2012 26th International Conference on Advanced Information Networking and Applications Workshops

A Discovery Method of Trend Rules from Complex Sequential Data

Shigeaki Sakurai
Advanced IT Laboratory
Toshiba Solutions Corporation
Tokyo, Japan
Sakurai.Shigeaki@toshiba-sol.co.jp

Kyoko Makino
Advanced IT Laboratory
Toshiba Solutions Corporation
Tokyo, Japan
Makino.Kyoko@toshiba-sol.co.jp

Shigeru Matsumoto
Advanced IT Laboratory
Toshiba Solutions Corporation
Tokyo, Japan
Matsumoto.Shigeru@toshiba-sol.co.jp

***Abstract*—This paper proposes a method that discovers trend rules from complex sequential data. The rules represent relationships among evaluation objects, keywords, and changes of numerical values related to the evaluation objects. The data is composed of numerical sequential data and text sequential data. The method extracts frequent patterns from transaction sets based on the changes. Also, it regards combinations of the patterns and the changes as trend rules. This paper applies the method to data sets composed of stock data and news headlines. Lastly, this paper compares the method with a method based on the random selection and shows the effect of the proposed method.**

***Keywords*-Trend rule, frequent pattern, numerical sequential data, text sequential data, and evaluation object.**

I. Introduction

As many kinds of sensors are smaller and cheaper, they are more easily buried in the real world environment. 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. Also, we 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, [8] 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. [9] 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. [11] proposes a method that discovers sequential patterns from sequential data with the tabular structure. In the case of the data, each item 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.

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.

Some existing methods dealing with complex sequential data have been proposed in the financial field. Existing methods [1][2][4] can analyze relationships between a specific numerical sequence and texts. The numerical sequence is the synthetic stock index such as Dow Jones Industrial Average (DJIA) or the change of specific currency exchange. The other existing methods [6][7][13] can analyze relationships between texts related to specific stock brands and their numerical sequences. It is difficult for these existing methods to simultaneously and respectively deal with many stock brands. Thus, this paper tries to consider an analysis method of the complex sequential data overcoming to this problem.

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, the former one is stock price sequences of evaluation objects and the latter one is sequences of news headlines describing them. Then, the evaluation objects are companies. This paper proposes a method that discovers trend rules from the data. The trend rules are used to extract attractive evaluation objects. This paper verifies the effect of the proposed method through experiments.

II. 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.

978-0-7695-4652-0/12 $26.00 © 2012 IEEE
DOI 10.1109/WAINA.2012.21

IEEE computer society

We focus on the change of the data. Then, we can observe 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 describes various meanings. We cannot 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. 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 explained 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 and explains it by separating the learning method of rules and prediction method based on rules.

B. Learning method of rules

This subsection explains the learning method of rules. The method acquires rules according to the outline as shown in the upper part of Figure 1. Firstly, the method applies a morphological analysis engine to text sequential data in order to separate texts in the data to words and analyze their parts of speech. This paper 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 texts. They are regarded as evaluation objects and attributes, respectively. A transaction composed of the evaluation objects and the attributes is generated from a text of the text sequential data. But, if the text does not include an evaluation object, a transaction cannot be generated from it. Next, the method extracts numerical sequential data related to the evaluation object. Two numerical values are extracted 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 Formula (1). In this formula, $f_{e,t}$ is a numerical value corresponding to both an evaluation object e and a target time t.

$$r_{e,t} = \frac{f_{e,t+1} - f_{e,t}}{f_{e,t}} \tag{1}$$

Next, the method identifies a class corresponding to a text including an evaluation object based on the change ratio. Transaction subsets with classes are generated by gathering transactions assigned the classes. Next, the method applies each transaction subset to the discovery method of frequent patterns. This paper uses the method proposed by [5]. Frequent patterns are discovered for each class. Lastly, the method regards the combination of a frequent pattern and a class 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 transactions related to classes. The disparity leads to the increase of discovery time. In our future works, we will try to apply the method into the data sets by easing the increase.

C. Prediction based on rules

This subsection explains the prediction method based on rules. The method predicts attractive evaluation objects according to the outline as shown in the lower of Figure 1. Firstly, the method applies the morphological analysis engine to real-timely distributed texts. Proper nouns and other nouns are extracted from each text. The extraction corresponds to the one in the learning method. Next, the prediction method compares a set of nouns extracted from a text with a set of nouns included in each rule. If the set of nouns included in a rule is included in the one extracted from a text, the text is regarded as a text matched to the rule. The evaluation object included in the text is identified. The transaction and the class of the rule are assigned to the evaluation object. The number of assigned transactions is accumulated for each evaluation object and each class. The extraction, the matching, and the accumulation are repeated until the next time stamp. Lastly, the method evaluates whether evaluation objects are attractive or not by referring to their numbers of assigned transactions. The attractive evaluation objects are shown to users. The users can judge whether countermeasures related to the attractive evaluation objects should be performed.

