evaluation objects are attractive or not by referring to their numbers of assigned evaluation transactions. The attractive evaluation objects are shown to a user. The user decides whether he/she performs countermeasures related to the attractive evaluation objects.

IV. Experiments

This section explains experiments based on real complex sequential data. The data is composed of stock price sequences and news headline sequences. The evaluation objects are stock brands. This section explains the experimental data, the evaluation criterion, the evaluation method, the experimental results, and the discussions in order.

A. Experimental Data

This paper collects news headline sequences as the text sequential data. The sequences are collected from five news distribution sites: Excite, Goo, Infoseek, Livedoor, and Yahoo. These sites deal with news headlines described in Japanese. The news headlines are collected in three intervals: August 28, 2010 $\sim$ January 31, 2011 (D1), February 1, 2011 $\sim$ April 6, 2011 (D2), and April 7, 2011 $\sim$ May 22, 2011 (D3). Also, the interval D2 is divided into two sub-intervals: February 1, 2011 $\sim$ March 10, 2011 (D2a) and March 11, 2011 $\sim$ April 6, 2011 (D2b). This is because it is anticipated that the big earthquake occurred in East Japan on March 11, 2011 gives a big impact to the change of stock prices and trend rules before the earthquake is different from the ones after it. Table I shows the numbers of collected news headlines. Each news headline has related information such as date, time, site name, and genre.

TABLE I
Collected news headlines

	D1	D2a	D2b	D3
Excite	132,878	38,761	26,993	44,580
Goo	143,062	30,593	22,070	22,269
Infoseek	240,141	62,407	49,755	68,927
Livedoor	233,773	66,740	46,904	75,393
Yahoo	253,619	70,184	50,037	83,556
Total	1,003,473	268,685	195,759	294,719

On the other hand, this paper collects stock price sequences of stock brands as the numerical sequential data. The sequences are collected from the storage site of stock price [15]. This site stores daily stock prices and daily stock market turnover for each stock brand in the Tokyo Stock Market over 250 business days. The data corresponding to each stock brand includes stock brand code, date, opening price, highest price, lowest price, closing price, and stock market turnover. It is stored with csv format. Its collection interval includes the ones of the news headline sequences. In the following experiment, this paper uses the opening price in order to decide a class.

This complex sequential data regards each stock brand as an evaluation object. This experiment deals with 1,680 stock brands included in the first section of the Tokyo Stock Market. The stock brand list is collected from two pages: [16] and [17].

B. Evaluation Criterion

It is anticipated that stock traders are interested in the changes of stock prices. This is because the stock traders can decide an appropriate trade operation by referring to the changes. Also, they cannot simultaneously pay attention to a lot of stock brands. It is sufficient for the recommendation task of the stock brands to recommend only parts of attractive stock brands. Even if some attractive stock brands are missed, the stock traders do not always care about the miss. On the other hand, it is important for the task to recommend valid stock brands. This is because the stock traders directly look at the recommended stock brands. If most parts of them are not valid, the stock traders may think that the recommendation is not valid. Therefore, this paper uses the precision defined by (3) as an evaluation criterion. The precision evaluates the precise ratio of recommended stock brands. But, this paper regards evaluation objects whose future change ratios are big as attractive evaluation objects. This experiment deals with daily stock prices. We evaluate whether each stock brand is attractive by referring to the stock price of the next day. If the attractive evaluation objects are recommended, recommended

