University of Science and Technology

Faculty of Post Graduate Studies and Academic Advancement

Omdurman - Sudan

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

Using Data Mining For Market Basket Analysis

By: Mustafa Alhadi Altaj

Supervisor: Dr. Atif Ali

February - 2014

Abstract

The advent of low-cost data storage technologies and the wide availability of Internet connections have made it easier for individuals and organizations to access large amounts of data. Such data are often heterogeneous in origin, content and representation, as they include commercial, financial and administrative transactions, web navigation paths, emails, texts and hypertexts, and the results of clinical tests, to name just a few examples. Their accessibility opens up promising scenarios and opportunities, and raises an enticing question: is it possible to convert such data into information and knowledge that can then be used by decision makers to aid and improve the governance of enterprises and of public administration?

There are group of algorithms makes indeed the answer of previous question is YES, we can say that business intelligence systems tend to promote a scientific and rational approach to managing enterprises and complex organizations. Even the use of an electronic spreadsheet for assessing the effects on the budget by fluctuations in the discount rate, despite its simplicity, requires a part of information help to make the best decision about financial flows.

This research review some of these algorithms in wide scope and puts enough focus on association rules data mining concepts, specially on Apriori algorithm. Java programming language is used to implement and solve the problem of a store selling accessories for cellular phones as a case study example. And as a result of this research, a new mining tool has been generated specifically for that.

المستخلص

ظهور تقنيات التخزين منخفضة التكلفة والتوفر الواسع لإتصالات شبكة الإنترنت جعل من السهل للأفراد والمنظمات الوصول إلى كمية كبيرة من البيانات. مثل هذه البيانات تكون غير متجانسة في الأصل والمحتوى وطريقة تمثيلها، لأنها تشمل المعاملات التجارية والمالية والإدارية، وعمليات التنقل بالويب، رسائل البريد الإلكتروني والنصوص والإرتباطات التشعبية، ونتائج الإختبارات السريرية، كلها أمثلة قليلة من كثير. تمهد لإمكانية الوصول لسيناريوهات وفرص واعدة، وتثير سؤالا يطرح نفسه: هل من الممكن تحويل مثل هذه البيانات إلى معلومات ومعرفة يمكن استخدامها من قبل صناع القرار للمساعدة والتحسين في تسيير المؤسسات وأمور الإدارة العامة ؟

هنالك مجموعة من الخوارزميات تجعل بالتأكيد الإجابة على السؤال السابق بـ نعم ، يمكننا القول بأن أنظمة ذكاء الأعمال تميل إلى تشجيع النهج العلمي والعقلاني في إدارة المؤسسات والمنظمات المعقدة. حتى إن استخدام الجداول الإلكترونية لتقييم التأثيرات على الميزانية من تقلبات سعر الصرف، على الرغم من بساطته، فإنه يتطلب جزء من المعلومات التي تساعد في اتخاذ أفضل قرار بشأن التدفقات المالية.

هذا البحث يستعرض بعضا من هذه الخوارزميات في نطاق واسع ويضع التركيز الكافي على مفاهيم قواعد العلاقات في تنقيب البيانات، بالتحديد على خوارزمية Apriori. استخدمت لغة جافا البرمجية لتطبيق وحل مشكلة لمتجر مبيعات لإكسسوارات الهواتف الخلوية كمثال لحالة دراسية. وكنتيجة لهذا البحث، تم إيجاد أداة تنقيب جديدة خصيصا لذلك.

1.1 : Introduction

In complex organizations, public or private, decisions are made on a continual basis. Such decisions may be more or less critical, have long- or short-term effects and involve people and roles at various hierarchical levels. The ability of these knowledge workers to make decisions, both as individuals and as a community, is one of the primary factors that influence the performance and competitive strength of a given organization. [1]

Most knowledge workers reach their decisions primarily using easy and intuitive methodologies, which take into account specific elements such as experience, knowledge of the application domain and the available information. This approach leads to a stagnant decision-making style which is inappropriate for the unstable conditions determined by frequent and rapid changes in the economic environment. Indeed, decision-making processes within today's organizations are often too complex and dynamic to be effectively dealt with through an intuitive approach, and require instead a more rigorous attitude based on analytical methodologies and mathematical models. [1]

As a result, business intelligence systems appear with main purpose to provide knowledge workers with tools and methodologies that allow them to make effective and timely decisions.