III. Experiments

This section explains experiments based on real complex sequential data. The data is composed of stock price sequences and news headlines. The evaluation objects are stock brands. It 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 headlines as the text sequential data. The news headlines 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). In this experiment, the interval D2 is divided into two sub-intervals:

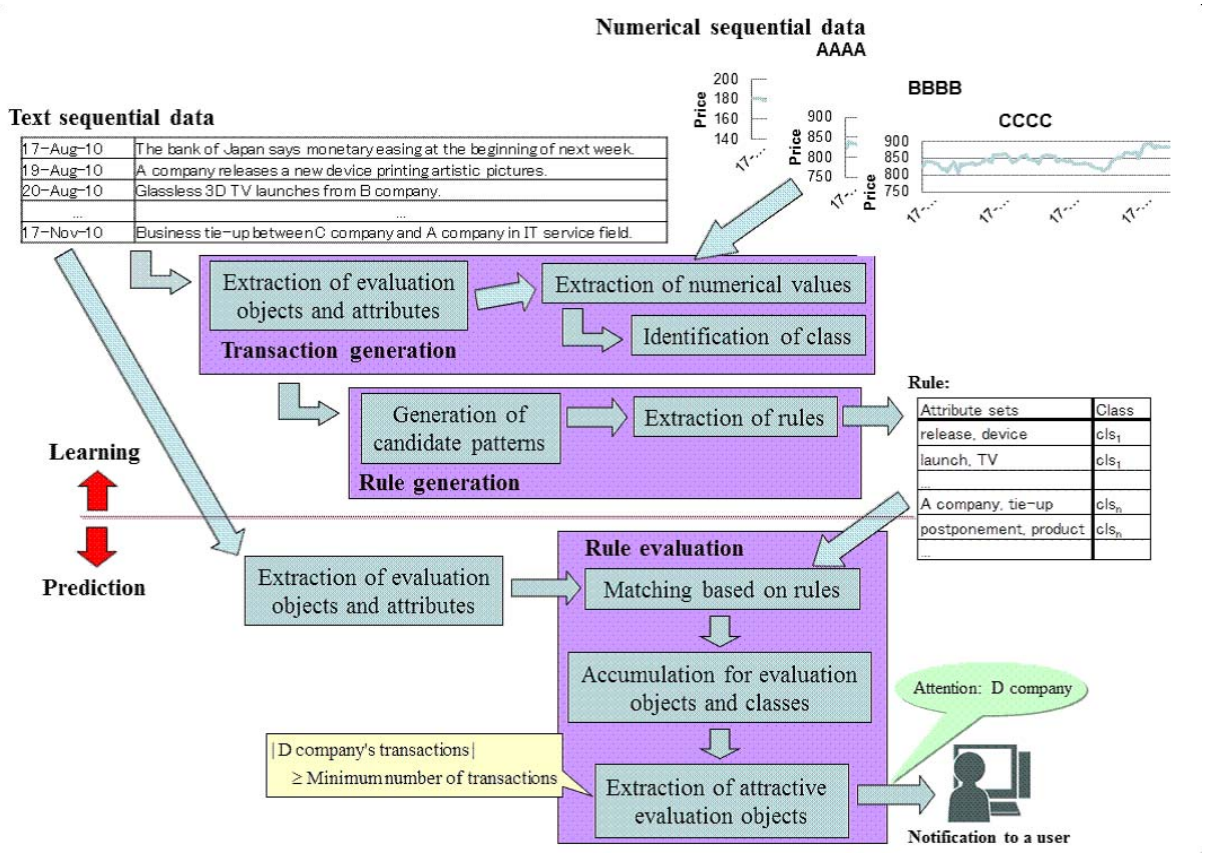


Fig. 1. An outline of the proposed method

February 1, 2011 ∼ March 10, 2011 (D2a) and March 11, 2011 ∼ 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 trend rules. Table I shows the numbers of collected news headlines. Each news headline is composed of related information such as date, time, site name, and genre.

	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

TABLE I
COLLECTED NEWS HEADLINE

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: http://www.geocities.jp/sundaysoftware/csv/keiretu.html. 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. They are stored with csv format. 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 the pages: http://www11.ocn.ne.jp/˜ kui168/link35.html and http://www11.ocn.ne.jp/˜ kui168/link37.html.