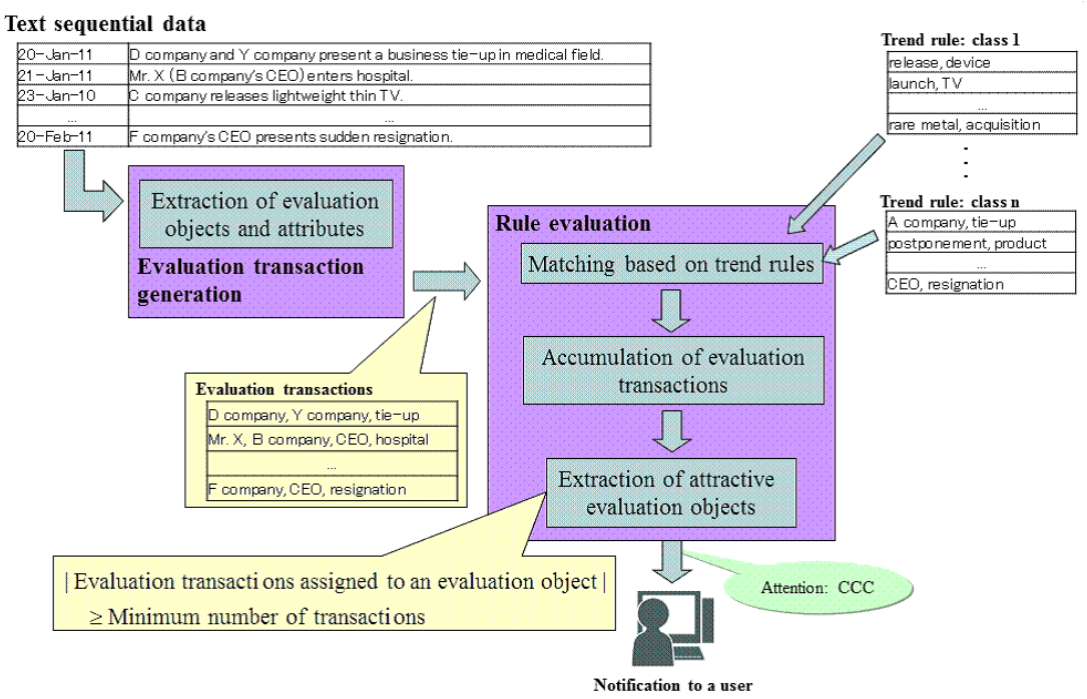


Fig. 2. An outline of prediction based on trend rules

evaluation objects are regarded as truly recommended ones.

$$\frac{\text{Number of truly recommended evaluation objects}}{\text{Number of recommended evaluation objects}} \quad (3)$$

Next, we consider the process time as other evaluation criterion. The criterion is selected due to the following three reasons. Firstly, the learning method requires dealing with more training transactions in order to acquire more valid trend rules. It is anticipated that the number of training transactions increases more and more. Secondly, the prediction method is required to process news headlines with high speed. This is because many news headlines are real-timely distributed from many distribution sites. Thirdly, it is required to evaluate the changes of stock prices within shorter time interval in near future, even if this experiment deals with daily changes. This is because the stock traders may repeatedly operate the same stock brand in a day. Therefore, the process time is an important criterion evaluating the possibility of real-time process. This paper simply measures the process time in order to roughly grasp it. That is, the DOS command "time" is run before and after running target processes where the command can measure the time with 100 millisecond unit. Also, the usual works are simultaneously performed on the same computer. The difference of the time displayed by the command is regarded as the process time. This experiment uses such a computer environment that the operating system is the Microsoft Windows XP Professional Version 2002 Service Pack 3 and the hardware is Dell Optiplex 960 including Intel Core2 Quad CPU Q9550 @ 2.83GHz 1.98GHz 3.25GB RAM.

C. Evaluation Method

Most of news headlines after March 11, 2011 are related to the earthquake due to the big earthquake occurred in East Japan. Our preliminary experiment shows that most of trend rules are related to the earthquake. However, the trend rules are not always usual trend rules because the big earthquake is one of rare events. The trend rules may not be appropriate in order to evaluate the effect of the proposed method. Thus, this experiment uses the data set D1 as the learning one and uses the data set D2a as the evaluation one in the case of the precision. On the other hand, it uses the data set D3 in order to evaluate the process time because the data set includes many news headlines.

This experiment classifies the changes of stock prices into three classes: "Drop", "Steady", and "Rise". "Drop", "Steady", and "Rise" correspond to three ranges: $(-\infty, -5\%]$, $(-5\%, +5\%]$, and $(+5\%, +\infty)$. We think that the stock traders are interested in two classes of them: "Rise" and "Drop". Also, we think that they would like to judge whether recommended stock brands are valid and decide which trade operations should be selected. Therefore, the difference of "Rise" and "Drop" is not so important. Thus, this experiment deals with only trend rules related to "Rise" and "Drop". Also, the prediction of stock brands does not care about their difference. That is, stock brands are selected as attractive stock brands if the total numbers of evaluation transactions with their classes assigned to them are larger than or equal to a predefined threshold.