There are many definitions for BI (Business Intelligence), and in general:

Business intelligence is represents the tools and systems that play a key role in the strategic planning process of the corporation. These systems allow a company to gather, store, access and analyze corporate data to aid in decision-making. [7]

A business intelligence environment offers decision makers information and knowledge derived from data processing, through the application of mathematical models and algorithms which are more dependable. As a result, decision makers are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way.

Enterprises operate in economic environments characterized by growing levels of competition and business domain standards. As a consequence, the ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.

Figure 1.1 illustrates the major benefits that a given organization may draw from the adoption of a business intelligence system. Decision makers ask themselves a series of questions and develop the corresponding analysis. Hence, they examine and compare several options, selecting among them the best decision, given the conditions at hand. [1]

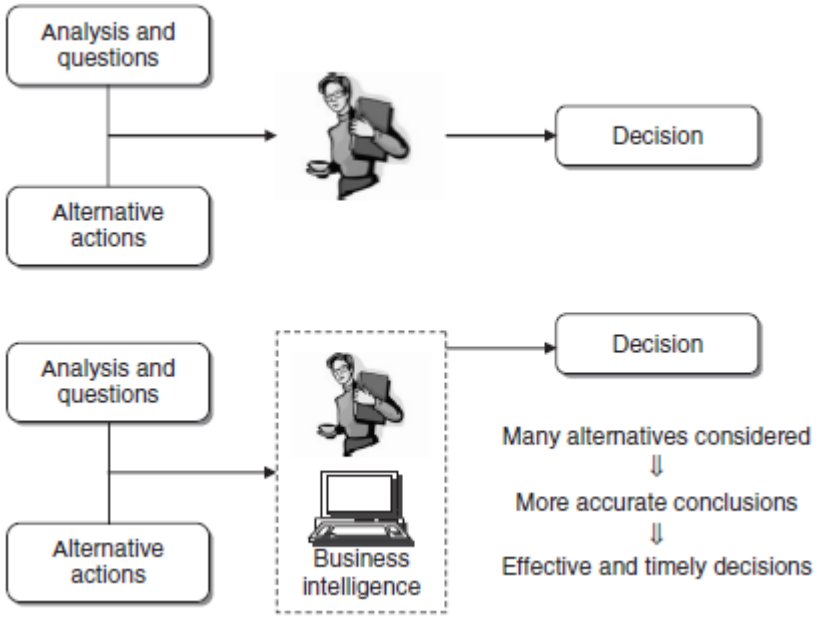

Figure (1.1): Benefits of a business intelligence system [1]

1.2 : Research Problem

In several areas of application, the systematic collection of data gives rise to massive lists of transactions that lend themselves to analysis through association rules mining technique in order to identify possible recurrences in the data, it is fairly simple and intuitive and are frequently used to investigate sales transactions in shopping baskets or navigation paths within websites.

The problem discussed in this research is how to analyze the products jointly purchased by customers, in three words it known as *market basket analysis*.

1.3 : Research Objective

The main objective of this research is to design a tool making the market basket analysis quite useful for marketing managers in planning promotional initiatives or defining the assortment and location of products on the shelves.

Association rules aim to identify which regular patterns and recurrences within a large set of transactions, and for reach this main objective, Apriori algorithm used to be a typical solution for this kind of problems.

There are sub research objectives mentioned below:

- Develop an application (tool) able to run on any operating system environment.
- Develop an effective data structure for memory management in Apriori algorithm which better by speed and performance generally, rather than popular hash-tree structure.
- Availability to showing the final results of stronger association rules in appropriate form for both users groups, detailed form to developers and experts, and simple integrated report form to ordinary end-users.

1.4 : Research Methodology

We used Java which is provided by Oracle Corporation because it is portable language so it can be works on different types of operating systems

and workstations, and we implemented it on the standard steps of Apriori algorithm, taking into account the sub research objectives above.

1.5 : Research Organization

Chapter two reviewed a background of data mining such definition, DM application, DM process and analysis methodologies like classification, association rules and clustering.

Chapter three described the CRISP model in DM and it's six phases in details.

Chapter four discussed the related works of Apriori algorithm, also discovered three different types of improvement on the algorithm to rise up performance and reliability.

Chapter five talked about Apriori algorithm in focus and explain all the processes of it by taking a case study and XLMiner mining tool as examples.

Chapter six explained the reason of selection Java language programming then a general view about the implementation, the results, programming concepts and most important fractions of codes.

Finally, chapter seven contains the conclusions and recommendations