Dynamic Programming Approach to Retrieving Similar Candlestick Charts for Short-Term Stock Price Prediction

Yoshihisa Udagawa
Computer Science Department, Faculty of Engineering,
Tokyo Polytechnic University
1583 Iiyama, Atsugi, Kanagawa, Japan
e-mail: udagawa@cs.t-kougei.ac.jp

Abstract— **The paper describes a method for a short-term stock price prediction based on candlestick chart techniques that are popular among stock traders using technical analysis. While the techniques have long history, there is still no consistent conclusion on the predictability of the techniques. We focus on the fact that a trend of stock prices often continues after intervals of several days because stock prices tend to fluctuate according to announcements of important economic indicators, economic and political news, etc. Typically, stock price movements in the period without important news are small, resulting in generating a series of noisy candlesticks. To cope with the noisy candlesticks, this paper focuses on a dynamic programming algorithm that allows us to perform partial matches on sequences of stock prices. We propose a model consisting of six parameters for retrieving similar candlestick charts in order to take into account where the stock price occurs in high/low price zones, in addition to a price change and a length of candlestick body. Experiments are performed on the daily NASDAQ composite index. We choose the daily time frame since important news that affects stock prices occurs on a daily basis. The statistics of the candlesticks are calculated to determine the parameter values of the proposed model based on the average and the standard deviation. Experimental results show that the proposed method is effective in predicting both uptrend and downtrend. Strictly, the prediction of the downtrend is a little bit difficult than that of the uptrend, probably reflecting the fact that the NASDAQ stock market is constantly growing.**

Keywords-Stock price prediction; Technical analysis; Candlestick charts; Longest common subsequence; Statistics of candlesticks.

I. INTRODUCTION

This research paper is based on the previously reported contribution on a candlestick chart retrieval algorithm for predicting stock price trend [1]. We provide details on implementation of the dynamic programming algorithm to eliminate noisy candlesticks that often occur when there is no noteworthy economic news. While experiments are performed on the daily Nikkei stock average (Nikkei 225) in the previous paper, this paper presents experimental results on the NASDAQ composite index [2] for showing the applicability of the algorithm.

Predicting the direction of future stock prices is a challenging topic in many fields including trading, finance, statistics and data mining in computer science. The motivation is to predict the direction of future prices so that stocks can be bought and sold at profitable positions. Fundamental analysis and technical analysis are two primary approaches to making investment decisions for successful stock trading [3].

Fundamental analysis is an approach involving a study of company fundamentals such as revenues and expenses, business concept and competition, and so on. To forecast future stock prices, fundamental analysis combines economic, industry, and company analysis to evaluate a stock's current fair value and forecast future value. Because of this analyzing processes, most people believe that fundamental analysis is mainly suitable for long-term prediction.

Meanwhile, technical analysis is a method of predicting the future direction of a stock price by studying historical stock price patterns. A technical analysis presumes that those price patterns tend to repeat themselves in the future. One of the important types of technical analysis is candlestick chart patterns [4]. The candlestick chart patterns provide short-term predictions for traders to make buy or sell decisions. While most techniques use patterns of stock prices during more than ten days, the candlestick charting technique focuses on patterns among several days of candlesticks formulated by opening, high, low, and closing prices within a specific time frame, such as minute, hour, day or week. Dozens of candlestick chart patterns are identified to be signals of up, down, and sideways of trend directions. These patterns consist of a single candlestick or a combination of multiple candlesticks usually less than four. In fact, the technique acts as a leading indicator with its capability to provide trading signals earlier than other technical indicators. It is also used by some real time technical service providers to provide quick signals for market's sentiments [5].

The candlestick charting technique probably began sometime after 1850 [4]. Despite of its long history and popularity, mixed results are obtained in the studies on candlestick charting. Negative conclusions to the predictability of candlesticks are reported [6]-[8], while positive evidences are provided for several candlestick chart patterns in experiments using the U.S., European and Asian stock markets [9]-[15].

It is also pointed out that candlestick chart pattern recognition is subjective [4]. The candlestick chart patterns are often qualitatively described using words and illustrations. The studies [6]-[15] adopt definitions using a series of inequalities with different parameters that specify candlestick patterns. Numerical definitions of candlestick patterns are still controversial issues.

In addition, the previous study [1] mentions that the candlestick patterns do not occur in time series in a strict sense because stock price fluctuation continues after intervals of several days depending on announcements of important economic indicators, economic and political news, etc. Because of these characteristics, the candlestick chart patterns are deemed to bring controversial results on predictability regarding future market trends even sort-term prediction.

The aim of the study is to estimate the predictability of candlestick patterns for future stock price trends. The proposed algorithm is applied to the daily NASDAQ composite index, while the daily Nikkei 225 (Nikkei stock average) is used in the previous study [1]. Daily historical stock prices are used because important news that affects stock prices occurs on a daily basis, and we plan to relate chart patterns to economic and political news in the future study.

The contributions of this paper are as follows:

(I) We propose a novel model for retrieving candlestick patterns that includes six attributes. The model takes account of 5-day moving average and 25-day moving average to decide whether the patterns occur in high or low price zones of a stock, which is original to the best of our knowledge,

(II) The values of the attributes of the proposed model are estimated based on statistical analysis of candlesticks so that the experimental results are evaluated in terms of statistics hopefully being applicable to world markets,

(III) The LCS (Longest Common Substring) algorithm [16] is improved to make an optimal matching of candlestick patterns that contain noisy candlesticks,

(IV) The proposed model devises a graphical representation to make evaluation of the retrieval results easy to depict the predictability for short-term trends.

The remainder of the paper is organized as follows. Section II gives some of the most related work. Section III describes backgrounds of the candlestick chart. Section IV proposes a model for retrieving similar candlestick charts. An augmented dynamic programing technique is adopted to implement the proposed model. Section V presents experimental results on both uptrend and downtrend of stock prices. Section VI concludes the paper with our plans for future work.

II. RELATED WORK

The principles of technical analysis were derived from the observation of financial markets over hundreds of years. In Europe, the dawn of technical analysis appeared in Joseph de la Vega's accounts of the Dutch markets in the 17th century. In Asia, the candlestick charting technique emerged during early 18th century, and probably established sometime after 1850 [4]. In the U.S., the Dow Theory traced back to 1884. However, the technical analysis is widely dismissed by academics in the 1970s. In particular, it is rejected by the weak form of the EMH (Efficient Market Hypothesis) formulated by Fama [17]. The EMH states that stock prices adjust rapidly to the arrival of new information, there is no way to outperform the market average. The studies of the last two decades show controversial results for supporting the EMH and opposing it.

Some studies [6]-[8] find that the candlestick charting is useless based on the experiments using the stock exchange markets' data in the U.S., Japan and Thailand. Horton [6] examines candlestick patterns for 349 stocks finding little value in the use of them. Marshall, Young, and Cahan [7] investigate the profitability of candlestick trading strategies using nine scenarios with different buy-and-hold strategies. They conclude that trading signals generated by the candlestick technical analysis do not have profitable forecasting power on the DJIA (Dow Jones Industrial Average) and the Japanese market, which is consistent with the EMH. Tharavanij, Siraprapasiri, and Rajchamaha [8] investigate the profitability of uptrends and downtrends of candlestick patterns consisted of one-day, two-day, and three-day candle sticks. Based on experiments using stock data in the SET (Stock Exchange of Thailand), they conclude that any candlestick patterns cannot reliably predict market directions even with filtering method using well-known stochastic oscillators [4].

