TJHSST Computer Systems Lab Senior Research Project "Research into Modelling of Complex Systems" 2006-2007

John Sherwood

October 31, 2006

Abstract

My project is involved with using data mining techniques on the internet in order to gather enough information for the use of a genetic algorithm in trend analysis of a complex system; e.g., the stock market.

1 Introduction - Elaboration on the problem statement, purpose, and project scope

1.1 Scope of Study

The most fundamental element of my program is creating a correlation between news about a company and its stock and the price of the stock itself. In order to do this, a huge amount of data on both stock prices and news regarding companies must be processed into a quantitative format, and then extensively analyzed.

My specific goal is to develop a program that can generate correlations and then, by analyzing recent events, extend those correlations into the future in order to extrapolate the future price of stocks.

1.2 Expected results

In doing this project, I expect to at the very least have a very useful genetic algorithm, that given a list of independant and dependant data, can generate

equations to create a tentative correlation. While the extremely chaotic nature of the specific application may prevent quantitative success in this instance, I do expect to have success on general terms.

1.3 Type of research

My project would most likely be categorized into use inspired basic research, as it has a very specific intent but is also searching for a fundamental understanding of fluctuations in the market.

2 Background and review of current literature and research

The background research I have done for my project has largely been reading on data mining techniques, as well as genetic algorithms, and, specific to my project itself, past (failed) attempts to model the stock market. By how lucrative a stock market predictor would be, there have been many attempts that have taken many different approaches, and I believe that internet data mining and genetic refinement of equations should proove to be the most useful.

3 Procedures and Methodology

My procedure is currently creating the necessary tools for my project, which includes an XML parsing algorithm in order to assist in the data mining, a data analysis tool in order to break the qualitative mined data into quantitative data usable by the predictor, and then of course the genetic algorithm to refine the equations I intend to use in my modeling program. The XML parser has a very specific definition of success; namely, if it can correctly parse XML/HTML documents, and the data analysis tool and algorithmic refiner will be evaluated by how accurate the end up being in prediction.

In terms of visuals, I intend to have, by the end of 2nd quarter, charts of dates and percent accuracy in predicting stock market flux over a period of time, hopefully increasing as time goes on.

In order to test my program, I intend to routinely run the algorithms to on various stocks and test how the data generated by the program compares

with the actual fluctuations in the market itself.

4 Expected Results

After I am finished with my project, I expect to have at the very least a quite extensive level of experience with regards to data mining as well as genetic algorithm refinement, and hopefully a stock market analyzer that will be accurate within a percentage point a day in advance.