A-011

Multi-solution Trend Analysis of Stock Price Change by Textual Information
ストリームデータ発掘手法に基づく記事イベントが株価変動における影響推測

范 薇† 渡邉 豊英† 朝倉 宏一‡
Wei Fan Toyohide Watanabe Koichi Asakura

1. Introduction

People read pieces of news to understand what is happening or what might happen in the future. Especially, for private investors, they pay more attentions on news articles which suggest why current economic performance is poor or predict an upturn in the economy in the coming months. News releases influence human behavior, and may indirectly affect fluctuations in financial market.

In our research work, we collect both of the online stock price data and time-stamped news articles, and aim to find the relationship between them. Stock price data are described into high-level features which we call trends. Then we align rise/drop trend of price data with news articles, and learn classifiers of these articles which are correlated with a given trend. A classifier determines the statistics of words usage patterns among the articles in the training set. According to these classifiers, we can predict rise/drop trend of stock price when given a newly released news article. The prediction results can be used by investors or an automated trading system as recommendations to buy or sell a particular stock. Additionally, our methods are special for realizing multi-solution trend analysis of stock price data satisfying both of short-term and long-term investments.

For example, as shown in Figure 1, we can find that the price data of *Tsuzuki Denki Co., Ltd.* has risen sharply after April 24, 2006. Refer to the news article about *Tsuzuki Denki Co., Ltd.* released in *Nikkei Shimbun* at the April 24, 2006 in Figure 2, we can see the words emphasized in the article were the influential factors to the rise of stock price data. Therefore, in this paper, we achieve to learn the patterns of words which are highly associated with rise or drop trend of price data, and identify news articles that are highly indicative of future trends. The ability of our methods for predicting forthcoming trends of stock price is evaluated by a market simulation. Results of our experiments demonstrate that our methods are capable of producing profits

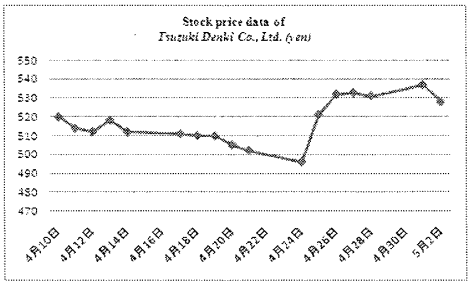


Figure 1 Stock price data.

†名古屋大学大学院情報科学研究科 Graduate School of Information Science, Nagoya University
‡大同大学 Daido University

2006年4月24日 日本経済新聞 (朝刊)

都築電、純利益７１％増――前期、物流拠点売却で税効果

都築電気は二十四日、二〇〇六年三月期の**連結純利益が前の期比七一％増**の二十一億円になったと発表した。従来予想は九億円だった。物流センターの売却に伴う税効果会計の適用で法人税等調整額を計上した。 **売上高は六％増**の千百六十八億円と従来予想を二十八億円上回った。情報通信システムは競争激化による**価格下落**で予想を下回ったが、**携帯電話用チップセットの売上高が伸びた**。

Figure 2 News article.

that are significant higher than fixed period approach.

The rest of this paper is organized as follows. Some of related works are studied in Section 2, and then in Section 3, we introduce the design of our framework and deliberate our methods. Section 3.1 describes our processes of stock price data: segmentation and labeling of fluctuation trends. Section 3.2 outlines the weighting rules for feature vectors of news articles. Then Section 3.3 explains how to construct classifiers of news articles associating with each type of trends. Section 3.4 extends a flexible multi-solution trend analysis gaining an advantage over other traditional algorithms. Finally, we evaluate our methods in Section 4.

2. Related Work

Recently, more and more research works on stock market prediction take the influence of textual information into account. The *AEnalyst* system was proposed by Lavernko et.al in [1] for predicting the change of stock price data based on analysis of online news stories. Mittermayer et.al [2] evaluated news articles based on analysis of influential factors to short-term fluctuation trends of stock price data. On the other hand, in [3], Izumi et.al contributed to factorial analysis of long-term financial markets' fluctuation based on monthly reports of Bank of Japan.

