Intelligent Stock Market Assistant using Temporal Data Mining

Gerasimos Marketos[1], Konstantinos Pediaditakis[2], Yannis Theodoridis[1], and Babis Theodoulidis[2]

[1] Database Group, Information Systems Laboratory,
Department of Informatics, University of Piraeus, Greece
{marketos, ytheod}@unipi.gr
http://isl.cs.unipi.gr/db
[2] School of Informatics, University of Manchester, UK
pediad@co.umist.ac.uk, babis.theodoulidis@manchester.ac.uk

Abstract. The stock market domain is a dynamic and unpredictable environment. Traditional techniques, such as fundamental and technical analysis can provide investors with some tools for managing their stocks and predicting their prices. However, these techniques cannot discover all the possible relations between stocks and thus there is a need for a different approach that will provide a deeper kind of analysis. Data mining can be used extensively in the financial markets and help in stock-price forecasting. Therefore, we propose in this paper a portfolio management solution with business intelligence characteristics.

1 Introduction

The main function of a stock market is the dealings of stocks between investors. Stocks are grouped into industry groups according to their primary business focus (e.g. IT, Banks, Manufacturing). A transaction is the willing of an investor to sell some stocks and the request of another to buy them.

Each stock is not only characterized by its price but also by many others variables. There is an interaction among all these variables and only a deep study could show the behavior of a stock over time. The main variables are shown in the table below.

Table 1. Stock Variables

Variable	Description
Price	Current price of a stock
Opening Price	Opening price of a stock for a specific trading day
Closing Price	Closing price of a stock for a specific trading day
Volume	Stock transactions volume (buy/sell)
Change	Opening and Closing stock value difference
Change (%)	Percentile Opening and Closing stock value difference

Maximum price	Maximum stock price within a specified time interval (day, month etc.)
Minimum price	Minimum stock price within a specified time interval (day, month etc.)

Initial research in financial and stock trading issues lead to the identification of some factors that are considered among experts to influence the price of a stock. At first, it is a reasonable thought that the behavior of an investor depends on the size of the owner company (blue chips-middle-small). Furthermore, we could identify the following influence factors:

Table 2. Possible stock price influence factors

Influence Factor	Description
P/E factor	Price per annual earnings
Volume	How many dealings are taking place
Business Sector	The sector in which a stock belongs
Historical Behavior	Fluctuation of a stock over time
Rumors	There is a rule suggesting to “buy on rumors sell on news” so that may cause some unpredictable behavior
Book (Net Asset) Value	The accounting value of a company
Stock Earnings	Percentile difference of the stock price value over a period of time
Financial position of a company	The financial status of a company
Uncertainty	Are there any unpredictable factors?

Obviously, all these factors cannot be easily modeled and embedded in a tool, since some of them are related with human psychology.

Data mining has found increasing acceptance in business areas which need to analyze large amounts of data to discover knowledge which otherwise could not be found. Temporal data mining provides some additional capabilities required in cases where the evolution of the existing data and their interactions need to be observed through the time dimension. The stock market is one of them.

We propose a tool that collects stock data and after analyzing and interpreting them, it will be able to act on the basis of these interpretations [1]. The capabilities of this tool are based on temporal data mining patterns, extracted from stock market data. The ultimate goal of such a tool would be to support stock market traders in their basic function, by suggesting possible stock market trading transactions, when strong evidence of possible profit from these transactions is available. It should also take into consideration the different types of users and their characteristics with respect to the trading strategy that a certain user possesses. The design of such a tool proposed in this paper consists of two specific parts.

In the monitoring part, the user is able to define stock portfolios, to view stock price values over time of companies with similar characteristics (e.g. same business sector, price range, P/E etc.). Other functionalities include access to market and company news, ability of the user to define alerts, which would be triggered on the basis of specific events happening or not.

