BIG DATA STREAM ANALYTICS FOR CORRELATED STOCK PRICE MOVEMENT PREDICTION

Wenping Zhang and Raymond Lau
Department of Information Systems
City University of Hong Kong
Hong Kong SAR
wzhang23-c@my.cityu.edu.hk

Chunping Li
School of Software
Tsinghua University,
100084 Beijing, China
cli@tsinghua.edu.cn

Abstract

While much research work has been devoted to big data analytics in the past few years, very few studies about big data stream analytics are conducted and reported in existing Big Data literature. Empowered by recent success on automated mining of dynamic business networks, this paper illustrates the design of a novel framework named BIDSTA that leverages big data stream analytics to predict correlated stock price movements. In particular, a business network-based model, named Energy Cascading Model is designed to predict stock price movements by taking into account the evolving sentiments of firms mined from continuous data stream generated by online social media. The business implication of our research is that business managers can apply our design artifacts to more effectively analyze and predict the potential business performance of targeted firms based on the business intelligence extracted from the big data readily available on the Web.

Keywords: Big Data, Data Stream Analytics, Twitter Sentiments, Stock Volatility.

1. Introduction

In the era of the Social Web, user-contributed contents have become the norm. The amounts of data produced by individuals, business, government, and research agents have been undergoing an explosive growth – a phenomenon known as the data deluge. For individual social networking, many online social networking sites have between 100 and 500 million users. By the end of 2013, Facebook and Twitter had 1.23 and 0.64 billion active users, respectively. The number of friendship edges of Facebook is estimated to be over 100 billion. For scientific research, the Large Hadron Collider (LHC) facility at CERN produced 13 petabytes of data in 2010 alone (Tan et al. 2013). The stream of extremely large volume of data, i.e., big data from online news, social media, search queries, and sensors has called for the research and development of a new generation of methods and tools to effectively process it. Big data is often characterized by three dimensions, named the 3 V’s: Volume, Velocity, and Variety (Boden et al. 2013). Currently, there are two main ways to deal with big data, namely distributed computing and streaming.

Most data originally produced from the Social Web is streaming data. For example, the data

1 http://www.statisticbrain.com/facebook-statistics/
representing actions and interactions among individuals in online social media, or the data denoting some events captured by sensor networks is the typical kind of streaming data. Other types of big data perhaps are just a snapshot view of the streaming data generated from a specific point of time. For big data streams, data arrives at high speed, and the methods of big data stream analytics must process it in one pass and under very strict constraints of space and time. Currently, algorithms of big data analytics often deal with big data in batch model, while algorithms designed to process big data stream in real-time or near real-time are rare. Big data analytics can be classified as distributed or single host approaches. For distributed methods, there are batch mode and streaming mode of processing. Although batch mode big data analytics approach (e.g., MapReduce) is the dominated method to date, online incremental algorithms that can effectively process evolving data stream are desirable to address both the “volume” and the “velocity” issue of big data on the Web. The main contribution of this paper is the design and development of a novel big data stream analytics framework that provides the essential infrastructure to operationalize a business network based inference model incorporating evolving data stream from online social media for correlated stock price movement prediction in near real-time.

Recent empirical studies have revealed that a firm’s relationships with other firms may have direct influence on its strategic competitiveness, and hence its business performance (Bernstein 2003; Lau et al. 2012; Ma et al. 2009). Such an influence may be explained with reference to the structural embeddedness theory (Gnyawali and Madhavan 2001) and Porter’s five forces model (Porter 1980). These theories provide the theoretical ground for the design of a network-based inference model for the analysis and prediction of business performance. For example, the structural embeddedness theory posits that a firm’s specific position in an inter-firm network influences its competitive behavior due to the firm’s unique way of accessing external assets and information through the network (Gnyawali and Madhavan 2001). In the past, researchers believed that stock price movements could be explained according to the random walk theory or the efficient-market hypothesis (Cootner 1964; Fama 1991). Accordingly, stock price movements of a firm are considered random and unpredictable. However, recent research findings reveal that stock price movements may not simply follow a random walk process, and stock volatility is predictable to a certain extent (Chaudhuri and Wu 2003; Lo and MacKinlay 2011).

It is generally believed that stock prices can reveal traders’ expectations about a firm’s business performance and other economic variables (Elton et al 1981; Feldman et al. 1997). According to the extended five forces model (Brandenburger and Nalebuff 1996), a firm’s rival force, supplier force, and complementary force are among the most significant factors that influence its competitiveness, and hence affect its ultimate business success. Accordingly, it is an intuitive assumption that a firm’s relationships (e.g., collaborative supply-chain relationships) with other firms in an inter-firm network can be used as the basis to infer its stock price movements. Another main contribution of this paper is the design of a novel business network-based inference model named Energy Cascading Model (ECM) for the analysis and prediction of firms’ directional stock price movements by leveraging big data stream analytics. The business implication of our research is that business managers and financial analysts can apply our design artifacts to more effectively analyze and predict the business performance of targeted firms in a timely manner. As a result, proactive corporate actions can be taken to streamline the business operations of these targeted firms, and enhance their competitive power. Moreover, business consultancy agencies can apply our design artifacts to better monitor their targeted firms and
extract accurate business intelligence from online social media in near real-time.