In fact, investors have flexible demands for short-term or long-term trend analysis of each stock, especially taking the data evolution of stock data into consideration. Therefore all of the above methods are failed to diagnosis the change of stock data flexibly. In this paper, we propose to detect changes of trends by analyzing data distribution of each stock price data, and store multi-solution estimates of data distribution in a hierarchical structure; consequently, we realize online and automatic monitoring of stock data without ignoring important change points of stock data lazily or wasting calculation resources for some stable stock data. Additionally, referred to the related work of [4], in order to avoid failure for learning insufficient news articles of some stocks, it is accurate to use expressions extracted by syntax analysis for characterizing news articles. Therefore, we focus on syntax analysis of sentences, especially dependency relation among words to characterize each news article.

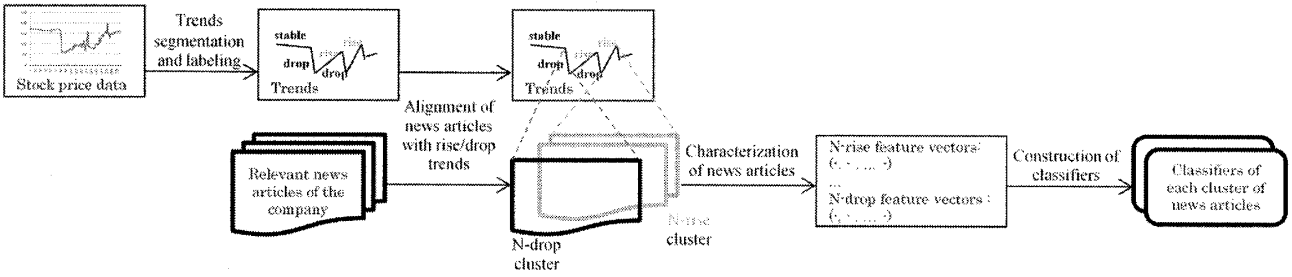


Figure 3 Framework design

3. Framework Design

Figure 3 illustrates the structure of our framework. We use both of the numerical historical stock price data and news articles about the companies. We have collection of 209,553 news articles for 108 stocks collected from *Nikkei Shimbun* from Journal 1, 2003 to December 31, 2006, as well as the stock prices for the 108 stocks over the same period of time. Using stock price data for a given company, we segment and label fluctuation trends. News articles are aligned with the labeled trends according to when the articles were released and when each trend occurred. We learn classifiers of the aligned articles for each type of trend based on feature vectors extracted from articles, and then use the classifiers to predict the newly released news article's influence on the future of stock price data. In the following subsections, we elaborate each process in our framework.

3.1 Processing of Stock Price Data

Taking data evolution of stock price into consideration, we can see that fixed periodical analysis of financial market as proposed in [3] is not appropriate. It is likely to cause lose of significant changes of stock data within the fixed period or waste of calculation resources for some stable stocks. On the contrary, some approximation algorithms of time-series with linear segments were also proposed. However, in these algorithms, setting of threshold for approximation error is very difficult for evolving stock price data. Therefore, an online and automatic method for monitoring the changes of stock price data is required. In this paper, we focus on estimating the data distribution of stock data in favor of change diagnosis of the time-series data. Change points of data distribution indicate the segments of stock data as well. Then we discuss how to diagnosis changes of data distribution and observe fluctuation trends of stock price data.

3.1.1 Segmentation

Given a time-series of stock price data described in Figure 4. Value a_t is the price data at current time instant t. In terms of user specified time horizon h_t, we propose methods to analysis the change of data distribution in the price data samples S which arrived in the time window (t-h_t, t). We calculate the rate of variation in data distribution $V_{(h_s,h_t)}(X,t)$at a spatial location X over time horizon h_t as defined in equation (1).

$$V_{(h_s,h_t)}(X,t) = (F_{(h_s,h_t)}(X,t) - R_{(h_s,h_t)}(X,t-h_t))/h_t \quad (1)$$

$$F_{(h_s,h_t)}(X,t) = \sum(t/h_t)K'_{h_s}(A) \quad (2)$$

$$R_{(h_s,h_t)}(X,t) = \sum(1-t/h_t)K'_{h_s}(A) \quad (3)$$

$$K'_{h_s}(A) = 1/(2\pi h_s^2)^{\frac{1}{2}} \cdot \exp\left\{-||a_t - X||^2/(2h_s^2)\right\} \quad (4)$$

In equation (1), spatial location X is a possible value within the range of price data, and $F_{(h_s,h_t)}(X,t)$ defined in equation (2) measures the density function of data samples S based on a kernel possibility estimation function which is defined as equation (4), where h_s is the smooth parameter. Similarly, $R_{(h_s,h_t)}(X,t)$ defined in equation (3) measures the density function based on the same data set reversely.