In the prediction part, the tool helps users to decide on their stock trades. A sequence mining algorithm is used in order to identify frequently occurring sequences of stock fluctuations and thus recommend some good trading opportunities, based on the extracted frequent patterns. However, the algorithm itself does not know what a good opportunity is. Therefore, we need to define interestingness measures that will allow the proposed system to discover such opportunities, based always on parameters that represent the user's trading strategy. The development of such a decision support tool introduces several challenging research issues:

- The incorporation of user-defined parameters into the system (preferences, orders)
- The pre-processing tasks that must be executed (definition of temporal hierarchies, generalization)
- The range of event types that will be used by the algorithm
- The store of the patterns produced by the temporal data mining algorithm
- The evaluation of these patterns (the weight of each variable) and how the results and the user defined parameters could affect the pre-processing tasks (optimization, generalization using different hierarchies)

The rest of the paper is organized as follows. In the next section we give a brief description of the temporal data mining research area and especially time series analysis and sequence mining. Section 3 presents an overview of the sequence mining approach that is used in our proposed system. Section 4 presents the prototype under development, describing its architecture and functionality. In section 5, we discuss related work and present other research prototypes. Finally, section 6 concludes with directions for future work.

2 Temporal Data Mining

Temporal data mining is a research field of growing interest in which techniques and algorithms are applied on data collected over time. According to Lin et al. [2], temporal data mining "is a single step in the process of Knowledge Discovery in Temporal Databases that enumerates structures (temporal patterns or models) over the temporal data".

Examples of temporal data mining tasks are classification and clustering of time series, discovery of temporal patterns or trends in the data, associations of events over time, similarity-based time series retrieval, time series indexing and segmentation. In the stock market domain, temporal data mining could indeed play an essential role. Identifying temporal patterns from the fluctuation of stock prices is a very complex problem. It is preferable to know the range of variation in both stocks prices, the period of time that this influence is likely to happen and also the statistical significance of the discovered rule.

2.1 Time Series

A sequence of continuous real-valued elements, such as stock prices is known as a time series. Time series form curves and can reveal trends through analysis, which leads to a large potential for analytical studies. The identification of trends takes place through the comparison of time series and the discovery of similar shapes between them, based on a predefined and domain-specific measure of similarity.

A fundamental problem that needs to be addressed before any attempt of trend discovery is the representation of the time series. The direct manipulation of continuous, high dimensional data in an efficient way is extremely difficult. The representation model selected, must also guarantee the compatibility between time series, since there can be scaling differences and gaps within the series among others that must be resolved before any similarity matching technique is applied. Piecewise linear transformation [3], Discreet Fourier Transformations [4] and Discreet Wavelet Transformations [5] are some of the techniques that have been extensively used to represent time series. The most common similarity measure used for the identification of trends is the Euclidean Distance.

2.2 Sequence Mining

Agrawal and Srikant [6] introduced the problem of mining sequential patterns from commercial databases. Our centre of attention will be restricted to the prediction of stock price behavior. This is a typical area where not only events of buying and selling stocks are interesting but also the sequence of them. It is considered that the fluctuation of a stock price is the result of previous stock events (buying, selling). Different events could lead to different prices. Subsequently, the idea is to predict such behaviors in order help the investors to optimize the management of their portfolios.

One basic problem [7], [8] in analyzing sequences of events is to find frequent event type subsequences, i.e. collections of event types that occur frequently with a certain order and within specific time spans. In the relevant literature many algorithms have been proposed giving different results. The proposed tool will use a novel sequence mining algorithm which is presented in [9].

The sequence mining algorithm used in our proposed system is able to extract statistically significant patterns of the form A→B[*t*] (event type B follows event type A within time *t*) from event sequences and focuses mainly on finding the optimal *t* for each of the extracted patterns.

3 Sequence Mining Approach

In the following paragraphs, some specific aspects and a more detailed description of the sequence mining approach used in our proposed system are given.