Other studies conclude that applying certain candlestick patterns is profitable at least for short-term trading [9]-[13]. Caginalp and Laurent [9] favorably evaluate the predictive power of eight three-day reversal candlestick patterns using S&P500 stock price data from 1992 to 1996. They propose to define candlestick patterns and price trend as a set of inequalities using opening, high, low, and closing prices. These inequalities are taken over in the later studies. Goo, Chen, and Chang [10] define 26 candlestick patterns using modified version of inequalities that are proposed by Caginalp and Laurent. They examine these patterns on stock data of Taiwan markets, and conclude that some of the candlestick trading strategies are valuable for investors.

Chootong and Sornil [11] propose a trading strategy combining price movement patterns, candlestick chart patterns, and trading indicators. A neural network is employed to determine buy and sell signals. Experimental results using stock data of the SET market show that the proposed strategy generally outperforms the use of traditional trading methods based on indicators.

Lu, Chen, and Hsu [12] apply candlestick trading strategies to the U.S. market data with several trend definitions. They claim that the trading strategies appear to possess predictive power on a price trend. Specifically, they indicate that three-day reversal patterns are profitable when the transaction cost is set at 0.5%. They also find that holding strategies play an important role to improve the effectiveness of candlestick charting.

Zhu, Atri, and Yegen [13] examine the effectiveness of five different candlestick reversal patterns in predicting short-term stock movements using stock data of two Chinese markets. The results of statistical analysis suggest that the candlestick patterns work out in predicting price trend reversals.

The following two studies conclude cautious results. Martinsson and Liljeqvist [14] give a set of inequalities to define candlestick patterns including the length of body, the change of price and the length of shadows. They evaluate the impact of the candlesticks trading strategies and the RSI (Relative Strength Index) to evaluate the trend of the market. While they have positive results on the Swedish OMXS30 exchange, they have negative results on the London FTSE100 exchange. Chin, Jais, Balia, and Tinggi [15] examine the predictive power of candlestick continuation patterns, which predict to continue in the direction of original trend, in a Malaysian stock market from 2000 to 2014. They conclude that only one downtrend continuation pattern shows significant predictive power during the 5-day holding period.

The studies [6]-[15] translate the candlestick verbal and visual descriptions into numeric formulas in order to be used in an algorithm. However, they fail to consider zones where the candlestick patterns of interest occur. The interpretation of candlestick patterns depends on the price zone, e.g., high, low, neutral. For example, the morning star pattern generally suggests an uptrend when it occurs in a low price zone. However, the morning star pattern is deemed to be less predictable when it occurs in a high price zone than it occurs in a low price zone.

Most importantly, the studies [6]-[15] do not discuss neutral or noisy candlesticks that often take place in charts because stock prices are apt to depend on important economic and political news and events.

III. Candlestick Chart and Patterns

This section introduces the formation of a candlestick. Candlestick patterns are a combination of one or more candlesticks [4]. Samples of well-known candlestick chart patterns are depicted. Because the candlestick patterns are described in natural language and illustrations, there are criticisms on their use for trend prediction by a computer.

A. Formation of Candlestick

A daily candlestick line is formed with the market's opening, high, low, and closing prices of a specific trading day. Figure 1 represents the image of a typical candlestick. The candlestick has a wide part, which is called the "real body" representing the range between the opening and closing prices of that day's trading.

If the closing price is above the opening price, then a white candlestick with black border is drawn to represent an uptrend candlestick. If the opening price is above the closing price, then a filled candlestick is drawn. Normally, black color is used for filling the candle to represent a downtrend candlestick.

The thin lines above and below the body represent the high/low ranges. These lines and are called "shadows" and also referred to as "wicks" and "tails." The high is marked by the top of the upper shadow and the low by the bottom of the lower shadow.

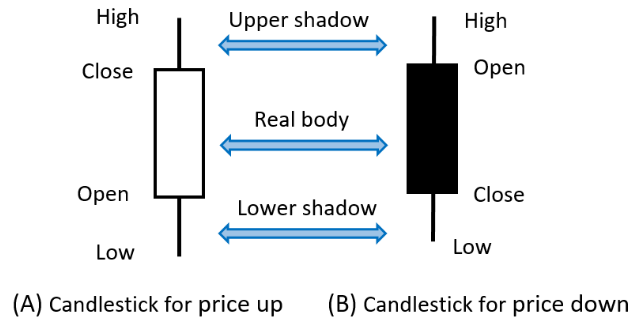

Figure 1. Candlestick formation.

Typically, a stock's opening price is not identical to its prior day closing price. This is because the time during which stock market is closed changes investor's emotions and expectations for the stock markets with different interpretations of economic news of the day and stock fluctuations.

B. Samples of Candlestick Patterns

Dozens of candlestick patterns are identified and become popular among stock traders [4]. These patterns have colorful names like *morning star*, *evening star*, *three white soldiers*, and *three black crows*.

Figure 2 shows the *morning star* pattern which is considered as a major reversal signal when it appears in a low price zone or at the bottom. It consists of three candles, i.e., one short-bodied candle (black or white) between a preceding long black candle and a succeeding long white one. The pattern shows that the selling pressure that was there the day before is now subsiding. The third white candle overlaps with the body of the black candle showing a start of an uptrend reversal. The larger the white and black candles, and the higher the white candle moves, the larger the potential reversal. The opposite version of the *morning star* pattern is known as the *evening star* pattern which is a reversal signal when it appears in a high price zone or at the end of an uptrend.

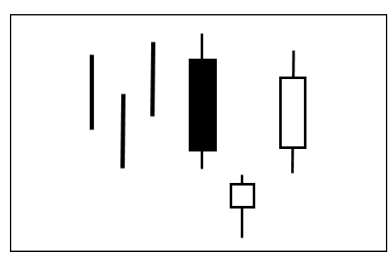

Figure 2. Morning star pattern.

Figure 3 shows the *three white soldiers* pattern which is interpreted as a strong indication of an uptrend market reversal when it appears in a low price zone. It consists of three long white candles that close progressively higher on each subsequent trading day. Each candle opens higher than the previous opening price and closes near the high price of the day, showing a steady advance of buying pressure.

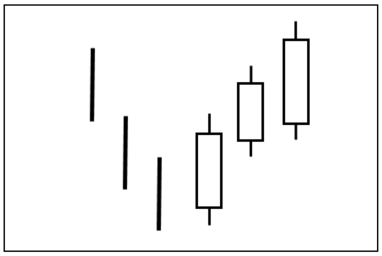

Figure 3. Three white soldiers pattern.

The opposite of the three white soldiers pattern is known as the *three black crows* pattern which is interpreted as a downtrend signal of market trend.

C. Criticism of Candlestick Patterns

The major criticism of the candlestick chart patterns is that the patterns are qualitatively described with words, such as "long/short candlesticks," "higher/lower trading," "strong/weak signal," supported by some illustrations [4]. What percentage of price change does "long/short" mean? Without modeling the candlestick patterns in a way that a computer can process and performing experiments comprehensively based on the statistics of candlesticks, arguments on the effectiveness of chart patterns would not come to an end.

Secondly, the candlestick chart patterns do not deal with market liquidity. Liquid market refers to a market in which there are many buyers and sellers and in which transactions of stocks rapidly take place. In a liquid market, stock price trend to change relatively small. The analogy holds for an illiquid market. It seems necessary to take account of the market's liquidity to improve the predictability of the candlestick chart patterns. We use statistics of candlesticks of market under study to cope with the first and second criticisms.

Finally, the candlestick chart patterns are described under the assumption that candlestick will occur consecutively, which is often not true for actual stock price movements. This study proposes an algorithm using dynamic programing technique for retrieving similar candlestick charts. The algorithm is designed to provide optimal matching between two given price sequences including several noisy candlesticks. The function of eliminating noisy candlesticks distinguishes this study from other ones.

IV. Proposed Model for Retrieving Candlestick Patterns

This section describes a model for retrieving similar candlestick charts. Following a problem definition, the principle of eliminating noisy candlesticks are described. A dynamic programing technique is used to implement the proposed model. Statistics of candlesticks are calculated in order to estimate the six parameters of the proposed model.