Consequently, it is noted that if a greater number of data samples which are closer to X arrive at the end of the interval (t-h_t, t), then the rate of variance in data distribution is positive. While, when a greater number of data samples which are closer to X arrive at the beginning of the interval, then the rate of variance in data distribution is negative. If the data distribution has largely remained unchanged, then the rate of variance at the location X will be almost zero.

According to the resultant rate of variance in data distribution at all of the possible spatial locations within the range of stock price, we can diagnose coagulation region where the rates of variance in data distribution are larger than user specified threshold in the fluctuation of price data. Similarly, we can define dissolution regions as well. Then in order to identify significant changes of data distribution within the time window (t-h_t, t), we compare the coagulation region [$cog_{start(t)}$, $cog_{end(t)}$] at each time instant t with dissolution region [$dis_{start(t-1)}$, $dis_{end(t-1)}$] at the previous time instant (t-1). If overlap exists between these two regions, it means that there is a change of data distribution at time instant t, and we can insert a segmentation point at t; otherwise, we can say that the data distribution remains the same.

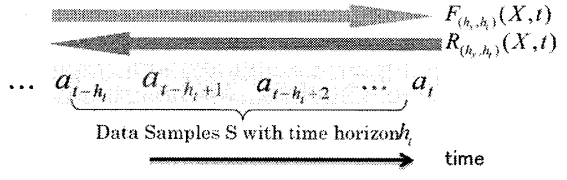
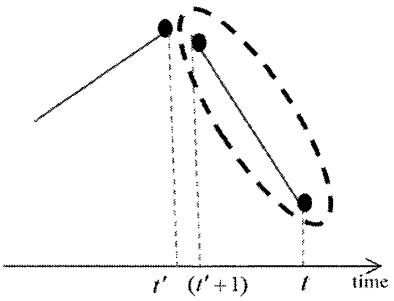


Figure 4 time series of stock price data.

3.1.2 Fluctuation trends labeling

According to the previous process, we get the segmentation points (t'+1) and t as shown in Figure 5, then a linear segment can be identified by the two data points: [(t'+1), mean($cog_{start(t'+1)}$, $cog_{end(t'+1)}$)] and [t, mean($cog_{start(t)}$, $cog_{end(t)}$)]. Here, mean(•) is the mean function of its two parameters.



We propose a hierarchical algorithm to cluster these segments

Figure 5 Segments of stock price data

into different interesting trends based on: slope m and coefficient of determination R^2. Here, each segment is represented by (m, R^2). These segments are merged according to minimum group average distance defined as

$$GAD(C_i, C_j) = \sum_{k \epsilon C_i} \sum_{l \epsilon C_j} d_{ij}(k, l) / |C_i||C_j|, \quad (5)$$

where $|C_i|$ and $|C_j|$ are the magnitudes of the cluster C_i and C_j respectively; $d_{ij}(k, l) = \sqrt{(m_k - m_l)^2 + (R_k^2 - R_l^2)^2}$ is Euclidean distance between the segments k and l which are inside of C_i and C_j, respectively. The clustering procedure terminates when the number of clusters are equal to three: rise, drop and steady. Those segments in the cluster having the maximum average slope are labeled as rise. Similarly, those segments in the cluster having the minimum average slope are labeled as drop. Segments in the remained cluster are labeled as steady.

3.2 Feature Vectors of News Articles

After trends are grouped into clusters, relevant news articles are then aligned to them. By alignment, we mean that the contents of the news articles would support and account for the happening of the trends. Obviously, not every piece of news announced under the time series would support the happening of the trends. In this section, we propose a new algorithm to filter out news articles that do not support the trends.

In the following discussion, let T-clusters be the clusters of trends and N-cluster be the clusters of news articles. First, all of the news articles that are released within T-cluster rise (drop) are grouped together. Each article is represented by a normalized vector space model $d_i=(w_1, w_2, \ldots w_n)$, where element w_t corresponds to the score of keyword t in the article d_i, and it is calculated by the standard tf-idf scheme: $w_t = tf_{d,t} \times \log \frac{N}{df_t}$, where $tf_{d,t}$ is the frequency of t in the article d_i; df_t is the number of articles containing the term t; N is the total number of articles containing in the particular T-cluster rise/drop. In our process, we use words in expressions extracted by syntax analysis of sentences, especially dependency relation among words to characterize each news article.