This paper compares the proposed method with the random method. The random method can give performance of base line. It randomly selects stock brands from all stock brands

by referring to a probability of the changes of stock prices. The probability is calculated by (4). Here, each stock brand is evaluated which class should be assigned in a day. The probability is the average value of all stock brands in the evaluation interval. It is anticipated that the precision of the proposed method is larger than the one of the random method.

$$\frac{\text{the number of stock brands assigned “Rise” and “Drop”}}{\text{the number of stock brands}} \quad (4)$$

D. Experimental Results

Table II shows training transaction sets for each site and each class. The sets are generated from news headlines in the data set D1. This table shows that evaluation objects are not picked up from many news headlines. We can consider two cases for no evaluation objects. One case shows that news headlines really do not include evaluation objects. The other case shows that the extraction of proper nouns fails to extract evaluation objects. The improvement of the extraction may lead to acquire more valid trend rules due to the use of many training transactions. In our future works, we will try to consider the improvement of the extraction.

TABLE II
TRAINING TRANSACTION SETS

	Rise	Steady	Drop	No evaluation object
Excite	58	7,096	75	123,390
Goo	128	7,576	101	132,740
Infoseek	208	17,312	151	219,865
Livedoor	166	13,918	168	214,773
Yahoo	298	20,464	389	228,163
Total	858	66,366	884	918,931

Table III shows evaluation transaction sets for each site. The sets are generated from news headlines in the data sets D3 and D2a. They are not directly identified classes by referring to the changes of stock prices, but the prediction based on trend rules identifies the classes. Only evaluation transactions whose classes are identified as “Rise” and “Drop” contribute to the recommendation of stock brands.

TABLE III
EVALUATION TRANSACTION SETS

	D3	D2a
Excite	41,926	37,440
Goo	20,905	29,520
Infoseek	64,677	60,129
Livedoor	70,935	64,209
Yahoo	78,765	67,835
Total	277,208	259,133

Fig. 3 shows the number of trend rules extracted from the data set D1. Fig. 3 (a) and (b) correspond to two classes “Rise” and “Drop”, respectively. In these graphs, horizontal axes show the minimum supports and vertical axes show the number of extracted trend rules. Each bar graph is accumulated the number of trend rules according to the number of nouns included in trend rules. These graphs show that the numbers of trend rules dramatically decrease as the minimum supports increase. We do not have the background knowledge how minimum supports are valid for the prediction of evaluation objects. We evaluate the validity of the prediction by applying some minimum supports: 0.005, 0.010, 0.020, and 0.030 to the trend rule learning.

Fig. 4 shows the number of evaluation objects whose classes are “Rise” or “Drop”. Each evaluation object is evaluated for each day included in the interval D2a. This figure shows the accumulated numbers of evaluation objects included in the classes. In this figure, a horizontal axis shows the change ratio of stock prices and a vertical axis shows the number of the evaluation objects. Here, change ratios of stock prices are smaller than or equal to the change ratio of stock prices in the case of the trend rule learning. This is because we think that news headlines related to high change ratio give a bigger impact. We can anticipate that the more valid trend rules are acquired from them.

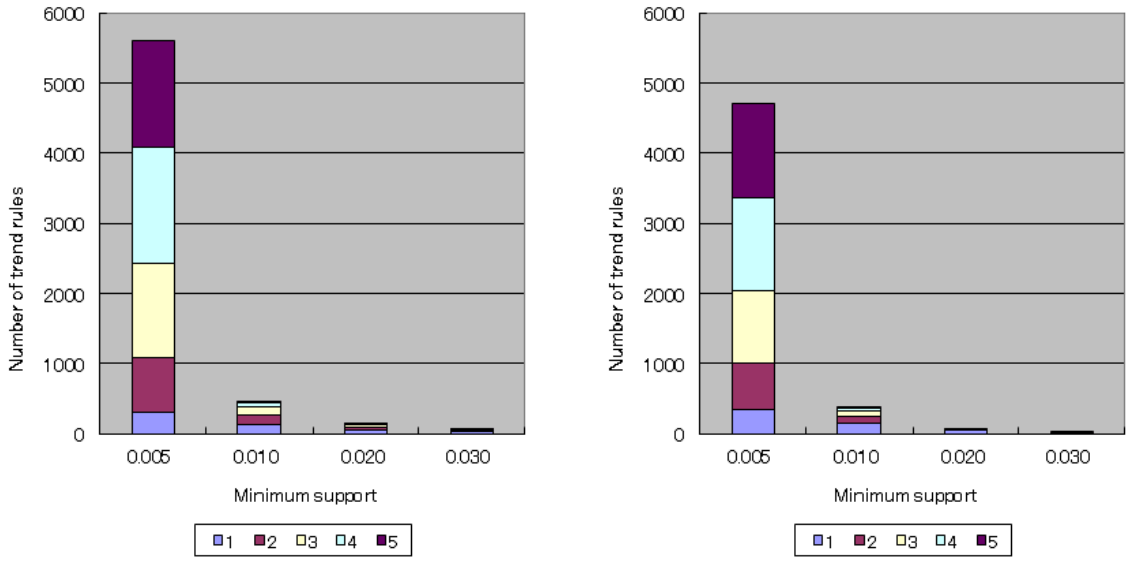
Fig. 5 shows precisions of the proposed method and the random method. In this experiment, the data set D1 and D2a are used for the learning and the prediction, respectively. In this figure, a horizontal axis shows the change ratio of stock prices and a vertical axis shows the precision. Three parameters of the proposed method are adjusted. They are the minimum support, the minimum number of nouns, and the minimum number of transactions. The prediction applies only trend rules whose numbers of nouns are larger than or equal to the minimum number of nouns to evaluation transactions. Also, it extracts evaluation objects whose total numbers of assigned evaluation transactions are larger than or equal to the minimum number of transactions as attractive evaluation objects. SxxIyTz represents a parameter set. Sxx means that the minimum support is 0.0xx, Iy does that the minimum number of nouns is y, and Tz does that the minimum number of transactions is z.