B. Evaluation criterion

It is anticipated that stock traders are interested in the changes of stock prices. This is because the traders can select 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 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 traders directly gaze at the recommended stock brands. If most parts of them are not valid, the traders may think that the recommendation is not valid. Therefore,

this paper focuses on the precision defined by Formula (2) 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. We evaluate whether they are attractive by referring to the stock price of the next day. If the attractive evaluation objects are recommended, recommended evaluation objects are regarded as truly recommended ones.

$$p = \frac{\text{Number of truly recommended evaluation objects}}{\text{Number of recommended evaluation objects}} \quad (2)$$

Next, we consider the processing time as other evaluation criterion. The criterion is selected due to the following three reasons. Firstly, the learning method requires dealing with more transactions in order to acquire more valid trend rules. It is anticipated that the number of training transactions increases more and more in near future. 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. This is because the traders may repeatedly operate the same stock brand in a day. The processing time is an important criterion evaluating the possibility of the real-time processing. This paper roughly measures the time. That is, the DOS command "time" is run before and behind running modules 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 processing time of respective processes. 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 rules are related to the earthquake. However, the rules are not always usual rules because the big earthquake is a rare event. The 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. Also, it uses the data set D2a for the evaluation of precision and the data set D3 for the evaluation of the processing time.

On the other hand, this experiment classifies the changes of stock prices into three classes: "Drop", "Steady", and "Rise". In this experiment, "Drop", "Steady", and "Rise" correspond to the ranges: (-∞, -5%], (-5%, +5%], (+5%, +∞). 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 they decide which operations should be selected. 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 are larger than or equal to a predefined threshold.

This paper compares the proposed method with the random method. The random method randomly selects stock brands from all stock brands by referring to a probability of the changes of stock prices. The probability is calculated by Formula (3). 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. We anticipate that the precision of the proposed method is larger than the one of the random method.

$$\frac{\text{the number of stock brands assigned "Rise" and "Drop"}}{\text{the number of stock brands}} \quad (3)$$

D. Experimental results

The left side of 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 most of 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. In our future works, we will try to consider the improvement of the extraction.

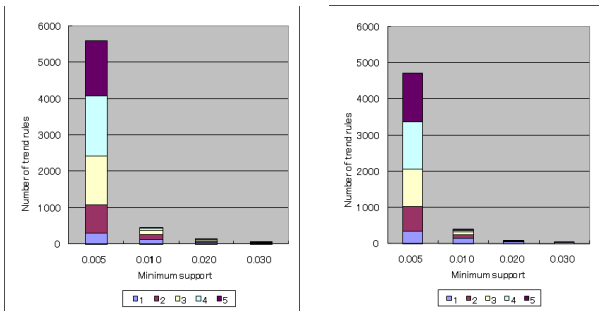
	Training				Evaluation	
	Rise	Steady	Drop	No Obj.	D3	D2a
Excite	58	7,096	75	123,390	41,926	37,440
Goo	128	7,576	101	132,740	20,905	29,520
Infoseek	208	17,312	151	219,865	64,677	60,129
Livedoor	166	13,918	168	214,773	70,935	64,209
Yahoo	298	20,464	389	228,163	78,765	67,835
Total	858	66,366	884	918,931	277,208	259,133

TABLE II
TRANSACTION SETS

The right side of Table II shows evaluation transaction sets for each site. The sets are generated from news headlines in the data sets D3 and D2a. Their classes directly identified by referring to the changes of stock prices, but are done by applying the sets into the prediction method. Only transactions identified "Rise" and "Drop" contribute to the recommendation of stock brands.

Figure 2 shows the number of trend rules extracted from the data set D1. Figure 2 (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 shows the accumulated number of trend rules for each number of items included in trend rules. These graphs show that the number

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 into the learning method.



(a) Rise (b) Drop

Fig. 2. Number of extracted rules

Figure 3 shows the number of evaluation objects whose classes are "Rise" or "Drop". Each evaluation object is evaluated for each day corresponding to the data set 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 training news headlines. This is because it is possible for a high change ratio of stock prices in the case of the training ones to acquire trend rules more validly representing "Rise" and "Drop".

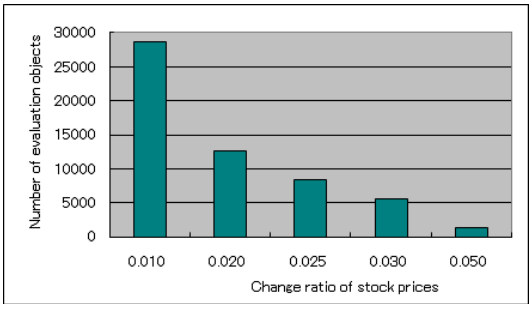


Fig. 3. Number of evaluation objects with high change ratios of stock prices

Figure 4 shows precisions of the proposed method and the random method in the case of data set D2a. In this figure, a horizontal axis shows the change ratio of stock prices and a vertical axis shows the precision. In this experiment, three parameters of the proposed method are adjusted. They are the minimum support, the minimum length of trend rules, and the minimum number of transactions. The prediction method applies only trend rules whose lengths are larger than or equal to the minimum length into evaluation transactions. Also, it extracts evaluation objects whose total numbers of assigned evaluation transactions are larger than or equal to the minimum number as attractive evaluation objects. SxxIyTz represents a parameter set. Sxx means that the minimum support is 0.0xx, Iy does that the minimum length is y, and Tz does that the minimum number is z.