A. Parameters Featuring Candlestick Patterns

As a preliminary stage of study, experiments only using the closing prices and the length of real bodies are performed. The experiments simply correspond to the conditions of the candlestick chart patterns [4]. The results are discouraging. Although mined stock price sequences are similar before the specified period of the reference date, trends of the sequences after the reference date are seemed to be random. Analyses of the results show that the randomness occurs due to the relative position among the stock price, the 5-day moving averages, and the 25-day moving averages.

Based on the results of the preliminary experiments, we propose a model for retrieving similar candlestick charts. Figure 4 depicts the model that consists of the six parameters as follows:

(1) Change of prices w.r.t previous closing price,
(2) Length of candlestick body,
(3) Difference from 5-day moving average,
(4) Difference from 25-day moving average,
(5) Slope of 5-day moving average,
(6) Slope of 25-day moving average.

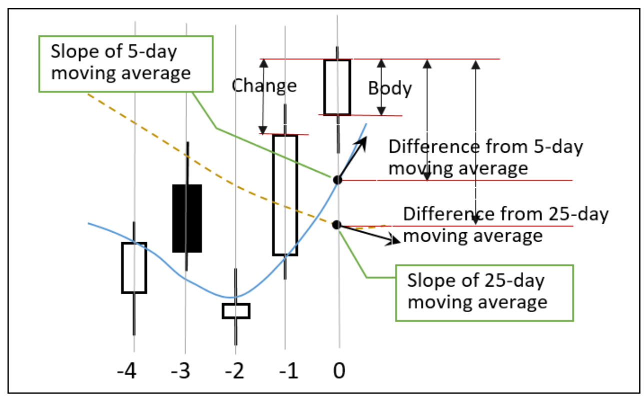

Figure 4. Candlestick pattern retrieval model.

The proposed model is unique because it uses two moving averages and their slopes, while the previous studies [6]-[15] do not deal with them. Relative position among a stock price, 5-day moving average, and 25-day moving average is significant to identify the zone where the candlestick pattern under consideration occurs, which is vital information for applying the candlestick pattern. The slopes of the moving averages are also important to identify their trends, e.g., an uptrend, a downtrend or a sideway (flat).

B. Problem definition and approach to solution

Let r_i ($1 \leqq i \leqq m$) and t_j ($1 \leqq j \leqq n$) represent candlesticks r_i and t_j which are defined by a vector of six parameters mentioned in (1) - (6). Let $R = (r_1,...,r_i,...,r_m)$ denote a reference candlestick pattern, and $T = (t_1,...,t_j,...,t_n)$ denote a test candlestick pattern with lengths of m and n, respectively. The problem we are dealing with is to find maximum matching of candlesticks r_i and t_j while eliminating unmatched ones.

Dynamic programming is a computer programming method that finds optimal solutions of a complicated problem. There are dozens of dynamic programming algorithms that implement optimal matching between sequences under certain criteria. Among them, we focus on finding the longest price sequence in common between stock price sequences. Because the LCS algorithm essentially finds the longest matched elements while discarding unmatched elements of sequences [16], the LCS algorithm fulfills our intention of eliminating noisy candlesticks.

The algorithm apparently satisfies the requirements of finding the longest price sequence. However, it is originally developed for strings of characters. The fact motivates us to modify the LCS algorithm to handle numeric sequences.

C. nLCS: LCS for Numerical Subsequences

The LCS algorithm is originally developed for character strings. Finding the LCS between two strings is described as follows. Given two strings, find the longest character subsequence that presents in both of them. Characters of the subsequence appear in the same relative order, but not necessarily contiguous. Figure 5 depicts the LCS of the two strings "246612" and "3651." Since elements of sequences are interpreted as characters that require an exact match, the LCS is "61."

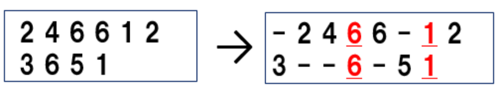

Figure 5. The LCS of two character sequences, "246612" and "3651."

It is rather easy to improve the LCS algorithm to deal with numerical sequences (nLCS) by interpreting each element as a number and using a tolerance given by a user. If the difference of two numbers is not greater than the given tolerance, then the two numbers are regarded as the same. For example, let the tolerance be set to one, and the two number sequences be "246612" and "3651." The nLCS are "2661" and "3651" as shown in Figure 6.

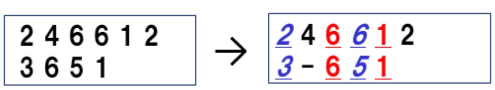

Figure 6. The nLCS of the two number sequences "246612" and "3651" for the tolerance of one.

The LCS and nLCS are formally defined as follows.

LCS algorithm: Let the input sequences be $\boldsymbol{X}$= (x[1],…, x[m]) of length m and $\boldsymbol{Y}$= (y[1],…,y[n]) of length n. Let D[i, j] denote the length of the longest common subsequence of x[i] and y[j] for $0 \leqq i \leqq m$ and $0 \leqq j \leqq n$.

A) If either sequence or both sequences are empty, then the LCS is empty, i.e., D[i, 0] = 0 and D[0, j] = 0.

B) If x[i] and y[j] match, i.e., x[i] = y[j], then the LCS is become longer than the previous sequences by one, i.e., D[i, j] = D[i−1, j−1] + 1.

C) If x[i] and y[j] do not match, i.e., x[i] ≠y[j], then the LCS is the maximum of the previous sequences, i.e., max(D[i−1, j], D[i, j−1]).

The value of D[m, n] is the LCS of the sequences $\boldsymbol{X}$= (x[1],…,x[m]) and $\boldsymbol{Y}$= (y[1],…,y[n]). The actual LCS sequence can be extracted by traversing the matrix D[i, j].

nLCS algorithm: The nLCS algorithm is derived from the LCS algorithm by replacing the match condition x[i] = y[j] with (x[i] − y[j]) ≦ diff where diff is a tolerance given by a user.

D. nLCSm: LCS for Subsequences with Multi Numerical Attributes

The idea of deriving the nLCS from the LCS can be further extend to the multi numerical attributes to obtain the nLCS for subsequences with multi numerical attributes (nLCSm).

nLCSm algorithm: Let p ($1 \leqq p$) denote the number of numerical attributes. Let C_q ($1 \leqq q \leqq p$) denote the match conditions for the q^{th} numerical attribute. The nLCS is extended with respect to the multiple numerical attributes, named the nLCSm. The nLCSm is derived by replacing the match condition of the nLCS, i.e., ((x[i] − y[j]) ≦ diff), with ($C_1 \wedge ... \wedge C_q \wedge ... \wedge C_p$).

Figure 7 shows an overview of implementation of the nLCSm algorithm. In principle, the nLCSm algorithm can be implemented by replacing the matching condition of the LCS algorithm with $C_1 \wedge ... \wedge C_p$ as shown in line 6 of Figure 7. However, the effort to implement the matching condition $C_1 \wedge ... \wedge C_p$ depends on the complexity of each matching condition.

```
1  for (int i = 0; i<= m; i++){ D[i][0] = 0;}
2  for (int j = 0; j<= n; j++){ D[0][j] = 0;}
3  // === Compute nLCSm
4  for (int i = 1; i<= m; i++){
5    for(int j = 1; j<= n; j++){
6      if ( C1∧...∧Cp ){
7        D[i][j] = D[i – 1][j – 1] + 1;
8      } else {
9        D[i][j] = Max(
                 D[i][j – 1], D[i – 1][j]);
10     }
11   }
12 }
```

Figure 7. Overview of implementation of the nLCSm algorithm.