Fig. 6 shows the number of evaluation objects extracted as attractive evaluation objects. In this figure, a horizontal axis shows the parameter set and a vertical axis shows the number of extracted evaluation objects.

Table IV shows the process time in the case of the learning and the prediction. Here, the prediction deals with the data set D3. In the learning, the training transaction generation process and the rule generation process are separately run due to the constraint of our experimental environment. Similarly, in the prediction, the evaluation transaction generation and the rule evaluation process are separately run. Each transaction generation process has the common extraction subprocess of evaluation objects and attributes.

TABLE IV
PROCESS TIME

	Process	Time(second)
Learning	Training transaction generation	1,607
	Rule generation	148
Prediction	Evaluation transaction generation	1,458
	Rule evaluation	357



(a) Rise (b) Drop

Fig. 3. Number of extracted trend rules

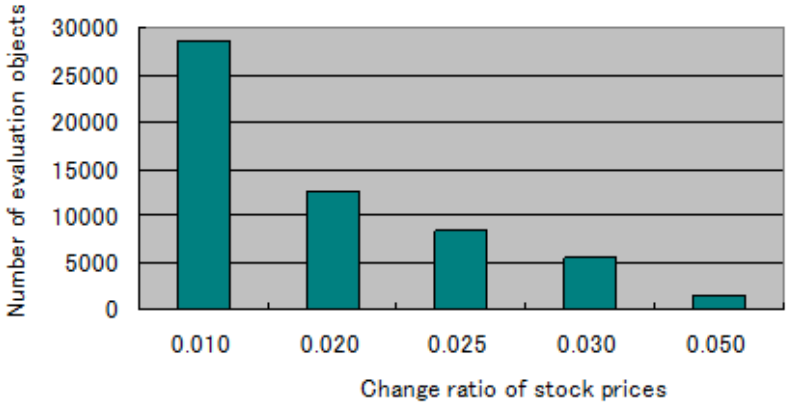


Fig. 4. Number of evaluation objects due to change ratios of stock prices

E. Discussions

This subsection discusses the effect of the proposed method with three viewpoints: the validity of prediction, the process time, and the variety of application tasks.

1) Validity of Prediction: In the case of the parameter set S30I2T1, the number of trend rules is very few and limited evaluation objects are extracted. Each trend rule gives an excessive impact to the precision. We think that small number of the trend rules leads to their deterioration. Also, Fig. 5 shows that the proposed method in most cases exceeds the random method. We think that the proposed method is better than the random method.

On the other hand, we think that the change ratio of stock prices 0.25 or 0.30 is valid in the case of the prediction. This is because the number of the evaluation objects regarded as attractive ones is moderate. These cases show that their precisions are distributed in [0.15, 0.32]. We cannot insist that the precisions are sufficiently high. We think that it is necessary to improve the precisions in near future.

