**الآية**

* بسم الله الرحمن الرحيم -

قال تعالى : "لَا يُكَلِّفُ اللَّهُ نَفْسًا إِلَّا وُسْعَهَا لَهَا مَا كَسَبَتْ وَعَلَيْهَا مَا اكْتَسَبَتْ رَبَّنَا لَا تُؤَاخِذْنَا إِنْ نَسِينَا أَوْ أَخْطَأْنَا رَبَّنَا وَلَا تَحْمِلْ عَلَيْنَا إِصْرًا كَمَا حَمَلْتَهُ عَلَى الَّذِينَ مِنْ قَبْلِنَا رَبَّنَا وَلَا تُحَمِّلْنَا مَا لَا طَاقَةَ لَنَا بِهِ ۖ وَاعْفُ عَنَّا وَاغْفِرْ لَنَا وَارْحَمْنَا أَنْتَ مَوْلَانَا فَانْصُرْنَا عَلَى الْقَوْمِ الْكَافِرِينَ"

* صدق الله العظيم -

سورة البقرة آية ﴿٢٨٦﴾

**Acknowledgments**

In such moments, the pen starts wondering what to write, and what to pick from letters, words and phrases, to fully describe his owner’s gratitude and appreciation for everyone made an impression that lighten the path of science and facilitate the obstacles.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Extend my sincere thanks and gratitude to the great man

**" Dr. Atif Ali "**

who oversaw this research and gave me the knowledge, advice and expertise, with a lot of patience until it came out in the image that it is on, may Allah bless and protect him .

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Also, I would like to thank all doctors who taught me in the best way, and gave me a strong boost in my academic career, they have all the thanks and respect.

To my dears colleagues, who I spent great times with them and shared the beautiful experiences, you were really the most sociable and helpful people, thank you my friends.

And of course, I will not forget to thank my parents, brother and sisters, and all my family who stood by me and helped to overcome the hard times of despair and frustration, and grabbed my hand and walked me to the end of the road and the beginning of the future.

Thank you all ...

