Halley’s comet

Science, Astrology

As the Internet of things unveils in the 21st century, almost everything seems to be connected and the world as a whole seems to be shrinking to a miniature chip with countlessly enormous new opportunities paralleled with an even more challenges related to privacy, and other Quality of service (QoS) challenges making it kind of mixed blessing. But in my current research I will focus on the blessing side of the opportunities inspired by the Internet of Things. The QoS challenges will be deferred to be part of my life-time research focus following my successful completion of my Doctoral Research.

As a consequence of the global economic down turn that hit and crippled many companies around the globe, a couple of years ago, the need for an efficient and more accurate Stock market prediction system has become a hot global issue that calls for the contribution and involvement of many scholars to achieve a reliable stock market prediction system which is multi-disciplinary in nature. Thus having such a reliable prediction system can save or at least alert the global companies of possible dangers of committing in non-profitable investments.

With the advent of online trading the application of algorithmic trading scripts to achieve a profitable instant sale or buy transaction has become a popular trading trend. But such instant or short time-interval trading algorithms are not suitable for big investors such us mining, manufacturing, agricultural, or any other companies that invest on a particular project whose profitability need to be predicted reliability and accurately. Such systems require a reliable and accurate prediction system that involves the prediction

of a stock market corresponding to longer periods ranging from a day to a couple of years.

Thus there arises a demand from such big investors for a reliable stock market prediction solution that accurately predicts what the future holds as regards to the rise or drop of the price of a particular stock market. A typical gold mining company, for instance, may require a reasonable prediction system that answers queries like: " Will the price for gold drop or rise tomorrow, a week later, a month later, or even a couple of years later?".

Thus predicting the price of a stock market at varying depths of the future is of paramount importance to company decision makers in general and to business decision makers in particular allowing them to foresee a preview or future snapshot of whether their investments will turn to be profitable or end-up in bankruptcy as severe as the one hit the world a couple of years ago and whose consequences have crippled many global companies and which has not yet been cured.

Thus business decision makers can benefit a lot from stock market prediction systems so as to mitigate the risks of possible business loses and save their company from going bankruptcy. Thus the need for stock market prediction systems is a global concern that deserves more cross-disciplinary scientific researches.

Now that I have clarified the need for such a reliable stock market movement prediction systems, I propose a novel approach for the design and development of a reliable stock market movement prediction system that

utilizes Multiple Machine Learning algorithms powered by a set of publicly available user-generated data corpus streamed from multiple Social Networks such as Twitter, Facebook, or other information services likeGoogleTrends. I propose to predict stock market movements relative to the more reliable Dow Jones Industrial Average (DJIA) market index with the help of a daily generated, rich supply of publicly available user- data streams from Social Networks.

To achieve this novel approach of stock market movement prediction every user-data corpus streamed from social networks will pass through a chain of advanced preprocessing stages to reduce the data corpus to a set of minimal useful user-data relevant for stock market movement prediction. Thus during the preprocessing stage user data paraphrasing and summarization will be used to make sense of whether the user data have some information about a stock market.

The Irrelevant user data will be filtered out and discarded saving storage and processing resource. As a result of this pre-processing we get a set of relevant user-data resulting in a minimal nominated user-dataset. The nominated user-data will then be subjected to a high level feature extraction process to achieve features that are inclusive, and which will result in higher-performance and higher prediction accuracy.

The paraphrasing and summarization done against the original user-data corpus, during the first stage of pre-processing, will play a role in achieving higher-quality and higher-level feature set for the nominated user-data. Thus following a feature extraction the resulting feature values or vectors will be

applied to machine learning algorithms such as SVM, Neural Networks, or Recursive Decision Trees which are trained to figure out or predict the future movement status of a stock market by computing the correlation value that exists between social Network user-data and Dow Jones Industrial Average Index.

Introduction

There are enormous ways the Internet of Things can be exploited to generate new knowledge or information which can be used by other users or systems for making informed and reliable decision making. Typical instances of such systems which can benefit from the Internet of Things are Investment companies. Such companies can make use of the huge user-generated data to predict their loss or profit related to a particular investment they make. Based on such prediction systems an investment company makes an informed management decision so that it wouldn't eventually end up in bankruptcy.

Machine Learning

It is always my habit to go back to mychildhood experiencewhenever I happen to come across a thread on Machine Learning either when discussing a topic with my students or when watching webcasts of the main players of this field of Machine Learning like Prof. Andrew Ng (of Stanford, now Baidu, and cofounder of Coursera) whose work has inspired me to have a new perspective or approach to solving almost any of the once complex and non-trivial real world problems.