2. The Big Data Stream Analytics Framework

An overview of the proposed framework that leverages Big Data Stream Analytics (BIDSTA) for directional stock price movement prediction is depicted in Figure 2. The BIDSTA framework consists of seven layers, namely data stream layer, data pre-processing layer, data mining layer, prediction layer, learning and adaptation layer, presentation layer, and storage layer. For these layers, we will apply well-proven methods and tools for rapid service prototyping. For instance, the open-source Distributed Data Stream Engine (DDSE) for big data, Strom2, is applied to process continuous messages fed from dedicated APIs and crawlers at the Data Stream Layer. For instance, the Topsy API3 is used to retrieve Twitter messages.

The Storage Layer leverages Apache HBase and HDFS for real-time storage and retrieval of big data. The Stanford Dependency Parser and the GATE NER module (Maynard et al. 2001) are applied to build the Data Pre-processing Layer. Our pilot tests show that the size of the multilingual social media data streams is within the range between 0.2 and 0.4 Gigabytes on a daily basis, and this size is steadily growing. At the data mining layer, the Affect Miner utilizes a novel community-based affect intensity measure to predict public’s moods towards stocks. Among the big six classes i.e., anger, fear, happiness, sadness, surprise, and neutral commonly used in affect analysis, we focus on the anger, fear, sadness, and happiness classes relevant for directional stock price movement predictions. The WordNet-Affect lexicon (Valitutti et al. 2004) extended by a statistical learning method is used by the Affect Miner. Since social media messages are generally noisy, one novelty of our framework is that we reduce the noise of the “affect intensity” measure by processing messages generated by each recognized financial investment community (e.g., the famous financial analysts) only.

Dynamic business networks are mined via our Latent Business Relationship Miner (Zhang et al. 2012). The dynamic business networks are mined based on a batch mode learning method. Previous research employed the HMM method to mine the latent “intents” of actors (Zhang et al. 2009). We exploit a novel and more sophisticated online generative model and the corresponding distributed Gibbs sampling algorithm to build our Latent Intent Miner that predicts the intents of the general investors in the stock market. The Sentiment Analyzer utilizes well-known sentiment lexicons such as OpinionFinder4 to predict the polarity of investment related messages. Finally, directional stock price movement prediction is performed based on a novel business network based inference model, named Energy Cascading Model (ECM). The predicted stock price movements given a user’s stock query are routed to the user via the presentation layer. Different modes of presentations (e.g., text, graphics, multimedia on desktops or mobile devices) are supported by our framework.

---

2 http://storm-project.net/
3 http://topsy.com/
4 http://mpqa.cs.pitt.edu/opinionfinder/
In addition, a novel parallel co-evolutionary genetic algorithm (PCGA) is designed so that the proposed prediction model is equipped with a learning and adaptation mechanism that continuously tunes the whole service with respect to possibly changing features of the problem domain. The PCGA can divide a large search space into some subspaces for a parallel and diversified search, which improves both the efficiency and the effectiveness of the heuristic search process. Each subspace (i.e., a sub-population) is hosted by a separate cluster. Three fundamental decisions are involved for the design a genetic algorithm (GA), that is, a fitness function, chromosome encoding, and a procedure that drives the evolution process of chromosomes (Lau et al. 2006). First, the fitness function of our PCGA is developed based on a performance metric (e.g., accuracy of stock price prediction). Second, since various components of the proposed service should be continuously refined, there are multiple sub-populations of chromosomes to be encoded and co-evolved simultaneously. During each evolution cycle, the best chromosome of a sub-population (e.g., prediction features, social media sources, system
parameters) is exchanged with that of other sub-populations. Armed with all the essential information, each chromosome of a sub-population represents a feasible prediction, and its fitness can be assessed accordingly.

3. Discussions and Summary

While some research work has been devoted to big data analytics recently, very few studies about big data stream analytics are reported in the literature. The main theoretical contributions of our research include the design and development of the big data stream analytics framework, named BIDSTA for directional stock price movement prediction. The business implication of our research is that business managers and financial analysts can apply our design artifacts to more effectively analyze and predict the business performance of targeted firms based on automatically mined business networks and dynamic sentiments mined from big data stream. Accordingly, they can take proactive business strategies to streamline the operations of these firms. One limitation of our current work is that the proposed framework has not been tested under an empirical setting. We will devote our future effort to evaluate the effectiveness and efficiency of the BIDSTA framework based on real-world financial data and social media messages collected from the Web.

References