**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.

**Contents**

|  |  |
| --- | --- |
| الآية **………………………………………………………………………………..** | i |
| Acknowledgments **..…………………………………………………………….** | ii |
| Abstract **………………………………………………………………………….** | iii |
| المستخلص **………………………………………………………………...………..** | iv |
| Contents **………………………………………………………………………...** | v |
| List of Figures **…………………………………………………………………..** | ix |
| List of Tables **……………………………………………………………………** | x |
|  |  |
| **1. Chapter One: Introduction** |  |
| 1.1 : Introduction **……………………………………………………………….** | 2 |
| 1.2 : Research Problem **……………………………………………………….** | 3 |
| 1.3 : Research Objectives **…………………………………………………….** | 3 |
| 1.4 : Research Methodology **…………………………………………………** | 4 |
| 1.5 : Research Organization **………………………………………………….** | 4 |
|  |  |
| **2. Chapter Two: Data Mining Background** |  |
| 2.1 : Definition **……………………………….………………………………….** | 6 |
| 2.2 : Models and Methods for Data Mining **………………………………….** | 6 |
| 2.3 : Data Mining, Classical Statistics and OLAP **…………………………..** | 7 |
| 2.4 : Applications of Data Mining **……………………………………………..** | 7 |
| 2.5 : Representation of Input Data **…………………………………………...** | 8 |
| 2.6 : Data Mining Process **……………………………….……………………** | 10 |
| 2.7 : Analysis Methodologies **……………………………….………………...** | 11 |
| 2.8 : Classification **……………………………….……………………………..** | 14 |
| 2.8.1 : Classification Problems **………………………………………….** | 14 |
| 2.8.2 : Taxonomy of Classification Models **……………………………** | 14 |
| 2.9 : Evaluation of Classification Models **…………………………………….** | 15 |
| 2.9.1 : Holdout Method **…………………………………………………..** | 16 |
| 2.9.2 : Repeated Random Sampling **…………………………………...** | 16 |
| 2.9.3 : Cross-Validation **………………………………………………….** | 17 |
| 2.9.4 : Confusion Matrices **………………………………………………** | 17 |
| 2.9.5 : ROC Curve Charts **……………………………………………….** | 17 |
| 2.9.6 : Cumulative Gain and Lift Charts **……………………………….** | 17 |
| 2.10 : Classification Trees **…………………………………………………….** | 17 |
| 2.10.1 : Splitting Rules **…………………………………………………..** | 19 |
| 2.10.2 : Univariate Splitting Criteria **……………………………………** | 21 |
| 2.10.3 : Development of a Classification Tree **………………………..** | 21 |
| 2.10.4 : Stopping Criteria and Pruning Rules **…………………………** | 21 |
| 2.11 : Bayesian Methods **……………………………………………………...** | 22 |
| 2.11.1 : Naive Bayesian Classifiers **……………………………………** | 22 |
| 2.11.2 : Bayesian Networks **…………………………………………….** | 22 |
| 2.12 : Logistic Regression **…………………………………………………….** | 23 |
| 2.13 : Neural Networks **………………………………………………………..** | 23 |
| 2.13.1 : The Rosenblatt Perceptron **……………………………………** | 23 |
| 2.13.2 : Multi-level Feed-forward Networks **…………………………...** | 23 |
| 2.14 : Support Vector Machines **……………………………………………...** | 24 |
| 2.14.1 : Structural Risk Minimization **…………………………………..** | 24 |
| 2.14.2 : Maximal Margin Hyperplane for Linear Separation **…………** | 25 |
| 2.14.3 : Nonlinear Separation **…………………………………………..** | 26 |
| 2.15 : Association Rules **………………………………………………………** | 27 |
| 2.16 : Apriori Algorithm **………………………………………………………..** | 29 |
| 2.16.1 : Generation of Frequent Itemsets **……………………………..** | 29 |
| 2.16.2 : Generation of Strong Rules **……………………………………** | 29 |
| 2.17 : General Association Rules **…………………………………………….** | 30 |
| 2.18 : Clustering **………………………………………………………………..** | 32 |
| 2.18.1 : Clustering Methods **…………………………………………….** | 32 |
| 2.18.2 : Taxonomy of Clustering Methods **…………………………….** | 33 |
| 2.18.3 : Affinity Measures **……………………………………………….** | 33 |
| 2.19 : Partition Methods **……………………………………………………….** | 35 |
| 2.19.1 : K-means Algorithm **……………………………………………..** | 35 |
| 2.19.2 : K-medoids Algorithm **…………………………………………...** | 37 |
| 2.20 : Hierarchical Methods **…………………………………………………..** | 37 |
| 2.20.1 : Agglomerative Hierarchical Methods **………………………...** | 38 |
| 2.20.2 : Divisive Hierarchical Methods **………………………………...** | 38 |
| 2.21 : Evaluation of Clustering Models **………………………………………** | 39 |
|  |  |
|  |  |
| **3. Chapter Three: CRISP-DM** |  |
| 3.1 : Introduction **……………………………………………………………….** | 41 |
| 3.1.1 : Business Understanding **………………………………………...** | 43 |
| 3.1.2 : Data Understanding **……………………………………………..** | 43 |
| 3.1.3 : Data Preparation **…………………………………………………** | 44 |
| 3.1.4 : Modeling **…………………………………………………………..** | 46 |
| 3.1.5 : Evaluation **………………………………………………………...** | 47 |
| 3.1.6 : Deployment **……………………………………………………….** | 47 |
| 3.2 : SEMMA **……………………………………………………………………** | 48 |
| 3.3 : Comparison of CRISP & SEMMA **………………………………………** | 49 |
|  |  |
| **4. Chapter Four: Related Works** |  |
| 4.1 : Introduction **……………………………………………………………….** | 51 |
| 4.2 : Apriori-TID Algorithm **…………………………………………………….** | 51 |
| 4.3 : Apriori-Hybrid Algorithm **…………………………………………………** | 51 |
| 4.4 : WFIM (Weighted Frequent Itemsets Mining) **………………………….** | 51 |
| 4.5 : Conclusion in Points **……………………………………………………..** | 52 |
|  |  |
| **5. Chapter Five: Apriori Algorithm** |  |
| 5.1 : Preface **…………………………………………………………………….** | 55 |
