Predicting Stock Trends with Time Series Data Mining and Web Content Mining

Gil Rachlin\(^1\) and Mark Last\(^1\)

\(^1\) Ben Gurion University of the Negev, Department of Information Systems Engineering, Beer-Sheva 84105, Israel, \{gilrach, mlast\} @ bgu.ac.il

**Summary.** This paper presents a new methodology for predicting stock trends and making trading decisions based on the combination of Data Mining and Web Content Mining techniques. While research in both areas is quite extensive, inference from time series stock data and time-stamped news stories collected from the World Wide Web require further exploration. Our prediction models are based on the content of time-stamped web documents in addition to traditional Numerical Time Series Data. The stock trading system based on the proposed methodology (ADimiral) will be simulated and evaluated on real-world series of news stories and stocks data using several known classification algorithms. The main performance measures will be the prediction accuracy of the induced models and, more importantly, the profitability of the investments made by using system recommendations based on these predictions.

**Key words:** Financial Intelligence, Data Mining, Web Content Mining, Text Mining, Classification, time-series analysis, decision trees, Efficient Market Hypothesis.

1 Introduction

The Efficient Market Hypothesis (EMH), as stated by Fama ([4, 5]), asserts that ‘Stock Prices fully reflect all their relevant information at any given point in time’. As the basis for growth and development of a modern economy this means that no information or analysis can be expected to result in out performance of the market and that stock prices follow ‘Random Walks’ ([1]), where a change in stock price over time is purely random and statistically independent of the stock price in the past. However, to this day no one can explain the anomalies in the market, which can be used to assure some short term predictive power ([1, 7, 9]).

In making their own forecasts most financial specialists try to exploit the time gap of the market’s adjustment to new information. They reduce their risk by combining both technical (base future price predictions on past prices) and fundamental (base predictions on real economy factors, such as inflation,
trading volume, organizational changes in the company etc.) analysis strategies, which were mentioned by Gidófalvi ([6]) and are fully explained in the literature by [1]. In order to obtain the data required for both strategies one can refer to various publicly available resources like the stock market itself, the companies, news papers or others.

A rather new source for information in the late 20th and the 21st centuries is, of course, the Internet. In order to exploit this relatively new media as an additional helping tool in our forecasting task, we need to combine techniques from both time series data mining and web content mining.

Most studies ([8, 11, 12]) agree that the process of Knowledge Discovery in Databases (KDD) involves iterating over four general steps, each using independent tools: 1) Data cleaning and preprocessing (create a common data representation from different sources and different data types, e.g: relational, transactional and spatial databases to large repositories of unstructured data such as the World Wide Web), 2) Test and discover relationships in the data using Artificial Intelligence and Statistical Analysis tools, 3) Postprocessing of discovered patterns, 4) Use the model to perform actions on real world data.

When adding the aspect of time to the Data Mining process, it is understood ([12]) that database records are time stamped and meaningful only as part of a time segment or time series.

In [12], Last et al. use a signal processing technique to pre-process the raw time series data. Then they construct an information theoretic connectionist network (IFN) to induce time series prediction rules, which are later reduced using fuzzification and aggregation. Finally, the rules are presented in natural language and used to predict future behavior of the time series. Their purpose is to predict the timing of Change Points. A Change Point is the point where a specific trend of values in the data is changed, for example, the time point from which a stock price starts to steadily increase after a steady decrease. Last et al. mention two common methods for finding change points: one is by recursive binary partitioning of the time segment, where likelihood criteria are used as the underlying model for each segment. A second method to finding the optimal number of linear segments in time series starts with large number of equal size segments (with 3 time points each), and proceed by merging two adjacent segments, which minimize the balance of error, which is calculated as a standard deviation of errors in all the segments.

Applying Data Mining techniques on the World Wide Web is usually referred to as Web Mining. It introduces new challenges to the Data Mining process on how to clean, categorize and utilize the information in order to create useful models, which can help in the future decision making process. Studies ([2, 10]) divide Web Mining into discovering intra-document structure (Web Content Mining) and to discovering inter-document structure (Web Usage Mining and Web Structure Mining). Web Content Mining is more relevant to our prediction task. It refers to finding relevant information on the Web, mainly, by using Information Extraction (IE) and Information Retrieval (IR) techniques for semi structured data. IE is viewed as trying to find the struc-