In this study, it takes approximately 600 code lines to implement the proposed model consisting of the six parameters shown in Figure 4, which is described in the next section.

E. nLCSm and candlestick pattern retrieval

Given the candlestick pattern model with six parameters as depicted in Figure 4, the nLCSm algorithm can be applied to implementing the model by assigning match conditions C_1 to C_6 for each candlestick as follows.

C_1: if a difference between closing price change of a reference candlestick and that of a test candlestick is within the change tolerance (*change_tol*), then C_1 is true.

C_2: if a difference between body length of a reference candlestick and that of a test candlestick is within the body tolerance (*body_tol*), then C_2 is true.

C_3: if a difference between a closing price of a reference candlestick and a 5-day moving average is within the tolerance (*av5diff_tol*), and a difference of a test candlestick and a 5-day moving average is within the tolerance, then C_3 is true.

C_4: if a difference between a closing price of a reference candlestick and a 25-day moving average is within the tolerance (*av25diff_tol*), and a difference of a test candlestick and a 25-day moving average is within the tolerance, then C_4 is true.

C_5: if a difference between a slope of a 5-day moving average of a reference candlestick and that of a test candlestick is within the given tolerance (*slope5_tol*), then C_5 is true.

C_6: if a difference between a slope of a 25-day moving average of a reference candlestick and that of a test candlestick is within the given tolerance (*slope25_tol*), then C_6 is true.

F. Statistics of candlesticks

In order to estimate the six parameters of the proposed model shown in Figure 4, we calculate the statistics of the candlesticks of the daily NASDAQ composite index from Jan. 2, 2009 to Aug. 20, 2018 of 2,425 business dates.

Figure 8 shows the frequency diagram of body length ratio. Let Open[i] and Close[i] denote the opening and closing price during the date i ($0 \leqq i < 2{,}425$), where i=0 means the current day. The body length ratio is computed by the following formula:

$$\text{Body length ratio} = (\text{Close}[i] - \text{Open}[i]) * 100 / \text{Close}[i] \quad (1)$$

The body length ratios are aggregated for every 0.1%. We see that the frequency diagram apparently follows the normal distribution [4]. The average is 0.03216%, while the standard deviation is 0.9330%.

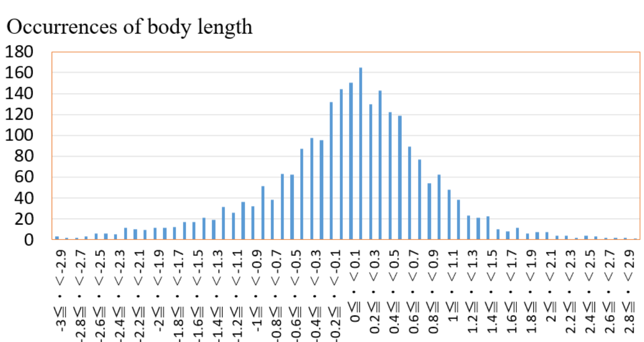

Figure 8. Frequency diagram of body length ratio distributions.

Based on the standard deviation, we divide the body length into seven categories in accordance with the classification of many literatures on candlestick charting [10]. Uptrend and downtrend candlestick bodies are divided into three each. The other one is a *Doji* [4], which is formed when the opening price and the closing price are nearly equal. Table I shows the seven categories of candlestick body. We define the values in Table I so that the candlestick bodies except for the *Doji* occur with the same probability of approximately one sixth. The candlestick categories are involved as a retrieval condition in addition to the tolerances of change, body length, difference from 5-day and 25-day averages, etc.

TABLE I. SEVEN CATEGORIES OF CANDLESTICK BODY.
SDV: Standard deviation of body length ratio.
BD: Parameter for Doji; 0.2% for uptrend, 0.25% for downtrend.

	Upper bound	Lower bound
Long price up body	∞	0.97*SDV
Medium price up body	0.97*SDV	0.44*SDV
Short price up body	0.44*SDV	BD
Doji	BD	− BD
Short price down body	− BD	−0.44*SDV
Medium price down body	−0.44*SDV	−0.97*SDV
Long price down body	−0.97*SDV	− ∞

Figure 9 shows the frequency diagram of change ratio distributions. Change is defined by the difference between the current value and the previous day's market close. The change ratio is computed by the following formula for each j ($0 \leqq j < 2{,}424$):

$$\text{Change ratio} = (\text{Close}[j] - \text{Close}[j+1]) * 100 / \text{Close}[j+1] \quad (2)$$

We see that the frequency diagram is in a form of the normal distribution [4]. The average is 0.07127%, while the standard deviation is 1.149%.

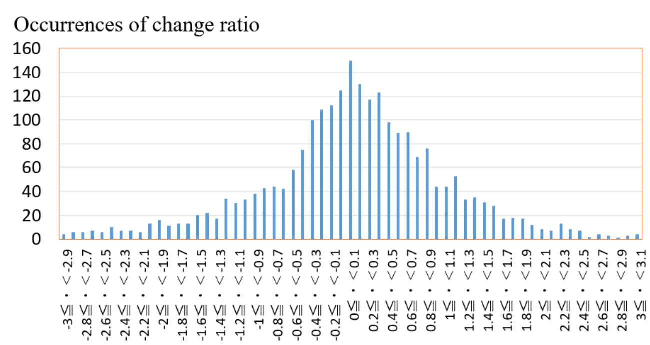

Figure 9. Frequency diagram of change ratio distributions.

Figure 10 shows the frequency diagram of 5-day average ratio distributions. Let Avg5[k] denote the 5-day average of the change ratios during the date k ($0 \leqq k < 2{,}420$). The 5-day average ratio is computed by the following formula:

5-day average ratio=

$$(\text{Avg5}[k] - \text{Avg5}[k+1])*100 \,/\, \text{Avg5}[k+1] \qquad (3)$$

The average is 0.0661%, while the standard deviation is 0.4880%. Since a 5-day average is the mean of five consecutive close prices, the 5-day average ratio statistically follows the standard deviation of the sample [4] of five close prices in theory. In fact, the standard deviation of the 5-day average ratio of 0.4880% roughly equals 1.149% / SQRT(5)= 0.5138% with the margin of error of 0.0258%. The error seems to occur because the change ratios do not strictly follow the normal distribution.

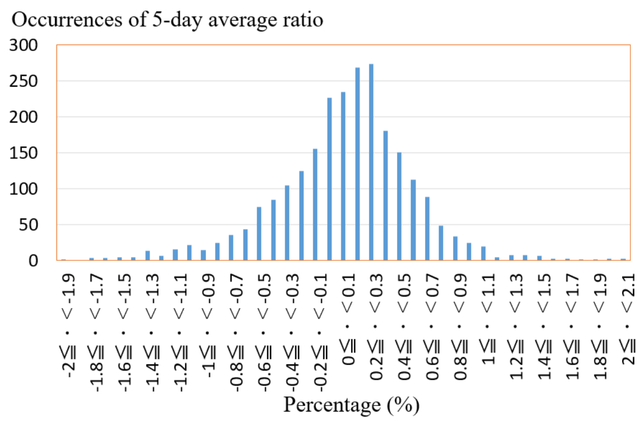

Figure 10. Frequency diagram of 5-day average ratio distributions.

Figure 11 shows the frequency diagram of 25-day average ratio distributions. Let Avg25[m] denote the 25-day average of the change ratios during the date m ($0 \leqq m < 2{,}400$). The 25-day average ratio is computed by the following formula:

25-day average ratio=

$$(\text{Avg25}[m] - \text{Avg25}[m+1])*100 \,/\, \text{Avg25}[m+1] \qquad (4)$$

The average is 0.0680%, while the standard deviation is 0.1910%. The standard deviation of the 25-day average ratio of 0.1910% roughly equals 1.149% / SQRT(25)= 0.2298% with the margin of error of 0.0388%.