But when I say this I don't disregard that Machine Learning is not a one fits all solution but at least it goes well with solving problems that would otherwise have no chance of being solved and even if solved would have been inefficient and computationally intensive demanding huge processing power and working memory. Thus referring back to my childhood days I try to figure out how a machine can learn to understand itsenvironmentand interact with it.

Thus as I said above the solution for many real world problems is embedded on the problems themselves. But solving these problems needs a new way of thinking and that's searching for patterns of solutions within the problem itself. Therefore in this research I will make an intensive use of Machine Learning Algorithms to achieve a reliable stock market prediction rate. Thus I am going to use SVM, Neural-Net, and Recursive Decision Trees and then compare and contrast the prediction results achieved by each of them so as to use the most reliable ML-algorithm fit for such application domains

Role of Social Media to stock market movement prediction

Nowadays, with so manysocial mediasites hosted and serving many virtual users according to their preferences has really created a Virtual Cyber community that outnumbers our real communities. Most of the people in the world have multiple accounts in different social media sites like Facebook, Twitter, LinkedIn, etc.

This online presence or virtual community presence gives rise to a new opportunity for solving the stock market movement prediction problem. Because every social media user has virtual friends stranded all around the

globe who communicate with each other regarding the price of a typical stock item, and their sentiments of that particular item.

So globally there will be a lot of threads that focus on stock item prices and accompanying personal sentiments which can be used to forecast how the stock market movement for particular items will get affected the next day, the next month, the next year, or couple of years later.

Motivationfor carrying out this research

I have three motivations that will keep me enthusiastic and energized throughout the successful completion Doctoral Research. Among these motivations are:

- It will act as a typicalcase studyfor understanding and applying the principles and practices of Intelligent Informatics and Distributed Computing. This will help me solve a myriad of none-trivial problems that will benefit the society which will play a great role in making the Internet of Things a real blessing to the current and the coming generations.
- I believethat the problem is solvable and deserves to be solved because I think that this solution will save and keep alert many global Investment companies from getting bankruptcy.
- Solving this problem would be a great contribution to the global body of knowledge especially for the field of Intelligent Informatics.
- The Multi-disciplinary nature of the research topic will expose me to an array of different fields of studies such as effective use of Machine

Learning toolsets, Social media engineering, Programming, databases and networks, and last but not least Business report analysis.

- Finally this research will alleviate my passion of being a good researcher in the Field of Intelligent Informatics and will reflect and contribute my expertise to bothacademicand industrial sectors.

Problem statement

Stock market movement prediction involves the use of previously generated stock market movement history in DJIA market index or any other reliable market indices in addition to the exploitation of daily generated but publicly available big data corpus produced by multiple Social Networks such as Twitter, Facebook, and other suitablesocial networks. Social network user-generated contents can be mined or forked out for possible patterns of future stock market movements which affect the daily, weekly or even monthly market movement.

The challenges associated with stock market movement prediction systems are overwhelmingly high but the solutions to such challenges are embedded on the challenges themselves. Therefore the connectedness of users to the Internet in general and to the social media in particular along with the soon to unveil, Internet of Things, the challenges will soon fadeout in magnitude. Thus the main challenges that I will address in my research are:

- Nominating potential user-data for prediction
- Optimal user-data feature selection
- Multi-lingual user-data handling to address user data in languages other than English

- Use of Multiple social networks to have a global sense of stock market movement
- Optimal configuration or setup of Machine Learning algorithms to achieve adequate modeling or learning of the social media user-data and the subsequent correlation process with the preferred market index in use.
- Using a realistic Market index that has been effective for many decades.

Proposed system

The novel approach for the design and development of a stock market movement prediction system will address the challenges facing the current stock market movement prediction systems using dedicated modules listed below:

1. Multi-Social Network public data capture
2. User-Content Nomination
3. Semantic Analysis
4. Feature Extraction
5. Training
6. Prediction

Proposed System Overview diagram

In this section I have outlined the general working principles of the proposed system block diagrams. There are two separate figures with Figure-1 depicting a high level context diagram showing the user-content sources originating from multiple social Networks such as Twitter, Facebook, Google

Trends, etc. It also shows the prediction system generating a report of the stock movement prediction for any day in the future such as tomorrow, next week, next month, or even next year depending on the user's requirements.

For Example a Gold mining company CEO may need to know in advance whether to commit to a mining contract in a particular gold mining site that takes more than a year's investment before the actual production and shipment to market begins. Thus the CEO needs options for predicting what will happen to the price of gold one year later by which his company starts producing and shipping Gold to Market. Thus the system should provide a range of options suiting the demands of different customers.