K-Means algorithm is then used for splitting the weighted article into two clusters. The centroid of the cluster C_i is defined as $C_i = \frac{1}{|S_i|} \sum_{d \epsilon S_i}$, where S_i is the set of articles within the cluster C_i and $|S_i|$ is the number of articles in this set. The similarity between the article d_i and centroid of C_j is determined by cosine measure:

$$\cos(d_i, C_j) = \frac{d_i \cdot C_j}{|d_i||C_j|},$$

where $|d_i|$ and $|C_j|$ is the magnitude of the article d_i and the cluster C_j respectively.

In order to differentiate the features appearing in one of the cluster but not the other, two coefficients are introduced: inter-cluster discrimination coefficient and intra-cluster similarity coefficient defined in equation (6) and (7) respectively:

$$CDC = \frac{n_{i,t}^2}{N_t}, \quad (6)$$

$$CSC = \sqrt{\frac{n_{i,t}}{n_i}}, \quad (7)$$

where $n_{i,t}$ is the number of articles in the N-cluster I containing keyword t; N_t is the number of articles containing t and n_i is the number of different keywords. Therefore, the weight of each feature in each document is finally calculated as follows:

$$w_t = tf_{t,d} \times CDC \times CSC$$

Finally, each news articles is represented by a vector-space model in which it is normalized to unit length, so as to account for documents with different lengths.

3.3 News Articles' Classifiers

The association between features of news articles and trends of price time-series data is generated based on Support Vector Machine (SVM) [5]. SVM is a new learning algorithm proposed by Vapnik to solve the two-class pattern recognition problem using the structural risk minimization principle. It obtains very accurate result in text classification, and outperforms many other techniques such as neural network and Naïve Bayes. We have a pair of classifiers. One is responsible for classifying whether a news article will trigger a rise event; the other is responsible for the event of drop.

3.4 Extension to Multi-solution Trend Detection

As discussed in previous section, in order to analysis trends of stock price data flexibly, we propose a hierarchical structure to store multi-solution data distribution estimates. As demonstrated in Figure 6, we store the estimates of data distribution as noted as black points at each level. In terms of user's request at arbitrary time instant with arbitrary time horizon, it is possible to calculate the demanded data distribution from the structure. For example,

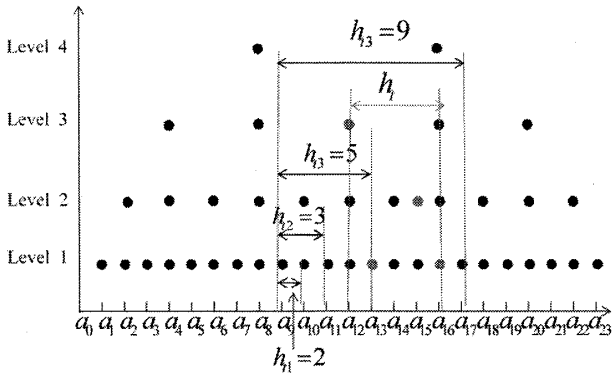


Figure 6 Multi-solution analysis of trend

at the time instant a_{16}, when a user requests to do trend analysis with time horizon h_t, we can estimate requested data distribution from the red points in Figure 6.

4. Evaluation

We use the collection of 209,553 news articles for 108 stocks collected from *Nikkei Shimbun* from Journal 1, 2003 to December 31, 2006, as well as the stock prices for the 108 stocks over the same period of time in our experiments. Features of news articles are extracted using *chasen* [6] and *cabocha* [7]. The training of the classifiers, as well as the prediction task, is carried out using the package of SVM^{light}.

4.1 Trends Discovery and Labeling

A typical result after the time series segmentation is shown in Figure 7. It is easy to see that shape of the stock data after segmentation is preserved, while the number of data samples would be reduced up to one-tenth of the original one.