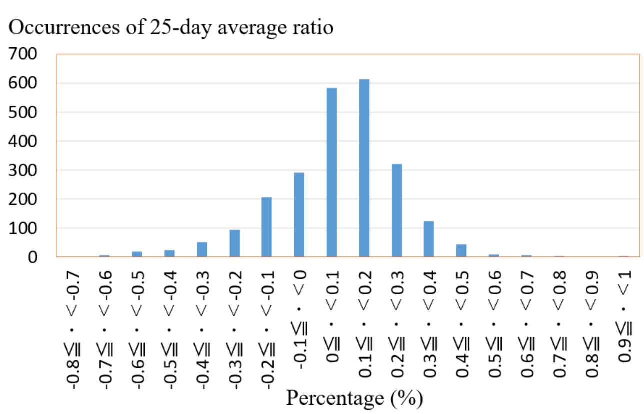

Figure 11. Frequency diagram of 25-day average ratio distributions.

The tolerance of 5-day moving average *av5diff_tol* is also statistically dependent on the change tolerance *change_tol*. In the proposed retrieval model, *av5diff_tol* and *av25diff_tol* are calculated by the following formulas as defaults according to the definition of the standard deviation of the sample [4].

$$\textit{av5diff\_tol} = \textit{change\_tol} \,/\, \text{SQRT}(5) = \textit{change\_tol} \,/\, 2.236 \qquad (5)$$

$$\textit{av25diff\_tol} = \textit{change\_tol} \,/\, \text{SQRT}(25) = \textit{change\_tol} \,/\, 5 \qquad (6)$$

Therefore, there are essentially four independent parameters in the proposed model, which still causes difficulties in setting parameters. Assuming that each parameter has 5 ranges of values representing, for instance, very high, high, the same level, low, and very low. The candlestick patterns of one candlestick have 5 to the power 4, i.e., 5^4= 625 cases of parameters. The patterns composed of two candlesticks have 5^(4*2)= 625*625= 390,625 cases. The patterns of tree candlesticks have 244,140,625 cases. These cases mean very wide varieties of candlestick charts leading difficulties even in setting parameters for retrieving a specific candlestick chart pattern. In fact, the experiments are performed by repeating trial and error while adjusting parameters.

V. EXPERIMENTAL RESULTS

The predictabilities of the *morning star* pattern indicating a reversal signal of starting uptrend trends, and the *downtrend engulfing* pattern indicating the end of uptrend trends [4] are evaluated through experiments. The experiments are performed on the daily historical stock prices of the NASDAQ composite index of 2,425 business dates from Jan. 2, 2009 to Aug. 20, 2018.

A. Data Conversion

The stock prices are converted to the ratio of closing prices to reduce the effects of highness or lowness of the stock prices. The formula below is used for calculating the ratio of prices in a percentage.

$$RCP_j = (CP_j - CP_{j+1})*100 \,/\, CP_{j+1} \;\; (1 \leqq j \leqq n) \qquad (7)$$

CP_j indicates the closing price of the j-th business date. CP_1 means the closing price of the current date. RCP_1 is the ratio of the difference between CP_1 and CP_2, and the closing price of CP_2, i.e., the closing price of the date before the current date. The similar calculations are performed to opening, high, and low prices. In addition, the 5-day and 25 day moving averages, and their slopes are calculated before the experiments. The number of valid 25 day moving averages, i.e., n in effect is 2,400 (= 2,425 − 25) because the 25-day averages can't be calculated to the last 25 days.

B. Experiments on Morning Star Pattern

In the previous contribution [1], we observe that the morning star pattern does not necessarily show stable uptrend signal, and one confirmation day after the pattern significantly improves the predictability of the pattern. This is true for the NASDAQ market. So we start this section with the morning star pattern plus one confirmation day.

Figure 12 shows the candlestick chart of the NASDAQ in which a strong uptrend starts on May 4, 2018 with a long uptrend body of 1.9966%. The short uptrend body of 0.3219% follows the next day, i.e., May 7, 2018. The candlestick arrangements from May 1 to May 7, 2018 form a morning star pattern with one confirmation day.

The first experiment is performed on the four candlesticks surrounded by a solid rectangle in Figure 12 as a reference for retrieval. This means the length of the reference candlesticks is four. We set the length of the candlesticks to be retrieved to eight, i.e., double the length of the reference.

The change tolerance varies from 0.3% to 1.8%. The tolerances of 5-day and 25-day moving average are calculated by (5) and (6). The slope of 5-day moving average is set to 0.4880% in an upward direction, i.e., the standard deviation of the 5-day average ratio, and −0.4880*0.25% in a downward direction. The slope conditions mean to select approximately 44% of the original candlesticks. The slope of 25-day moving average is set analogously.

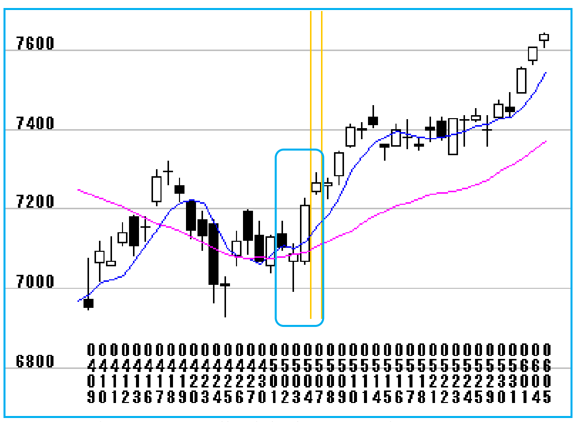

Figure 12. Candlestick chart around May 7, 2018.

Figure 13 shows the number of retrieved business dates that have a candlestick pattern similar to that of May 7, 2018 within a change tolerance ranging between 0.3% and 1.8%. The number of data gradually increases as the change tolerance rise, but it does not rise linearly. The algorithm retrieves almost the same number of business dates for tolerance between 0.8% and 1.2%, and between 1.3% and 1.8%.

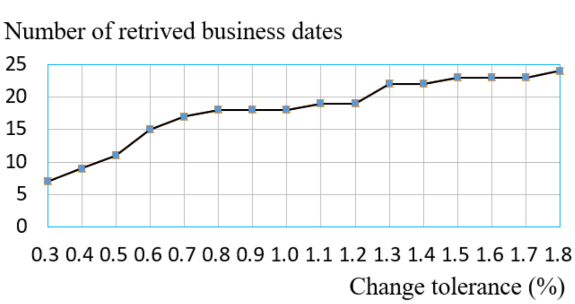

Figure 13. Number of retrieved business dates and change tolerance.

Table II shows a list of the business dates retrieved under the change tolerance of 1.1%. "nLCSm" shows the length of candlesticks that match the key for retrieval. "Length" represents the length of the retrieved candlesticks. "nLCSm/Length" indicates the *"density"* of the matched candlesticks in the retrieved ones. So, "nLCSm^2/Length" means the *"weight"* of the retrieved candlesticks and used as a key to sort in descending order.

TABLE II. RESULTS CONCERNING MORNING STAR PATTERN PLUS ONE CONFIRMATION DAY ON MAY 7, 2018 UNDER THE CHANGE TOLERANCE OF 1.1%.

