A Highly Efficient Big Data Mining Algorithm Based on Stock Market

Jinfei Yang, School of Economics, Minzu University of China, Beijing, China
Jiajia Li, School of Computer Science, South China Normal University, Guangzhou, China
Qingzhen Xu, School of Computer Science, South China Normal University, Guangzhou China

ABSTRACT

This article proposes a new algorithm which includes two stages. First, the Pearson correlation coefficient is used to calculate the similarity data, and the activity of stock money flow was calculated by combined the probability generating function (P.G.F.) of stationary waiting time and stationary queue length. Second, the discrete time Geo/G/1 queue with a Bernoulli gated service is proposed in calculating money flow by data mining of stock. The new algorithm could calculate data in real time, and each investor could see the real-time data mining graphics. Investors could establish their quantitative trading strategies based on the new money flow model. The proposed algorithm exploits the nature behind stock data. The experimental results show that the authors’ approach can be automatically implemented by the investment strategy and know the future trend of the stock market, as well as the economic development of the region, according to the results of the stock data mining in a certain region.

KEYWORDS

Big Data, Data Mining, Eastern Region, Geo/G/1 Queue, Money Flow, Western Region

1. INTRODUCTION

The traditional method is mainly from several parameters such as geographical location, GDP, and population to study the economic development level of a region. In addition to geographical location, other indicators can be false easily. This undoubtedly increases the difficulty of studying the level of economic development in the region. For example, GDP data in Northeast China is false. This brings us to the second question we studied. The stock market is the most fair in the world, and it is the most stringent regulation. The cost of fraud is very large, which guarantees the authenticity of a regional economic development. This paper attempts to analyze and judge the economic development of a region from the data mining. We have to recognize that a region’s economic development level of people’s income. It will directly determine people’s consumption levels such as the number of stock, stocks’ activity, etc. This paper is to study the economic development level of a region by data mining of their stocks.

A data analytics tool that enables knowledge discovery through information retrieval (i.e., terms) from document-append style storage is designed (Lomotey & Deters, 2015). YU KM et al. proposed multi-core architectures and two high efficiency load balancing parallel data mining methods based

DOI: 10.4018/IJGHPC.2018040102
on the Apriori algorithm (Yu et al., 2015). Jaein Kim et al. presented a novel data mining algorithm called CanTree-GTree (Jaein & Buhyun, 2016). The CanTree-GTree algorithm is improved based on a sliding-window and batch-unit move. Their results show that it can reduce about 35% and 26% of operating run-time cost. The new algorithm wins in many other algorithms. They contribute to the field of real-time stream data mining.

To predict stock return and risks accurately, they proposed a novel approach by three stage method (Sasan & Mohammad, 2015). Firstly, they analyze the most important factors that affect stock returns and risks. Then stocks return and risk can be predicted by data mining technology, as follows:

\[
R = \sqrt{1 + \frac{r_1}{100} \left[ 1 + \frac{r_2}{100} \right] \ldots \left[ 1 + \frac{r_n}{100} \right]}
\]  

(1)

where rn is real return of nth periods. The standard deviation of the stock returns, as follows:

\[
\sigma = \sqrt{\frac{1}{n-1} \sum_{i=0}^{n} (r_i - E(r))^2}
\]  

(2)

Finally, hybrid feature selection algorithm is improved for the Risk and return forecasts. The model of this paper has the advantages of various algorithms, and can accurately predict the stocks of Tehran Stock Exchange (TSE) from 2002 to 2011. Dr. Sasan do made a contribution in predicting stock returns and risks.

A regression model is proposed to study the relationship among gold mining stocks, gold prices and oil prices (Jonathan, & Cetin, 2017) as follows:

\[
RGDX_i = a_0 + a_1 RSPY_i + a_2 RGDL_i + a_3 ROIL_i + \varepsilon_i
\]  

(3)

At last, they find that there is no causality between the gold price and the prices of gold mining stock. The price of crude oil and gold mining shares are positively correlated. Oil stocks fell, and gold mining shares fell.

The representation and conversion method of Bezier surfaces in Multivariate B-Form is proposed (Ruomei & Xiaonan, 2006). Professor Wang presented an efficient numerical method for describing a garment’s mechanical behavior during wear (Ruomei & Yu, 2011). The algorithm simulates vivid dress. They improved the algorithm and improved the research level in the application field.

An exercise thermo physiology comfort prediction model is proposed. They coupled the thermal interactions among human body, clothing, and environment (HCE) (Jia & Yu, 2016). They combined the inner side of the clothing close to the skin and outer side of the clothing exposed to the environment algorithms. A complicated prediction model is improved to a mobile application platform. Each mobile phone user can predict their own thermo physiology comfort analysis, and this was a bold move. They did a very interesting study.

Of course, there is much space to explore. In order to better study of the economic development of a region, this paper summarizes the previous research, which we try to use a regional stock data mining to study the level of economic development.

The specific contributions of this paper include:

- A literature survey about various existing data mining algorithms, and analyze their advantages and disadvantages;
Collaborative e-Learning and ICT Tools to Develop SME Managers: An Italian Case
www.igi-global.com/article/collaborative-learning-ict-tools-develop/46051?camid=4v1a