4.2 Market Simulation

For the task of market prediction, the ultimate evaluation of performance is whether or not the framework would be able to make a profit. In the following simulation we construct a strategy which mimics the behavior of a day trader who would use the resultant prediction in a very simple fashion: if the resultant prediction indicates that a news article is likely to precede a rise trend for the stock, he invests in that stock; if the resultant prediction trend is drop, he sells the stock. To simulate this very simple strategy, we induced a separate classifier for each stock for each trend type using 3 months worth of training data between October and December 2005. Then, for 40 days starting on January 2nd, we have our framework monitor the news. Every time a news article appears for some company, our framework determines which trend model is most likely to have generated it.

We compare our framework with the fixed period alignment approach. Figure 8 shows a comparison between our methods and the fixed period approach with the frequency of news articles announced versus resulting profit. All of the stocks are ranked based on the total number of news articles that are associated with in the training period. In the figure, the smaller the number of the x-axis is, the fewer number of news articles are aligned to that stocks. In general, our approach is highly superior to the fixed period approach. The reason is that we use all news articles,

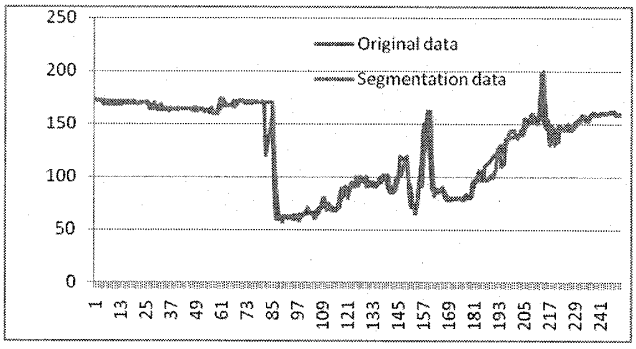


Figure 7 A stock price data before and after segmentation

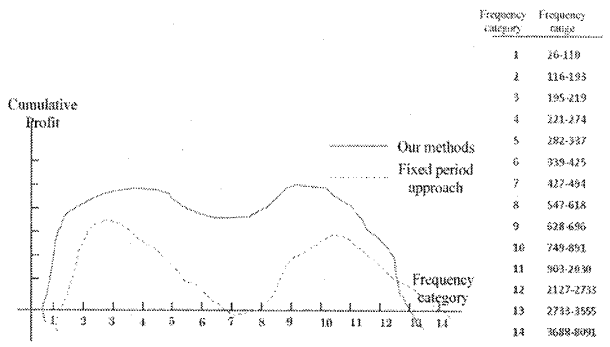


Figure 8 The relationship between the frequency of news articles and the resulting profit

but fix period approach only uses news articles within a fixed interval preceding the happening of a trend. However, when a stock received too many news articles, fix period approach outperforms us due to the fact that the probability of having noise would be higher.

5. Conclusion

In this paper, we demonstrated a sophisticated framework which monitors the stock market and predicts its future behaviors. The major difference between our framework and the existing forecasting techniques is that our approach does not need any assumptions that require a fixed period for aligning news articles to a trend. Data distribution estimation based algorithm and a hierarchical structure are used for discovering and labeling multi-solution trends respectively. A new algorithm is used for news articles filtering and alignment. The new weighting scheme is formulated to identify the important features within the collection of news articles. Finally, a market simulation using a very simple trading strategy based on the Buy-and-Sell test is carried out, and the results indicated that our approaches are profitable.

REFERENCE

[1] Lavernko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J.: "Mining of Concurrent Text and Time Series", Proceeding of KDD Conference Text Mining Workshop, pp. 37-44 (2001).

[2] Mittermayer, M. A.: "Forecasting Intraday Stock Price Trends with Text Mining Techniques", Proceeding of 37th Hawaii International Conference on System Sciences, pp. 64-73 (2004).

[3] 和泉 潔, 後藤 卓, 松井 藤五郎, "テキスト情報による金融市場変動の要因分析", 第 23 回人工知能学会全国大会, (2009).

[4] 張 へい, 松原 茂樹,:"株価データに基づく新聞記事の評価", 第 22 回人工知能学会全国大会, (2008).

[5] Joachims, T.: "Making large-scale SVM Learning Practical", Advanced in Kernel Methods-Support Vector Machine, (1999).

[6] Chasen, http://chasen.naist.jp/

[7] Cabocha, http://chasen.org/taku/software/cabocha/