3.1 Event and Event Sequence Definitions

In the literature [10], [11] there are different definitions of the event concept. We could consider a given set $R = \{ A_1, \ldots, A_m \}$ of event attributes with respective domains $D_{A1}, \ldots, D_{Am}$. An event e over R is a (m + 1)-tuple ($a_1, \ldots, a_m$, t), where $a_i \in D_{Ai}$ and $t \in \aleph$, being the occurrence time of e. An event sequence S is a collection of events over R. For example in the stock market data, each event would refer to a certain stock, the price of the stock and the time of the transaction (i.e. we could choose focus on days, weeks or months)

In this approach, a simplified model of events has been adopted. Each event has a type and a time of occurrence. Therefore, given a class E of elementary event types, an event is a pair (e, t), where $e \in E$ and $t \in \aleph$. An event sequence S is an ordered sequence of events, i.e.:

$S = (e_1, t_1), (e_2, t_2), \ldots, (e_n, t_n)$ where $e_i \in E$ and $t_i \leq t_{i+1}$ for all $i \in \{1,\ldots, n-1 \}$.

We also assume in our simplified event model, that events happen spontaneously in time, i.e. they do not exhibit any duration. Moreover, the idea of the same exact event happening multiple times on the same time unit is of no particular meaning in our model.

Considering the event types A, B, C and D an example of such an event sequence would be:

$$S = (B, 4), (D, 6), (C, 6), (A, 8), (B, 12), (D, 15)$$

3.2 Simple Frequent Pattern Discovery in Event Sequences

The sequence mining approach used in the proposed system deals with the discovery of temporal associations between pairs of event types in an event sequence. The temporal aspect of these associations is that one of the event types happens after the other within a certain time frame. These correlations are supported by statistical measures, which denote the frequency of occurrence of these associations within the sequence. A real life example of such a temporal association is like the following: *"On 77% of the times the IBM stock decreases by 3.5%, the increase of Yahoo's stock by 2.5% follows within 1 week"*. A formal description of the problem of finding such temporal associations between pairs of event types in event sequences follows:

Let there be an event sequence S as defined in the previous paragraph, a user-defined minimum frequency threshold *min_freq*, and a maximum temporal difference boundary $\Delta T_{\max}$, also defined by the user. Assuming that there is a finite set E of different event types that are contained in the event sequence S, the proposed approach discovers frequent patterns that belong to the pattern class shown below:

$$X \rightarrow Y\ [\Delta T_S, \Delta T_E] \tag{1}$$

Where X, Y are event types that belong to E and $[\Delta T_S, \Delta T_E]$ is a relative time frame, within which Y happens after X with a frequency ≥ *min_freq*. ΔT_S, ΔT_E are positive integers with $\Delta T_S \leq \Delta T_E \leq \Delta T_{\max}$.

As in [8], [10] we are not interested in an 'absolute' frequency, but rather in a frequency relative to the LHS (Left-Hand Side) event type of the aforementioned pattern class. Therefore, the frequency *f* of a pattern is defined as the number of times the pattern occurs for a different occurrence of X, divided by the total number of times X occurs in the sequence. Note also that all the occurrences of the pattern using the same occurrence of X are counted as one.

4 The Prototype

4.1 Architecture

As a basis for an intelligent stock market assistant following the above requirements, we propose the architecture presented in Fig. 1. In this section we will present the components of the prototype.

The update component consists of a set of agents who communicate with the web sources to retrieve data and store them locally. The data format is not the same for all the sources so there is the need for dedicated, to the sources, agents. Obviously each agent should check if the data have been already updated by another source so that data consistency is guaranteed. The agents can be scheduled to update automatically the data every business day. Finally, each agent will be responsible for the cleaning and preparation of the data before their storing in the main database of the system.

The ETL (Extract-Transform-Load) component consists of a set of tools that are responsible for the preparation of the data before a data mining algorithm uses them. Of course these tools depend on the specific algorithms hence each algorithm needs its own tool. However, the first version of this tool will use only one temporal data mining algorithm.