Figure-2 depicts a little bit of detail in terms of the processes or modules involved to realize thegoalsof the system. Therefore there are six-core modules abstracting further implementation and design details.

3. 2 Proposed System component wise Description

In this section I provide an overview of how the proposed system is organized to achieve its goals of stock market movement prediction.

Multi-Social Network public data capture

This module is concerned with the capturing of user-content that will act as the basic raw material based on which the prediction system will predict the future stock movements. Therefore I propose to use Twitter, Facebook, Google Trends, and other user-content serving web-services as I see them fit for my purpose to realize the success of my research. Using Multiple Social Networks has the advantage of having global coverage of user-contents related to a particular stock item.

This will enable the system to achieve an absolute global prediction instead of giving local prediction influenced by small number of virtual communities. For example if we happen to use Twitter user-contents only then we lose the large user base in Facebook, or any other social site that is common to a particular community.

User-Content Nomination

This module is concerned with the first hand data-corpus pre-processing. This Module will take care of the burden of filtering out user data that have no significance or relevance in relation to the stock market. The Task of pre-processing will be split in to two sub-modules each of which can performed their assigned specific tasks.

a. Basic Garbage Data Filter: this sub module will filter out user-contents based on regular expressions patterns. Garbage data involves encrypted content, or any data generated from social media threads which have nothing to do about stock items.

b. User-Content Language Normalizer: This sub module handles the task of detecting the language used to present the current user-content. If it is not an English content the sub-module would use Google Translate web service API to translate the content.

Semantic Analysis

This module needs to apply some degree of intelligence to make sense of the actual meaning of the user content so that it can represent the user-content in a simpler format by introducing the concept of user-content paraphrasing and summarization. Therefore this module processes a

relatively bulky user-content and generates relatively compact user content without affecting the actual meaning of the original user-content. This module makes it easier for the subsequent modules of the stock market movement prediction system to process the content.

a) Paraphrasing: the nominated user-content will be paraphrased or re-worded to build a standard set of phrases to represent user-content so that the bulky and yet worst unstructured user-content will be optimally minimized and restructured but of course without losing the original meaning of the user-content. Rewording has the advantage of minimizing occurrences of ambiguous words which threatens prediction systems.

b) Summarization: this sub module requires more intelligence in order to makes sense of the meaning of the user content and then produce a digest or summary of the user-content so as to achieve more compact and more clear idea of what the user-content is saying about a particular stock market item.

Feature Extraction

The challenge in this module is the selection of the best or optimal feature of a user-content which will be used for modeling the user content. The better the feature sets selected for user-content, the better the Machine Learning algorithm will learn or model the user-content.

Therefore this module will require more efforts in selecting high quality features that result in better modeling of user-content and less processing overhead to the Machine Learning algorithm in use. Therefore the Semantic analysis described above will help this Feature Extraction Module a lot in

selecting a higher level feature sets instead of the cumbersome word-by-word count based feature extraction system used in existing stock market movement prediction systems.

Training

The Training Module is all about modeling the Feature vectors using machine learning algorithm. Thus, this module relays feature vectors corresponding to a user-content obtained from the Feature Extraction Module to a particular Machine Learning (ML) algorithm. The ML-Algorithm will process the feature vectors and updates its knowledge base to reflect the effect of the current feature vector to the already learned ones. Thus, once the ML-Algorithm is trained with a richer supply of user content feature vectors, the system will be ready to predict the stock market movement.

Prediction

This module is concerned with the actual prediction of the stock market movement withrespectto a particular credible market index such as Don Jones Industrial Average (DJIA) market index. Therefore this module will use the now learned and hence expert ML-Algorithm trained above in the Training module to predict the stock market movement of tomorrow or any time in the future by correlating today's user-content obtained from social media users with the near real-time Dow Jones Industrial Average.

Conclusion

Finally, stock market prediction is a hot research topic that requires an advanced knowledge and information engineering. The design and implementation of a reliable and accurate stock market prediction system is

inspired by the large-scale connectedness of users to the Internet in general and to social media in particular.

With the unfolding of the Internet of Things, more devices will be plugged into the Internet with more valuable data flooding into the Internet from almost any device or object. As a consequence of this user-data influx, more valuable marketing data will be available on the Internet inspiring the solution to many non-trivial real world problems based on the principles of Intelligent Informatics.

Therefore the research topic will expose me to the latest technologies available in the booming field of intelligent informatics and I hope will contribute a lot to this booming field through ahard workthat even extends beyond my Doctoral Research. Therefore this research topic is a little bit multi-disciplinary in nature that involves Intelligent Informatics, distributed computing some basic knowledge of market data analysis.