We note two related works [2] and [8]. Reference [2] shows that 87.6% parts of the changes of stock prices in DJIA are explained by the emotion “Calm” extracted from Twitter messages. It deals with the synthetic changes of stock prices and is easier than the task in this paper. This is because the task deals with many respective changes of stock prices. However, it is important that the analysis of 10 millions messages brings in high precision. We can anticipate that the increase of training transactions leads to higher precision. On the other hand, [8] shows that rules related to three classes “Rise”, “Steady”, and “Drop” are discovered and they can predict the

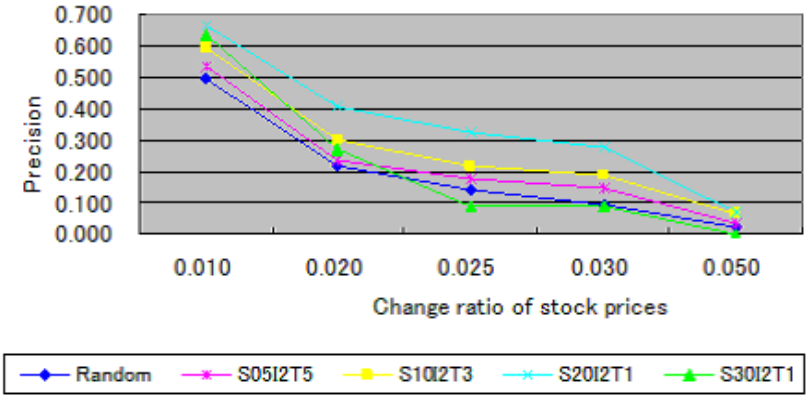


Fig. 5. Proposed method vs. random selection

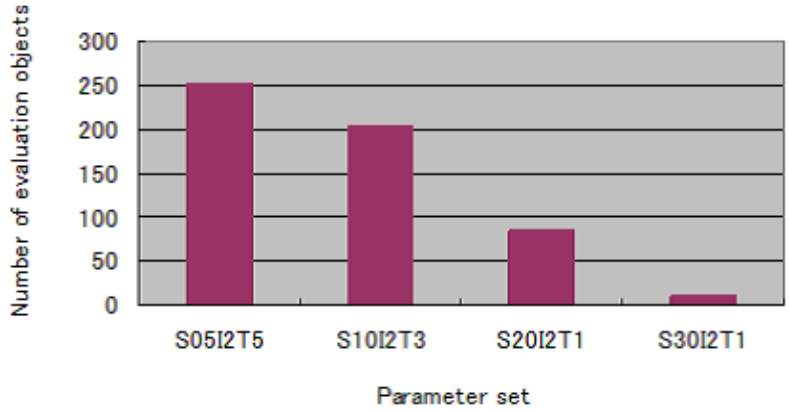


Fig. 6. Number of extracted evaluation objects

classes with over 50% probability, even if [8] discovers only rules related to limited brands of currency exchange. Also, it uses keyword pairs manually designated by human experts in order to discover the rules. We can note the use of the background knowledge. If we use the knowledge in order to identify classes or evaluation objects, we can anticipate higher precision ratio. In our future works, we will tackle on these improvements.

2) Process Time: Firstly, we note the extraction subprocess of evaluation objects and attributes. We think that the subprocess can be performed in parallel by applying it to each news headline. The parallel process can realize shorter process time. In our other research, the effect of the parallel process based on Hadoop is verified to some extent. Also, we think that it is possible to perform the rule evaluation process in parallel with two viewpoints. That is, one is the division of trend rules and the other is the division of news headlines. The prediction of attractive evaluation objects can real-timely be processed.

Next, we note the rule generation process. It is not easy to real-timely perform the process, even if training transactions can be easily divided based on classes. This is because the number of classes is small and the effect of the parallel process is limited. If we aim at giving the effect of high parallel process, it is necessary to in parallel discover frequent patterns. The parallel process requires comparatively complicated calculation process. On the other hand, it is not always necessary to real-timely update trend rules. This is because the trend is comparatively stable and does not real-timely change. We think that the parallel process of the learning is not important in near future.

F. Variety of Application Tasks

This paper applies the proposed method to the application task which predicts attractive stock brands. However, the method is not limited to the application task. For example, it can be applied to a smart community field. In the case of this field, evaluation objects, numerical sequential data, and text sequential data are areas, consumption sequences of electricity, and message sequences in community bulletin boards site. The method may be able to smart control the consumption of electricity in the areas. Also, the method can be applied to a healthcare field. Persons, result sequences of physical

examination, and comment sequences by doctors are dealt with in order to improve the health conditions of the persons. In addition, the method can be applied to a machine maintenance field. Target equipment, its measurement sequences of tests, and comment sequences by maintenance persons are dealt in order to early detect broken equipment. The proposed method has various application fields.