The sequence mining algorithm that was used (data mining engine component) needs a specific input in order to run. The ETL component should prepare a table in the database with just two fields: Event description, Event timestamp. The Event description represents the description of the event, i.e. the event type that occurred on the Event timestamp. The form of the results produced by the algorithm is: *LHS Event Type* $\rightarrow$ *RHS Event Type* $[\Delta T_S, \Delta T_E]$*, Frequency.*

The evaluation component is the intelligent part of the whole system. It takes the set of rules produced by the sequence mining algorithm and it tries to understand the meaning of these. Specifically, it tries to adapt the results to the user's preferences and needs and to provide suggestions. A major task of this component is to decide whether a produced rule (pattern) strengthens or not an already stored one. The tool includes an algorithm which searches for equal patterns (common LHS and RHS) and decides to update or not the one that is already stored. In fact, there are three parameters that should be tested: ΔT_S, ΔT_E and the frequency value.

The system includes a pattern warehouse where the patterns produced by the evaluation component are stored. Each pattern consists of many events. An event can be related with a stock transaction or with a different type of event that the user can introduce to the warehouse. The pattern warehouse already includes some standard events types: the volume of a stock, the open/close fluctuation and the high/low fluctuation. Obviously, these types are related with a stock and this has been represented in the pattern warehouse. Fig. 2 represents the pattern warehouse schema.

We have used the notation that is usually used in the data warehouse representation.