Bus.Date	nLCSm	Length	nLCSm/Length	nLCSm^2/Length
20180507	4	4	1.000	4.000
20120119	4	5	0.800	3.200
20180709	4	6	0.667	2.667
20160711	4	6	0.667	2.667
20160525	4	6	0.667	2.667
20130503	4	6	0.667	2.667
20150319	4	7	0.571	2.286
20131223	4	7	0.571	2.286
20121129	4	7	0.571	2.286
20110705	4	7	0.571	2.286
20171218	3	4	0.750	2.250
20151016	3	4	0.750	2.250
20130411	3	4	0.750	2.250
20121120	3	4	0.750	2.250
20110527	3	4	0.750	2.250
20100914	3	4	0.750	2.250
20090427	3	4	0.750	2.250
20150123	4	8	0.500	2.000
20100308	4	8	0.500	2.000

Figure 14 shows overlapped closing prices whose business dates are listed in Table II for graphically representing the future stock trend. All business dates are aligned on the origin to make the comparison easy. The thick black line represents the closing prices of the reference date, i.e., May 7, 2018. Thin solid lines represent the closing prices of business dates listed in Table II except for the reference date. The thick light blue line indicates the average of the candlestick charts plotted by thin solid lines, which suggests at most 1% increase in the coming 10 days.

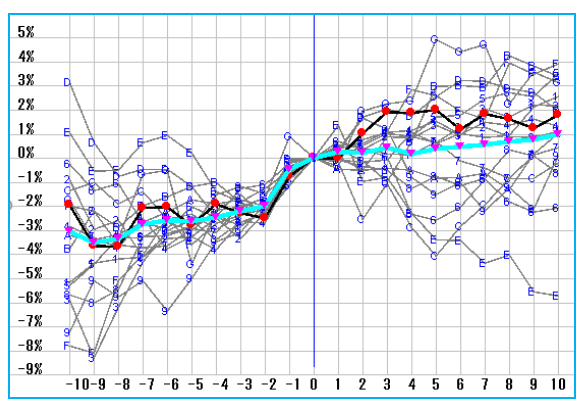

Figure 14. Overlapped closing prices of retrieved candlesticks using May 7, 2018 as key.

Figure 15 shows overlapped closing prices of business dates whose *"weight"* are not less than 2.286. Six out of nine closing prices suggest an uptrend, while the other three out of seven ones suggest a downtrend. The average of the nine closing prices is almost the same as that of eighteen closing prices shown in Figure 14 since nine closing prices including those showing extreme fluctuations of both an uptrend and downtrend are deleted.

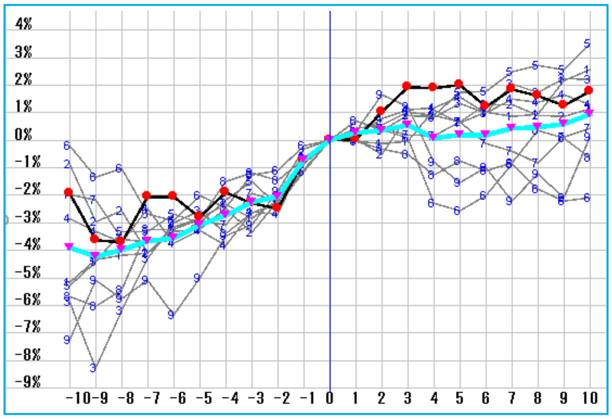

Figure 15. Overlapped closing prices of business dates whose weight are not less than 2.286.

Figure 16 shows overlapped closing prices of business dates whose weight are not less than 2.667. All five closing prices suggest an uptrend. This result may well be noteworthy for traders to identify buying opportunities with expectation of approximately 2.0% profits on average in the next ten days.

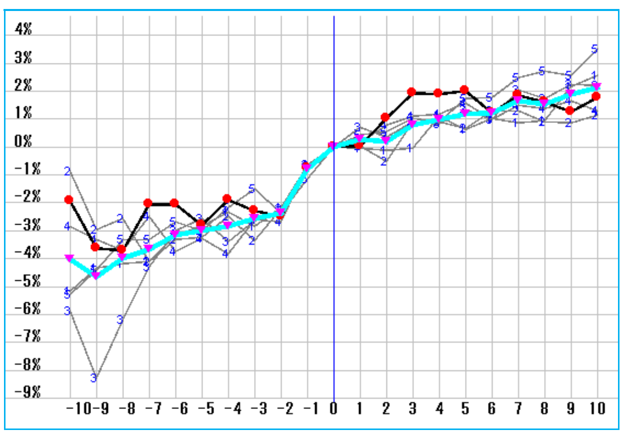

Figure 16. Overlapped closing prices of business dates whose weight are not less than 2.667.

We should learn from the retrieval results shown in Figure 14 through 16 that there are price movements opposite to what candlestick pattern predicts. As the number of the retrieved candlestick sequences increases, the average value of the retrieved sequences tends to approach 0, meaning that candlestick pattern technique has no predictive power on price movements. It is vital to control the number of retrieved data to be appropriate for prediction.

Figure 17 shows a candlestick chart of the NASDAQ around Jan. 19, 2012. The proposed nLCSm algorithm returns four matching candlesticks in five business days, which is the best match in the list shown in Table II. Though the category of the candlestick body is different, the same pattern takes place on Jan. 6, 2012 showing the good opportunity to buy stocks as indicated by an arrow on Figure 17.

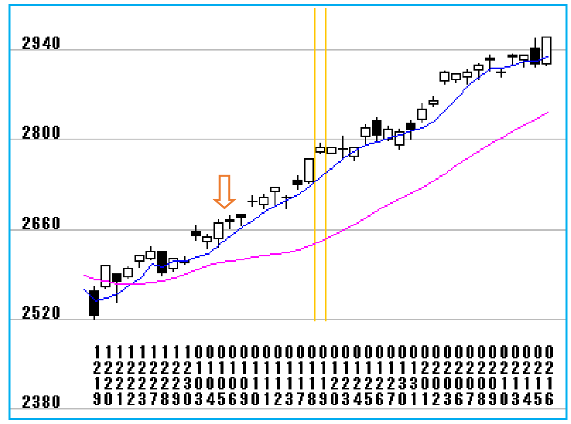

Figure 17. Candlestick chart around Jan. 19, 2012.

Figure 18 shows the candlestick chart around July 11, 2016. This chart is similar to that of May 7 in Figure 12 in the sense that a few days passed after the 5-day average exceeds the 25-day average. The chart is worthy of the investor to pay attention to.

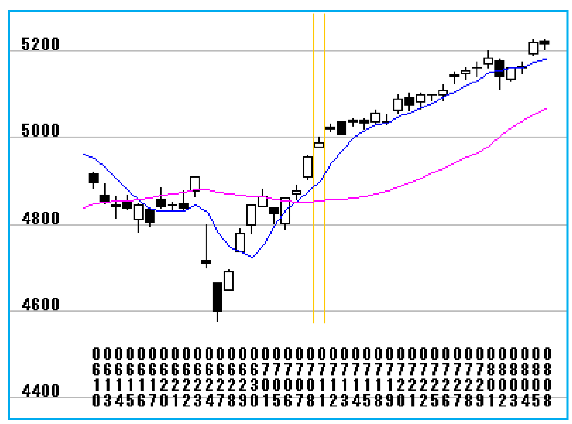

Figure 18. Candlestick chart around July 11, 2016.

Figure 19 shows the candlestick chart around July 9, 2018. The key candlestick pattern of retrieval just comes up on the day that the 5-day average exceeds the 25-day average, which seems to imply a rather weak uptrend. In fact, the opening and closing prices move in a zigzag way, and draw a downtrend after 14 days. The candlestick chart around May 25, 2016 in Figure 20 shows the same feature and changes the trend after 10 days.

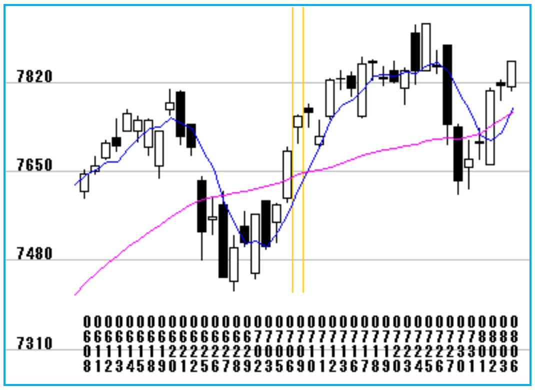

Figure 19. Candlestick chart around July 9, 2018.