| 5.2 : Motivation and Structure of Association Rules **………………………..** | 55 |
| 5.2.1 : Market Basket Analysis **………………………………………….** | 55 |
| 5.2.2 : Web Mining **……………………………………………………….** | 55 |
| 5.2.3 : Purchases with a Credit Card **…………………………………..** | 56 |
| 5.2.4 : Fraud Detection **…………………………………………………..** | 56 |
| 5.3 : Mining Frequent Patterns and Association Rules **…………………….** | 57 |
| 5.4 : Apriori **……………………………………………………………………...** | 57 |
| 5.5 : Example - Synthetic Data on Purchases of Phone Faceplates **……..** | 60 |
| 5.6 : Generating Candidate Rules **……………………………………………** | 60 |
| 5.7 : Selecting Strong Rules **…………………………………………………..** | 62 |
| 5.7.1 : Support and Confidence **………………………………………...** | 62 |
| 5.7.2 : Lift Ratio **…………………………………………………………..** | 63 |
| 5.8 : Data Format **………………………………………………………………** | 64 |
| 5.9 : The Process of Rule Selection **………………………………………….** | 65 |
| **6. Chapter Six: Implementation** |  |
| 6.1 : Why Use Java **……………………………………………………...........** | 69 |
| 6.2 : General View **……………………………………………………………..** | 69 |
| 6.3 : The Implementation Results **…………………………………………….** | 75 |
| 6.4 : Programming Concepts **…………………………………………………** | 76 |
| 6.5 : The Main Class Methods – apriori class **………………………………** | 79 |
| 6.5.1 : Method - main(String[]) {lines 563-576} **……………………….** | 79 |
| 6.5.2 : Method - getAssociationRules(int,int,String) {lines 506-561} | 79 |
| 6.5.3 : Method - next(Pineapple.Slice,char) {lines 107-117} **………..** | 80 |
| 6.5.4 : Method - max(int) {lines 119-128} **……………………………...** | 81 |
| 6.5.5 : Method - candidate(Pineapple) {lines 130-170} **………………** | 81 |
| 6.5.6 : Method - prune(Pineapple) {lines 172-217} **…………………..** | 81 |
| 6.5.7 : Method - supportK1(Pineapple) {lines 219-230} **……………..** | 81 |
| 6.5.8 : Method - support(Pineapple) {lines 232-280} **…………………** | 82 |
| 6.5.9 : Method - print(Pineapple) {lines 282-306} **…………………….** | 82 |
| 6.5.10 : Method - readHeader( ) {lines 436-460} **……………………..** | 82 |
| 6.5.11 : Method - getRules(Pineapple,int,int) {lines 308-434} **………** | 82 |
| 6.5.12 : Method - sortRules( ) {lines 462-504} **………………………..** | 83 |
|  |  |
| **7. Chapter Seven: Conclusions and Recommendations** |  |
| 7.1 : Conclusions **……………………………………………………………….** | 85 |
| 7.2 : Recommendations **……………………………………………………….** | 85 |
|  |  |
| **References ……………………………………………………………………..** | 86 |
| **Appendix A …………………………………………………………………….** | 87 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Description** | **Page No.** |
| **1.1** | Benefits of a business intelligence system | **3** |
| **2.1** | Data mining process | **10** |
| **2.2** | Example of a classification tree | **18** |
| **2.3** | Univariate split for a binary attribute | **20** |
| **2.4** | Univariate split for a nominal attribute | **20** |
| **2.5** | Univariate split for a numerical attribute | **21** |
| **2.6** | Example of neural networks | **24** |
| **2.7** | The lines represent possible separating hyperplanes in the tow-dimensional space for the points in the diagram | **25** |
| **2.8** | Maximal margin separating hyperplanes and canonical hyper-planes for a linearly separable dataset | **26** |
| **2.9** | Nonlinear separation achieved by means of transformations in the feature space | **26** |
| **2.10** | Euclidean distance (full line) and Manhattan distance (dotted line) between two points in the plane | **34** |
| **2.11** | Contingency table for a binary attribute | **34** |
| **2.12** | An example of application of the K-means algorithm | **36** |
| **3.1** | CRISP-DM life cycle | **41** |
| **3.2** | CRISP-DM phases and tasks | **42** |
| **3.3** | SEMMA | **48** |
| **3.4** | Comparison of CRISP & SEMMA | **49** |
| **5.1** | Recommendations Based on Association Rules | **56** |
| **5.2** | Apriori Algorithm | **58** |
| **5.3** | Association Rule Generation Algorithm | **59** |
| **5.4** | Transactions for Purchases of Different Colored Cellular Phone Faceplates | **60** |
| **5.5** | Transactions for purchases in binary matrix format | **64** |
| **5.6** | Item sets have a count of at least 2 | **65** |
| **5.7** | Association Rules from the item set {red,white,green} | **66** |
| **5.8** | Association Rules for Phone Faceplates Transactions- XLMiner Output | **67** |
| **6.1** | Header file format | **70** |
| **6.2** | Dataset file format | **70** |
| **6.3** | The program running in Windows OS terminal | **71** |
| **6.4** | Message dialog box titled by “Association Rules:” with report | **72** |
| **6.5** | Same case study report but with less confidence 50% | **72** |
| **6.6** | Same case study terminal but with less confidence 50% | **73** |
| **6.7** | The report look organized after copy it to a file | **74** |
| **6.8** | Eclipse platform and on right showing apriori main class methods | **78** |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Description** | **Page No.** |
| **2.1** | Data mining, classical statistics and OLAP | **7** |
| **2.2** | Example of a dataset consisting of transactions defined over the set of objects | **27** |