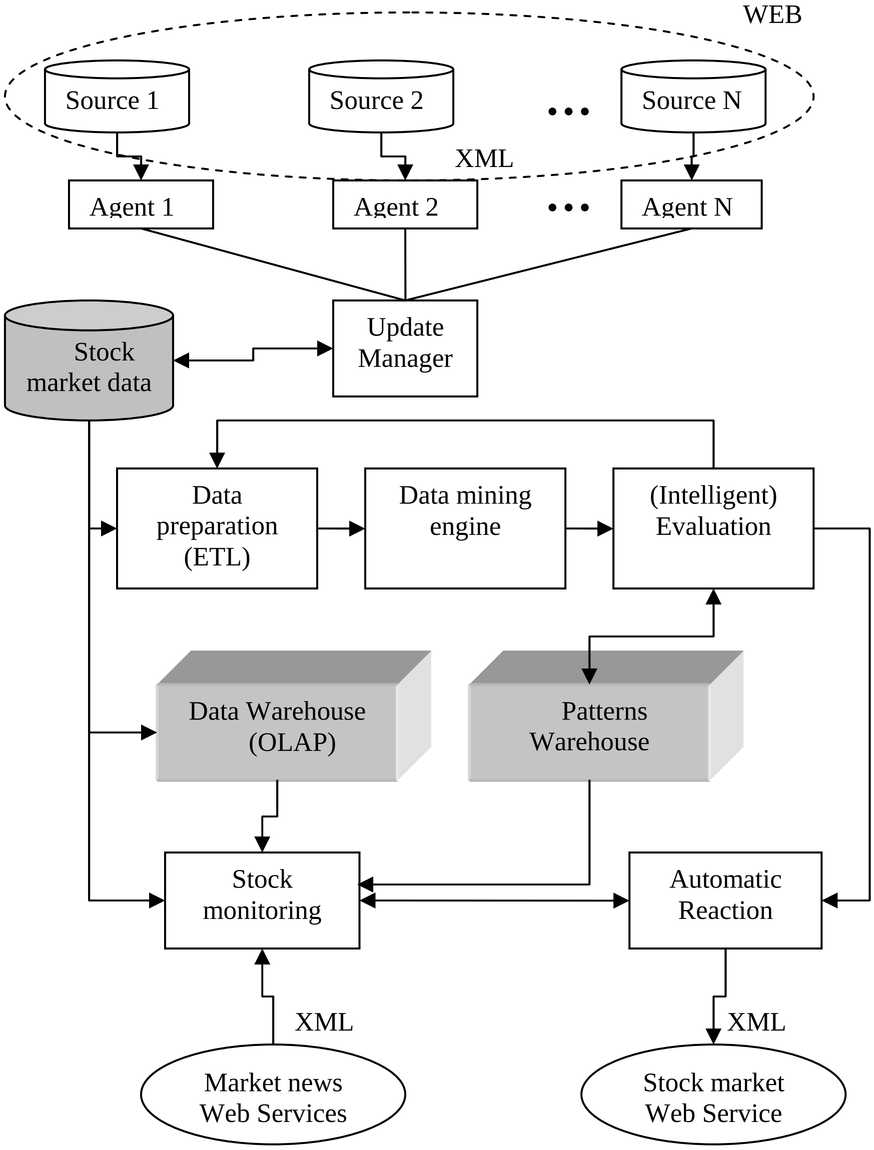

Fig. 1. Architecture of the System

Except of a pattern warehouse, the system could provide data warehousing capabilities. Data warehouses are used to support the integration of many, distributed data sources and the application of OLAP technology. Decision making needs aggregated, statistical data and not raw data that are stored and used only for operational purposes.

The “automatic reaction” component has been included for a more complete view of the system. It could be a future characteristic which will be based on alerts and triggers that will act when some predefined events occur.

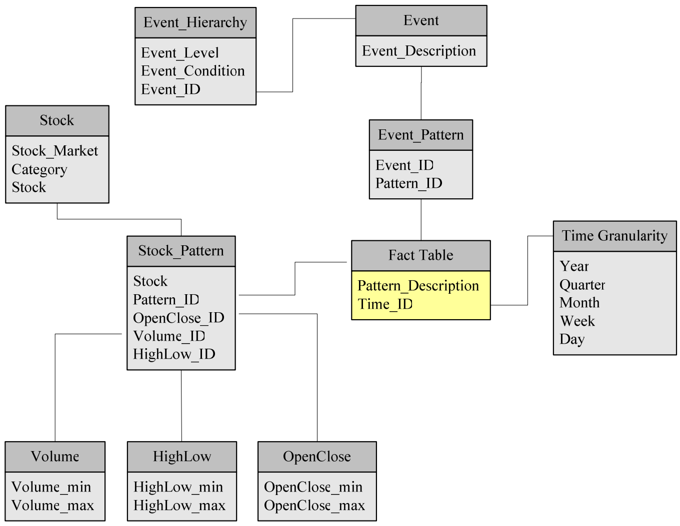

Fig. 2. Pattern Warehouse Schema

4.2 Functionality

This section presents the current main capabilities of the proposed prototype. There is a main form (see Fig. 3) where the user can see market news, stock prices (imported from Yahoo Finance Web Service [12]) and perform a number of operations using either the menu bar or the tabs.

The user can create portfolios that contain stocks belonging to many different business sectors of the stock market. Each portfolio is related with monthly, weekly or daily prices and a risk is specified for it. Users can select to view the stocks per portfolio, per category (sector) of a portfolio or per category (sector). In any case, the system will produce a chart that will show graphically the row data which are also provided.

What-if scenarios analysis is an important part of modern decision support tools. It provides the user with the capability to create custom scenarios and to receive answers to their questions. In the case of our proposed system, the user can create scenarios that include the stock prices behavior and other external events. An example is “What will happen when the price of oil reaches the season highest?”.

The user can select to include some standard events (close/open fluctuation level, high/low fluctuation level and volume level) or some external events (elections, oil prices). The formulated SQL statement is executed against the patterns warehouse which includes the sequences of events produced by the execution of the main sequence mining algorithm. After the execution of the SQL statement the results are presented in a user friendly tree grouped by the different stocks (or other event types). Therefore, a user can ask what will happen when event type A occurs and receive categorized answers.