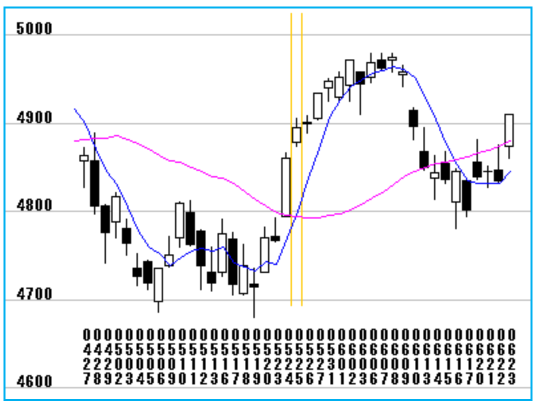

Figure 20. Candlestick chart around May 25, 2016.

The candlestick chart around May 3, 2013 shows the similar uptrend pattern.

C. Experiments on downtrend engulfing pattern

Figure 21 shows the candlestick chart of the NASDAQ around June 21, 2018 with a long downtrend body preceded by a short candlestick. The pattern is named the *downtrend engulfing* pattern [4] suggesting downtrend reversal. The second experiment is performed on the four candlesticks surrounded by a solid rectangle in Figure 21 as a reference for retrieval. We set the length of the retrieved candlesticks to eight, which is the same as the first experiment.

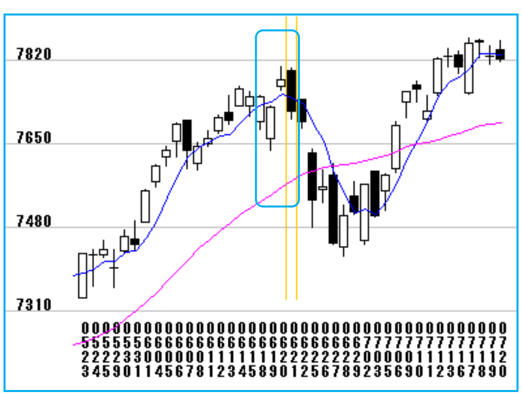

Figure 21. Candlestick chart around June 21, 2018.

Figure 22 shows the number of retrieved business dates that have a candlestick pattern similar to that of June 21, 2018 within a change tolerance ranging between 0.3% and 1.8%. The number of data gradually increases as the change tolerance rise, but it does not rise linearly.

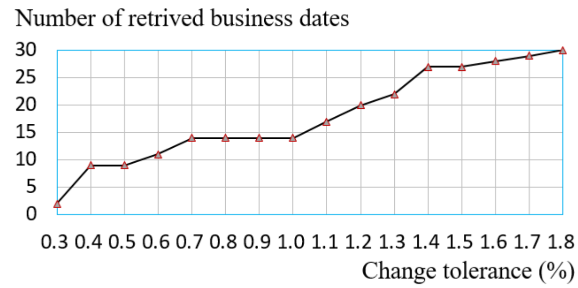

Figure 22. Number of retrieved business dates and change tolerance.

Table III shows a list of the business dates retrieved under the change tolerance of 0.8%. The contents of Table III are sort in descending order by the “nLCSm^2/Length” column.

TABLE III. RESULTS CONCERNING MORNING STAR PATTERN PLUS ONE CONFIRMATION DAY ON June 21, 2018 UNDER THE CHANGE TOLERANCE OF 0.8% .

Bus.Date	nLCSm	Length	nLCSm/Length	nLCSm^2/Length
20180621	4	4	1.000	4.000
20110408	4	5	0.800	3.200
20161130	4	6	0.667	2.667
20161027	4	6	0.667	2.667
20170727	3	4	0.750	2.250
20100621	3	4	0.750	2.250
20150806	4	8	0.500	2.000
20120720	4	8	0.500	2.000
20090923	4	8	0.500	2.000
20131118	3	5	0.600	1.800
20130522	3	5	0.600	1.800
20141201	3	6	0.500	1.500
20120801	3	6	0.500	1.500
20090814	3	7	0.429	1.286

Figure 23 shows overlapped closing prices whose business dates are listed in Table III. The thick light blue line, representing the average of the retrieved candlesticks, indicates that the sideways are expected after two-days short dips with the wide price fluctuations ranging between 3.5% and −8.9%.

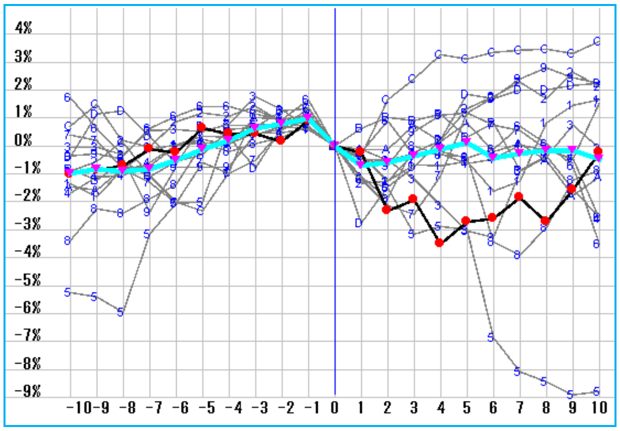

Figure 23. Overlapped closing prices of retrieved candlesticks using June 21, 2018 as key.

Figure 24 shows overlapped closing prices of business dates whose weight are not less than 2.0. Five business dates with the weight between 1.286 and 1.80 are discarded including one forming the upper bound. As a result, the average closing prices of the retrieved candlesticks decreases gradually to 1%.

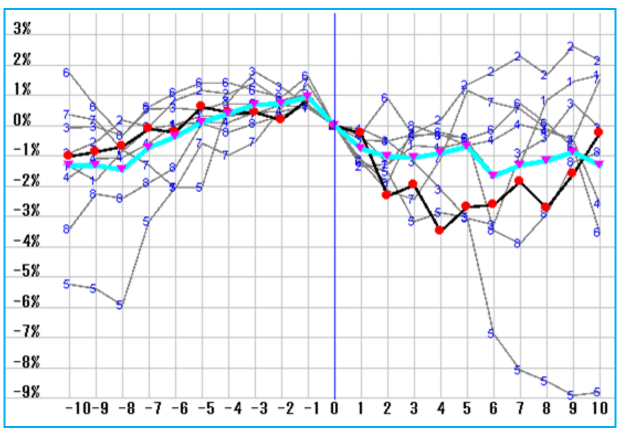

Figure 24. Overlapped closing prices of business dates whose weight are not less than 2.0.

Figure 25 shows overlapped closing prices of business dates whose weight are not less than 2.25. The prices of the business dates with the weight between 1.286 and 1.80 are located between the upper bound and the lower bound. Consequently, the average closing prices of the retrieved candlesticks seems to remain the same as that of Figure 24.

Figure 25 suggests two possible price trends after two-day drips. One is an uptrend reversal and the other is a continuous downtrend trend as shown by the upper and lower bounds of Figure 25, i.e., lines marked by "2" and "5".

Figure 26 shows the candlesticks around Nov. 30, 2016, which forms the upper bound depicted in Figure 25. Figure 27 shows the candlesticks around June 21, 2010 forming the lower bound.

