بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Graduate Studies

Lecture Timetable System's Data Mining Case Study: SUST Timetable System

A Thesis Submitted in Fulfillment for the Requirements of MSc Degree in Computer Science

By:

Enas Awad Abd Elnabi Ali

(BSc Computer Science, Sudan University of Science and Technology)

Supervisor:

Dr. Mohemmed Alhafiz Mustafa Musa

June, 2013

Acknowledgement

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor, **Dr. Mohemmed Alhafiz Mustafa Musa,** Dean of College of Computer Science and Information Technology, Sudan University of Science and Technology, for giving me the opportunity to do research and providing in valuable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance. I am extremely grateful for what he has offered me. I would also like to thank him for his friendship, empathy, and great sense of humor.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my husband and my son for their love, understanding, prayers and continuing support to complete this research work. Also I express my thanks to my sisters, brothers, for their support and valuable prayers. My Special thanks go to my friends for their support to complete this thesis successfully.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Abstract

There are many computerized systems developed and used for many years in Sudan University of Science and Technology (SUST). These systems stored huge amount of data in auxiliary storage. In this research, the data from both SUST Timetable System specially Arabic and English lectures and SUST Registration System of academic years from 2009/2010 - 2011/2012 are used in mining, where the twostep algorithm is used to cluster the data as descriptive mining technique. SPSS Clementine is used to build the model, and Microsoft SQLServer is used to store the data set.

The process of data mining led to the discovery of interesting information that may help in having better timetable or can be used for another purposes. This discovered information focuses about teachers, classrooms and their capacity, location of classrooms verses location of student's college, location of the college of teacher verses location of student's college, and other information.

.

المستخلص

يوجود العديد من الأنظمة المطورة والمستخدمة منذ سنوات عديدة بجامعة السودان للعلوم والتكنولوجيا، حيث أتاحت هذه الأنظمة وجود قدر وفير من البيانات. تم إختيار البيانات المستخلصة من نظامي وضع جداول المحاضرات وبالأخص محاضرات اللغة العربية واللغة الإنجليزية، وكذلك نظام تسجيل الطلاب بالجامعة في الأعوام الدراسية من 2010/2009 إلى 2012/2011 لإجراء عمليات التنقيب عليها. كما تم استخدام خوارزمية الخطوتين لتجميع البيانات، حيث أنها تستخدم تقنية التنقيب غير الموجه. تم اختيار الأداة كلمنتاين لإجراء عملية التجميع، كما تم استخدام مايكروسوفت سيكول سيرفر لحفظ البيانات.

إن عملية التنقيب أدت لإكتشاف مجموعة من المعلومات التي قد تساعد في وضع وتنظيم الجداول كما أنها يمكن الاستفادة منها في أغراض أخرى. حيث ركزت هذه المعلومات على الأساتذة، القاعات الدراسية وسعاتها، موقع القاعات الدراسية التي استخدمت للمحاضرة مقارنة مع موقع الكلية التي يدرس بها الطلاب، موقع الكلية التي يعمل بها الأستاذ مقارنة مع موقع الكلية التي يدرس بها الطلاب، ومعلومات أخرى.

Table of Contents

Content	Page No.
Acknowledgement	I
Abstract	II
المستخلص	III
Table of Contents	IV
List of Figures	VI
List of Tables	VII
1 Introduction	1
1.1 Motivation	2
1.2 Problem Definition	2
1.3 Objectives	2
1.4 Methodology	2
1.5 Research Organization	3
2 Clustering	4
2.1 Introduction	5
2.2 Data Mining	5
2.3 Knowledge Discovery from Data (KDD)	5
2.4 Uses of Clustering	6
2.5 Clustering Techniques	7
2.6 Clustering Methods	8
2.7 Classes of Clustering Tasks	9
2.8 Characteristics of a Good Clustering	9
2.9 Outlier Analysis	11
3 Twostep Algorithm	12
3.1 Introduction	13
3.2 Twostep Clustering Procedures	13
3.3 Accuracy	15
3.4 Distance Measure	15
3.4.1 Log-Likelihood Distance	16
3.4.2 Euclidean Distance	16
3.5 Performance of the Twostep Clustering Algorithm	16
3.6 Why Twostep is used in research	17
4 Case Study	18

4.1 Introduction	19
4.2 Timetable System	19
4.3 Registration System	21
4.4 Mining the Data	22
4.4.1 Problem specification and The Application Domain	22
4.4.2 Resourcing	22
4.4.3 Data Cleaning and Preprocessing	22
4.4.4 Data reduction and projection	23
4.4.5 Data mining	24
4.4.6 Experimental Result	26
4.4.7 The Discovered Knowledge	27
5 Conclusion	34
5.1 Conclusion	35
5.2 Recommendation and Future Work	35
References	38

List of Figures

Figure	Page No.
1.1 Knowledge Discovery from Data	6
4.1 The SUSTs Clustering model	26
4.2 Cluster number 18, it shows English lecturers who work permanently on the College of Engineering	27
4.3 Cluster number 5, it shows the part timer English lecturers in college of Business Studies	28
4.4 Cluster number 14, it shows Arabic lectures in the college of Agricultural Studies	29
4.5 Cluster number 14, it shows the class capacity vs. average number of students in first year of college of Agricultural Studies	30
4.6 Cluster number 16, it shows the class capacity vs. average number of students in second year of the department of Electrical Engineering.	31
4.7 Cluster number 2, it shows the gender of students in first year of college of Education VS. their Arabic lecturer's gender	32
4.8 Cluster number 5, it shows position of English lectures VS. Position of college of Business Studies Department of Banking and Finance, also shows the room capacity VS. the number of students in first year.	33

List of Tables

Table	Page No.
3.1 Performance of the twostep clustering algorithm	16
4.1 The raw dataset of lectures timetable in Sudan University of Science and Technology	20
4.2 The raw dataset schema obtained from the Registration System	21
4.3 All attributes in data set after join the dataset from the two systems	25