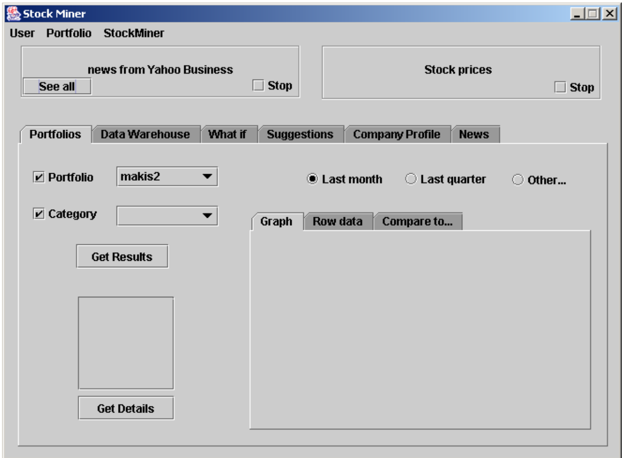

Fig. 3. The main form of the prototype

The proposed tool is an intelligent stock market assistant so it should provide capabilities of suggesting to buy or to sell stocks. Furthermore, alarms are used to remind users of these suggestions. Three major factors are taken into consideration; the risk level, the buying price and the need for cash. When the user asks for suggestions then the stored patterns are examined and these that are according to user preferences are presented. Fig. 4 presents the interface of this capability.

If the user needs cash immediately then the system looks for rules that determine the fall of some stocks. The investors can sell this stock before it falls so that they earn some more money. Of course the buying price should be taken into consideration and this is the reason why the investor is informed about the profit or the loss. In general suggestions follow the format: *Action* + *Stock* +*" because it will go"* + *Fluctuation* + *"in"* + $[\Delta T_S, \Delta T_E]$.

The most intelligent capability of this system is to protect the investors from taking risky decisions. A mechanism is used to examine rules such as "When IBM goes down 3% (event type A) then Oracle goes up 5% (event type B) in 0-20 days". The obvious behavior of the tool would be to suggest to the users to buy Oracle stocks so that when B happens they sell the Oracle stocks and they gain profit. However, during the period 0-20, the stock of Oracle could have lost 30% after A happened and then earned 5% (so B happens). Obviously, this does not mean that the investors make profit. The system checks the historical prices and presents the profit level to the user.

The last two tabs provide users with some additional capabilities. In the company profile tab, users can select a stock and view all the other that affect (or are affected by) the stock price of this one. In the news tab, users can see market news from a number of sources.

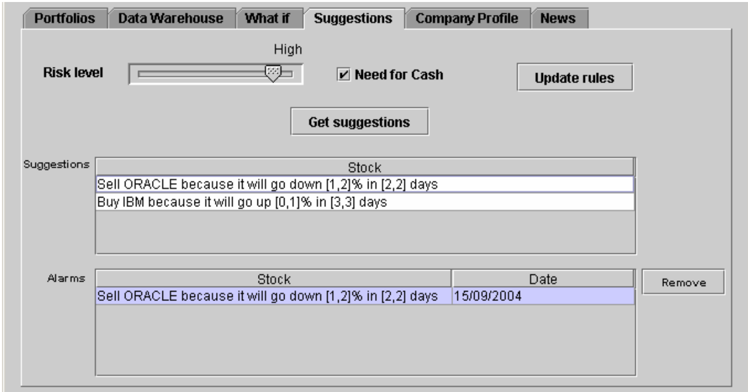

Fig. 4. Suggestions

5 Related Work

Surprisingly, little work has been done on applying sequence mining techniques on the stock market case study. However, there is a lot of interest in the database community in time series data mining. A clustering algorithm is used in [13], in order to discover temporal patterns in financial data. The presented approach is concerned with the analysis of the impact of trade-specific and market-specific features on trading styles in the T-bond futures market. The data analysis dealt with the values of trade profit and the time until expiration.

