University of Science and Technology

Faculty of Computer Science and Information Technology

Postgraduate Studies

Master of Information Systems Batch (4)

Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of Information Systems

Implementation of K-Means Algorithm on Web Mining

Prepared By:

Adeel El-Tigani Adeel Osman

Supervised By:

Dr. Atif Ali Mohamed

December 2014

Abstract

Data mining is a one of powerful new technology with great potential to help companies to focus on the most important information that will extract from their data. They have collected about the behavior of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal. The Saudi Arabia stock markets has dynamic huge data needs to be analyzed and categorized in term to be helpful for the customer to make the best decision either to buy stocks or not based on the company's profits. To apply the concept of web mining it was used many software tools; MS-Excel application was used to extract data from the official website of Saudi Arabia stock markets also used Rapid Miner software to import the collected data then to apply the K-Means algorithm to clustering collected data. K-Means algorithm had created three clusters each cluster contains on group of data that had approximate values. The clustered data used to divide the Saudi Arabia's companies into three clusters based on profits (high, medium and low profits). This classify based on K which was passed to K-Means algorithm and their mathematical operations.

المستخلص

التنقيب في البيانات هو واحد من التقنيات الجديدة الفعالة والتي تحتوي على إماكنيات كبيرة تساعد الشركات في التركيز على أهم المعلومات التي يتم إستخلاصها من البيانت الخاصة بهم، والتي لها علاقة بسلوك العملاء الحاليين والمستقبليين، حيث تقوم هذه التقنية بالرد على الإستفسارات ووتزويدهم بالتقار التي لا يمكن الكشف عنها بالطرق التقليدية . أسواق المملكة العربية السعودية للأسهم المالية لديها مجموعة ضخمة من البيانات الديناميكية (المحدثة آنياً) والتي تحتاج للتحليل والتصنيف لكي تساعد العملاء في إتخاذ أفضل القرارات بشراء الأسهم الخاصة بشركة اولا بناءَ على ارباح تلك الشركة. لتطبيق مفهوم التنقيب في الويب تم استخدام مجموعة من الأدوات البرمجية، فقد تم إستخدام برنامج MS-EXCEL لإستخلاص البيانات من الموقع الرسمي لأسواق المملكة العربية السعودية للأسهم، ومن ثم أستخدام برنامج Rapid Miner لتطبيق خوارزمية K-Means على البيانات التي تم إستخلاصها. قامت خوارزمية K-Means بإنشاء ثلاث مجموعات كل مجموعة تحتوي علي مجموعة من البيانات ذات القيم التقريبية. البيانات المصنفة أستخدمت لتقسيم الشركات الى ثلاث مجموعات بناءً على الربح (العالي ، المتوسط ، المنخفض). تم هذا التصنيف إعتماداً على عدد المتغير K الذي تم تمريرها لخوارزمية K-Means والعمليات الرياضية الخاصة بها للتقسيم .

1.1 Introduction

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The term Web Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web mining to utilize that information in their best interest.

Saudi joint stock companies had their beginnings in the mid 1930’s, when the "Arab Automobile” company was established as the first joint stock company. By 1975 there were about 14 public companies. The rapid economic expansion, besides the Saudisation of part of the foreign banks capital in the 1970’s led to the establishment of a number of large corporations and joint stock banks.

The market remained informal, until the early 1980’s when the government embarked on forming a regulated market for trading together with the required systems. In 1984, a Ministerial Committee composed of the Ministry of Finance and National Economy, Ministry of Commerce and Saudi Arabian Monetary Agency (SAMA) was formed to regulate and develop the market. SAMA was the government body charged with regulating and monitoring market activities until the Capital Market Authority (CMA) was established in July 2003 under the Capital Market Law (CML) by Royal Decree No. (M/30). The CMA is the sole regulator and supervisor of the capital market, it issues the required rules and regulations to protect investors and ensure fairness and efficiency in the market. [1]

1.2 Problem Statement

In Saudi Arabia’s stock markets many individual and groups of companies are involved on it with different levels in term of their own Profits and Losses for each which depend on many aspects. The stock markets used variable data that represent the statue for these companies which verify in term of time. There are large amount of data should be analyzed in order to predict the future of stocks price which may indicate to the future of these companies. To use prediction methods on such dynamic data with no arrange this data in certain manner the prediction will be very difficult.

1.3 Main Objectives

The main objectives of this research are:

- Study the theoretical basic principles of Web-Mining.
- Collect dynamic data from web site (Saudi Arabia’s stock markets web site).
- Use Rapid Miner application to arrange the dynamic data into a particular groups or clusters based on the level of profit (using K-means algorithm).

1.4 Sub Objectives

- Use MS-Excel application to extract data from web site (Saudi Arabia’s stock markets web site).
- Use Rapid Miner application to implement the K-means algorithms.

1.5 Methodology Overview

In this research it was used the descriptive analysis to build and apply a mechanism that able to arrange a dynamic data into clusters. MS-Excel application was used to extract the data from the website, then through Rapid miner software we applying K-means algorithm to create clusters each cluster contain on group of data that had approximate values. The K-means algorithm divides the Saudi Arabia’s companies into clusters based on profits.

1.6 Chapters Layout

This research consists of five chapters in addition to the references. **Chapter two**, concerns on the basic concept of Big Data, techniques, basic principles of web mining that are used with it. **Chapter three** concentrates on Methodology, concept of clustering and algorithms which are applied to it. **Chapter four** represents the Implementation of web mining and the explanation of K-means algorithm that used in the Implementation. **Chapter five** represents the results and the discussion of these results and recommendations for future works.