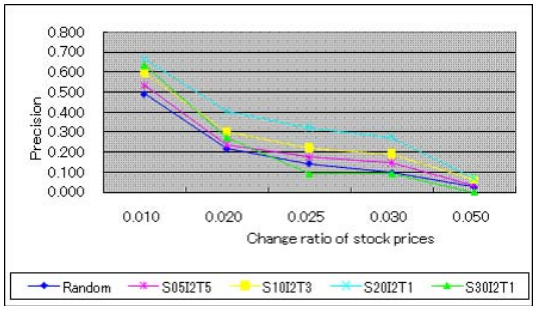


Fig. 4. Proposed method v.s. random selection

Figure 5 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.

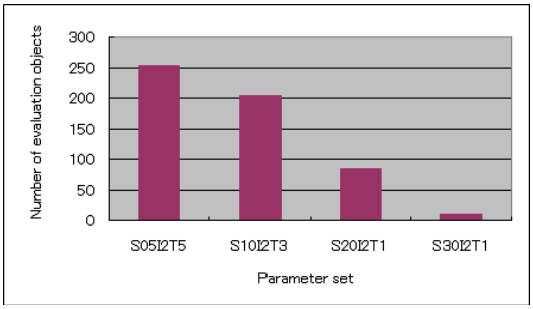


Fig. 5. Number of extracted evaluation objects

Table III shows the processing time of the learning method and the prediction method. Here, the prediction method deals with the data set D3. This is because the interval of the data set is longer than the one of the data set D2a. In the learning method, the transaction generation process and the rule generation process are separately run. Also, in the evaluation method, the extraction process of evaluation objects and attributes, and the rule evaluation process are separately run. The separation is based on the constraint of our experimental environment.

E. Discussions

This section discusses experimental results with two viewpoints: the validity of prediction and the process time.

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 rule gives an excessive impact to the precision. We think that the excessive small number of the rules leads to their deterioration. Also, Figure

	Process	Time(second)
Learning	Transaction generation	1,607
	Rule generation	148
Prediction	Extraction of evaluation objects and attributes	1,458
	Rule evaluation	357

TABLE III
PROCESS TIME

4 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 their precisions are distributed in [0.15, 0.32]. We cannot insist that the precisions are sufficiently high. It is necessary to improve the precisions in near future.

We note two related works [2] and [7]. [2] shows that 87.6% parts of the changes of stock prices in DJIA are explained by the feeling "Calm" extracted from Twitter messages. It deals with the synthetic changes of stock prices and is easier than our target task. This is because our target task deals with many respective changes of stock prices. However, it is important that the analysis of 10 million messages brings in high precision. We can anticipate that the increase of training transactions leads to higher precision. Also, [7] shows that rules related to three classes "Rise", "Steady", and "Drop" are discovered and they can predict the classes with over 50% probability, even if [7] discovers only rules related to a few brands of currency exchange. On the other hand, 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 process of evaluation objects and attributes. We think that the process can be performed in parallel by applying it into each news headline. The parallel process can realize shorter process time. In our other study, 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 strategies. That is, the parallel process can be performed based on the decomposition of both trend rules and news headlines. The prediction of evaluation targets can be real-timely processed.

Next, we note the rule generation process. It is not easy to real-timely perform the process, even if 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 discover frequent patterns in parallel. The parallel process requires comparatively complicated calculation process. On the other hand, it is not always necessary to real-timely perform the learning process. This is because the trend does not real-timely change. We think that the parallel process of the learning process is not important in near future.

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

IV. 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 into real data sets collected from Web sites. In the data sets, evaluation targets, text sequential data, and numerical sequential data are stock brands, news headlines, and stock prices, 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.

In our future works, we will try to improve the precision. Also, we will try to improve the update of new attributes and to develop a method which flexibly designs classes. In addition, we will try to apply the proposed method into other application fields. For example, we consider the analysis of complex sequential data collected from smart communities and machine maintenance field. 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] 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.

[5] 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.

[6] 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.

[7] D. Peramunetilleke and R. K. Wong, *Currency Exchange Rate Forecasting from News Headlines*, Proc. of the 13th Australasian Database Conference, vol.5, pp.131-139, 2002.

[8] 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.

[9] 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.

[10] S. Sakurai and R. Orihara, *Discovery of Important Threads from Bulletin Board Sites*, Intl. J. of Information Technology and Intelligent Computing, vol.1, no.1, pp.217-228, 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.