The Worcester Polytechnic Institute runs a project [14] exploring temporal associations in the stock market. The system used was a combination of the WPI WEKA data mining system as well as an event identification pre-processing module implemented as an extension of an existing algorithm.

6 Conclusions - Further Work

In this paper, we have proposed an Intelligent Stock Market Assistant after having investigated the area of sequence mining. The current version of the tool integrates a sequence mining algorithm and has a pre-processing and a pattern evaluation part.

Future work could include the enrichment of the current tool and its expansion with a component that will combine data mining and technical analysis capabilities. The core intention of a successful investor is to catch trends in their early beginning or to technically capture it when it is still in progress. The aim is not to buy cheap stocks but these that present an upward tendency. After a medium term interval and when the stocks that were bought start to present a downward tendency, the investor sells the stocks and earns profit. Technical analysis can provide the user with answers about these tendencies. The answers are hidden on the charts; this is the philosophy of technical analysis. In fact, technical analysis focuses on the chart of a stock and does not try to relate one stock with some others in order to discover some common or correlated behavior.

Traditional technical analysis can be injected with techniques and tools from the area of data mining. The upwards/ downward tendencies can be considered as event types and be combined with others (e.g. announcements about dividends of shares, Central Banks announcements). Investors need to know not only that a stock will go up but also how much will change its price and when the tendency will change.

References

1. Adriaans, P., Zantinge, D.: Data Mining. Addison Wesley Longman, Harlow, England (1996) 1-10
2. Lin, W., Orgun, M., Williams, G.: An Overview of Temporal Data Mining. Proc. of the Australasian Data Mining Workshop, ADM02, Sydney, Australia (2002) 83-90
3. Das, G., Gunopulos, D., Mannila, H.: Finding Similar Time Series. Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining (KDD ’97), Newport Beach, California, USA (1997) 88-100
4. Agrawal, R., Faloutsos, C., Swami, A.: Efficient Similarity Search in Sequence Databases. Proc. 4th Int. Conf. Foundations of Data Organizations and Algorithms (FODO ’93), Chicago, IL USA (1993) 69-84
5. Chan, K.-P., Fu, A.W.-C.: Efficient Time Series Matching by Wavelets. Proc. 15th Int. Conf. Data Engineering, Sydney, Australia (1999) 126-133
6. Agrawal, R., Srikant, R.: Mining sequential patterns. Proc. 11th Int. Conf. Data Engineering (ICDE ’95), Taipei, Taiwan (1995) 3-14
7. Mannila, H., Toivonen, H., Verkano, I.: Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, Vol. 1:3 (1997) 259-289
8. Bettini, C., Wang, X.S., Jajodia, S., Lin, J.L.: Discovering Temporal Relationships with Multiple Granularities in Time Sequences. Technical Report, George Mason University (1996)
9. Pediaditakis, K.: A Temporal Data Mining Approach for the Extraction of Optimal Temporal Constraints of Sequential Patterns on Event Sequences. MPhil Thesis, School of Informatics, Faculty of Humanities, University of Manchester (2005)
10. Bettini, C., Wang, X.S., Jajodia, S.: Mining temporal relationships with multiple granularities in time sequences. Data Engineering Bulletin (1998) 21:32-38
11. Manilla, H., Ronkainen, P.: Similarity of Event Sequences. Proc. 4th Int. Workshop Temporal Representation and Reasoning (TIME ’97), Daytona Beach, FL USA (1997) 136-139
12. Yahoo! Finance Web site: http://finance.yahoo.com
13. Weigend, A., Chen, F., Figlewski, S., Waterhouse, S.R.: Discovering Technical Trades in the T-Bond Futures Market. Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD ’98), New York, NY USA (1998) 354-358
14. Worcester Polytechnic Institute, Knowledge Discovery and Data Mining Research Group: Exploring temporal associations in the stock markets. http://www.cs.wpi.edu/~ruiz/KDDRG/sequence_mining.html