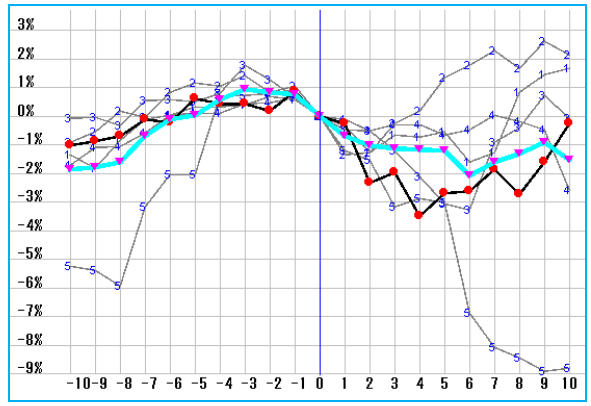

Figure 25. Overlapped closing prices of business dates whose weight are not less than 2.25.

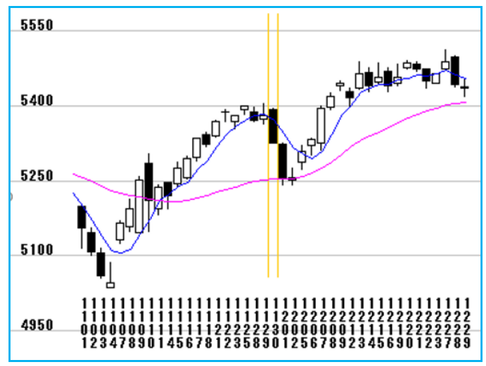

Figure 26. Candlestick chart around Nov. 30, 2016.

Figure 27. Candlestick chart around June 21, 2010.

The candlestick chart in Figure 26 shows the downtrend engulfing pattern on Nov. 30, 2016, while forms the *morning star* pattern three days later, i.e., Dec. 5.

Figure 27 shows a continuous downtrend. To the best our knowledge, the candlestick charting can tell when the specified trend begins, but *cannot* tell when the reverse of the specified trend begins. The results suggest the limitation of the stock price prediction by the candlestick charting.

In the study, we find that the prediction of the downtrend trend is more difficult than that of the uptrend trend. The fact seems to reflect that the NASDAQ stock market is constantly growing. In fact, the NASDAQ index on Aug. 20 2018 is 2.16 times the price on Aug. 20 2013.

VI. CONCLUSION AND FUTURE WORK

Extracting stock price change patterns and predicting future stock prices are an interesting task for traders as well as financial data analysts. Typically, when there is no outstanding news, stock price is apt to show small price change during the period. These small price changes are generally interpreted as market indecisiveness and need to be eliminated when retrieving similar stock prices.

In this paper, we propose an algorithm for matching candlesticks while skipping small price changes. A numerical sequence version of the LCS (Longest Common Substring) algorithm, which makes use of dynamic programming technique, is devised for the purpose.

This paper also proposes a model with six parameters that provide bases for retrieving similar candlestick patterns. The proposed model is original because it deals with slopes of the 5-day and 25-day moving averages to identify their trends, in addition to 5-day and 25-day moving averages to decide whether the price occurs in high or low price zones.

In this study, we employ statistical values of the candlesticks to estimate the parameters of the model. The standard deviation of the sample [4] is substantially helpful to approximate the parameters concerning the 5-day and 25-day moving averages.

Experiments are performed on the daily NASDAQ composite index. The results of the experiments seem to show that the forecast for the uptrend trend is more accurate than that of downtrend, which reflects that the NASDAQ stock market is constantly growing.

We also find limitations of the current approach though the experiments. First, the determination of the value of the parameters is still immature. Actually, it is done by trial and error all through the experiments. We need to develop a systematic method to estimate the value of the parameters and the number of retrieved candlestick data. Second, the proposed model fails to support trading volume. According to literature on technical analysis [4], the volume is an important indicator to confirm the strength or weakness of price movements. Finally, the proposed model need to support parameters on upper and lower shadows. Support of shadow parameters allows the model to examine various shadow sensitive patterns known as the *hammer, dragon fly* patterns [4].

As the current study is limited to an average or composite index of stock markets, future researches will focus on analyzing stock prices of industry sectors and individual companies in international stock markets.

Our future plans also include developing an automatic pattern mining method to discover frequently repeated patterns for stock trend prediction. Since candlestick patterns proposed so far have been found from human experience, there can be unknown profitable candlestick patterns with more complicated structures than ever known. Finding noticeably profitable patterns stimulates our interest in researches of data mining.

ACKNOWLEDGMENT

This research is supported by the JSPS KAKENHI under grant number 16K00161.

REFERENCES

[1] Y. Udagawa, “Design and Implementation of Candlestick Chart Retrieval Algorithm for Predicting Stock Price Trend,” The Fourth International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2018), pp. 19–25, 2018.

[2] Yahoo! Finance, “NASDAQ Composite, Historical Data,” https://finance.yahoo.com/quote/%5EIXIC/history?p=%5EIXIC, Aug. 2018.

[3] V. Drakopoulou, “A Review of Fundamental and Technical Stock Analysis Techniques,” Journal of Stock & Forex Trading vol. 5, pp. 1–8, Nov. 2015.

[4] “Technical Analysis,” Cambridge Univ. pp. 1–179, Available from: http://www.mrao.cam.ac.uk/~mph/TechnicalAnalysis.pdf, Feb. 2011.

[5] Fusion Media Limited. “Nikkei 225 Futures,” https://www.investing.com/indices/japan-225-futures-candlestick, Aug. 2018.

[6] J. M. Horton, “Stars, crows, and doji: The use of candlesticks in stock selection,” Quarterly Review of Economics and Finance, vol. 49, pp. 283–294, Nov. 2007.

[7] R. B. Marshall, R. M. Young, and R. Cahan, “Are candlestick technical trading strategies profitable in the Japanese equity market?,” Review of Quantitative Finance and Accounting, vol. 31, pp. 191–207, Aug. 2008.

[8] P. Tharavanij, V. Siraprapasiri, and K. Rajchamaha, “Profitability of Candlestick Charting Patterns in the Stock Exchange of Thailand,” SAGE journals, pp. 1–18, Oct. 2017.

[9] G. Caginalp and H. Laurent, “The predictive power of price patterns,” Applied Mathematical Finance, vol. 5, 1998, pp. 181–206.

[10] Y.-J. Goo, D.-H. Chen, and Y.-W. Chang, “The application of Japanese candlestick trading strategies in Taiwan,” Investment Management and Financial Innovations, vol. 4, pp. 49–79, Jan. 2007.

[11] C. Chootong and O. Sornil, “Trading Signal Generation Using A Combination of Chart Patterns and Indicators,” International Journal of Computer Science Issues, vol. 9, pp 202–209, Nov. 2012.

[12] T.-H. Lu, Y.-C. Chen, and Y.-C. Hsu, “Trend definition or holding strategy: What determines the profitability of candlestick charting?,” Journal of Banking & Finance, vol. 61, Dec. 2015, pp. 172–183.

[13] M. Zhu, S. Atri, and E. Yegen, “Are candlestick trading strategies effective in certain stocks with distinct features?” Pacific Basin Finance Journal, vol. 37, pp. 116–127, Apr. 2016.

[14] F. Martinsson and I. Liljeqvist, “Short-Term Stock Market Prediction Based on Candlestick Pattern Analysis,” Examensarbete Inom Teknik, Stockholm, Sverige, June, 2017, pp. 1–36.

[15] C.-L. Chin, M. Jais, S. S. Balia, and M. Tinggi, “Is candlestick continuation patterns applicable in Malaysian stock market?,” SHS Web of Conferences 34, 2017.

[16] “Longest common subsequence problem,” https://en.wikipedia.org/wiki/Longest_common_subsequence_problem, Sept. 2018.

[17] E. F. Fama, “Efficient capital markets: A review of theory and empirical work,” Journal of Finance, vol. 25, 1970, pp. 383–417.