According to these discussions, we believe that the proposed method is efficient.

V. Conclusion

This paper proposed an analysis method of complex sequential data. The data is composed of numerical sequential data and text sequential data. This paper applied the method to real data sets collected from Web sites. In the data sets, evaluation objects, text sequential data, and numerical sequential data are stock brands, news headline sequences, and stock price sequences, respectively. This paper verified the effect of the proposed method by comparing it with the random method. In addition, this paper showed the possibility of real-time prediction and the one of various application tasks.

We will try to improve the precision. For example, we will do it by the increase of training transactions. This is because large amount of training data tends to acquire more valid rules. Also, we will do it by the improvement of the rule generation process. On the other hand, we will try to in detail verify the effect of the proposed method. For example, we will evaluate based on the other evaluation criteria and compare the proposed method with existing related researches. Lastly, we will try to apply the proposed method to other application fields such as smart communities, healthcare, and machine maintenance. We believe that the proposed method will become more attractive by these improvements and the expansion of the application fields.

References

[1] W. Antweiler and M. Z. Frank *Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards*, J. of Finance, vol.59, no.3, pp.1259-1294, 2004.
[2] J. Bollen, H. Mao, and X. -J. Zeng, *Twitter Mood Predicts the Stock Market*, http://arxiv.org/PS_cache/arxiv/pdf/1010/1010.3003v1.pdf, October, 2010.
[3] Chasen, http://chasen.naist.jp/hiki/ChaSen/, 2010 (in Japanese).
[4] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann, *Can Blog Communication Dynamics be Correlated with Stock Market Activity?*, Proc. of the 19th ACM Conf. on Hypertext and Hypermedia, 2010.
[5] G. P. C. Fung, J. X. Yu, and W. Lam, *News Sensitive Stock Trend Prediction*, Proc. of the 6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, pp.481-493, 2002.
[6] J. Han, J. Pei, and Y. Yin, *Mining Frequent Patterns without Candidate Generation*, Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data, pp.1-12, 2000.
[7] M. -A. Mittermayer and G. F. Knolmayer, *NewsCATS - A News Categorization and Trading System*, Proc. of the 6th IEEE Intl. Conf. on Data Mining, pp.1002-1007, 2006.
[8] D. Peramunetilleke and R. K. Wong, *Currency Exchange Rate Forecasting from News Headlines*, Proc. of the 13th Australasian Database Conf., vol.5, pp.131-139, 2002.
[9] S. Sakurai and K. Ueno, *Analysis of Daily Business Reports based on Sequential Text Mining Method*, Proc. of the 2004 IEEE Intl. Conf. on Systems, Man and Cybernetics, vol.4, pp.3279-3284, 2004.
[10] S. Sakurai, Y. Kitahara, and R. Orihara, *Sequential Mining Method based on a New Criterion*, Proc. of the Artificial Intelligence and Soft Computing 2006, pp.1-8, 2006.
[11] S. Sakurai, Y. Kitahara, R. Orihara, K. Iwata, N. Honda, and T. Hayashi, *Discovery of Sequential Patterns Coinciding with Analysts' Interests*, J. of Computers, vol.3, no.7, pp.1-8, 2008.
[12] S. Sakurai, *An Efficient Discovery Method of Patterns from Transactions with their Classes*, Proc. of the 2010 IEEE Intl. Conf. on Systems, Man and Cybernetics, pp.2116-2123, 2010.
[13] Y. -W. 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, January, 2004.
[14] 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 - Social and Behavioral Sciences, vol.26, pp.55-62, 2011.
[15] http://www.geocities.jp/sundaysoftware/csv/keiretu.html
[16] http://www11.ocn.ne.jp/˜ kui168/link37.html
[17] http://www11.ocn.ne.jp/˜ kui168/link39.html

Shigeaki Sakurai is with the IT Research and Development Center, Toshiba Solutions Corporation, Tokyo, 183-8512 Japan. He acquired the Ph.D. degree from Tokyo University of Science on 2001.

Kyoko Makino is with the IT Research and Development Center, Toshiba Solutions Corporation, Tokyo, 183-8512 Japan.

Shigeru Matsumoto is with the IT Research and Development Center, Toshiba Solutions Corporation, Tokyo, 183-8512 Japan.