MINING RARE RELATIONAL PATTERNS FROM MULTI RELATIONAL TABLES

SITI HAIRUL NITA SELAMAT

MASTER OF SCIENCE

KOLEJ UNIVERSITI SAINS DAN TEKNOLOGI

MALAYSIA

2006

MINING RARE RELATIONAL PATTERNS FROM MULTI RELATIONAL TABLES

SITI HAIRUL NITA SELAMAT

Thesis Submitted in Fulfillment of the Requirement for the Degree of Master of Science in the Faculty of Science and Technology Kolej Universiti Sains dan Teknologi Malaysia

October 2006

TABLE OF CONTENTS

To my lovely mom,
Thank you for your great selfless love..

Abstract of thesis presented to the Senate of Kolej Universiti Sains dan Teknologi Malaysia in fulfillment of the requirement for the degree of Master of Science

MINING RARE RELATIONAL PATTERNS FROM MULTI RELATIONAL TABLES

SITI HAIRUL NITA SELAMAT

October 2006

Chairperson : **Associate Professor Muhammad Suzuri Hitam, Ph.D.**

Member : **Professor Mustafa Mat Deris, Ph.D.**

Faculty : **Science and Technology**

In recent years, the amount of data that stored in any organizational databases has been dramatically increased. Hence, there is a need to discover valuable hidden knowledge from these huge database resources. This is where data mining techniques used for analyzing the data to derive hidden knowledge that has obvious benefits for any organizations.

Previous researches on mining association rules in relational tables only focus on discovering the relationship among large data items in a database. However, association rule for significant rare data item that appear infrequently in a database but are highly related with other data item is yet to be discovered. Differently with single table case, multi relational tables consist of co-occurrences of attributes that some are related with specific attributes in other tables.

In this thesis, a newly proposed algorithm named Extraction Rare Pattern (ERP) had been proposed for extracting the rare relational patterns from multi relational tables that

employs a couple of predefined minimum support thresholds. The results have been compared with existing Relative Support Apriori Algorithm (RSAA) technique. It is found out that ERP technique is not only able to extract rare data item, but also its run time is much lower as compared to the exsting RSAA technique. Result from the implementation to a sample relational tables reveals that the proposed algorithm is capable of mining rare data item in multi relational tables. Discovering these patterns from multi relational tables can provide significant and useful knowledge for support in effective organizational decision-making.

Abstrak tesis yang dikemukakan kepada Senat Kolej Universiti Sains dan Teknologi Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENGEKSTRAKAN PATEN HUBUNGAN TERKECIL DARIPADA JADUAL HUBUNGAN BERBILANG

SITI HAIRUL NITA SELAMAT

Oktober 2006

Pengerusi : Profesor Madya Muhammad Suzuri Hitam, Ph.D.

Ahli : Profesor Mustafa Mat Deris, Ph.D.

Fakulti : Sains dan Teknologi

Sejak beberapa tahun kebelakangan ini, jumlah pangkalan data bagi organisasi telah mengalami peningkatan data secara mendadak. Sehubungan dengan itu, terdapat keperluan untuk menjana informasi dan pengetahuan yang tersembunyi dan bernilai di dalam pangkalan data tersebut. Oleh sebab itu, teknik-teknik perlombongan data telah digunakan untuk menganalisis data-data tersebut bagi menjana aturan yang tersembunyi yang mana secara jelas dapat memberi manfaat kepada sebarang organisasi.

Sebelum ini, perlombongan aturan penyatuan dalam jadual hubungan berbilang hanya memfokuskan penemuan mengenai hubungan antara item-item data yang kerap di dalam pangkalan data. Walau bagaimanapun, aturan penyatuan bagi penemuan item data yang sedikit bilangannya tetapi penting iaitu item data yang jarang dan penting di dalam pangkalan data tetapi mempunyai hubungan yang sangat kuat dengan item-item data tertentu yang lain masih belum diterokai lagi. Berbeza dengan kes jadual tunggal, jadual hubungan berbilang mempunyai atribut-atribut yang berkait dengan atribut tertentu di dalam jadual-jadual yang lain.

Sehubungan dengan itu, di dalam tesis ini mencadangkan satu algoritma baru yang dinamakan algoritma Pengekstrakan Corak Jarang (ERP) daripada jadual hubungan berbilang yang mana ia menggunakan dua nilai sokongan minimum untuk mengesktrak corak hubungan bagi item-item data yang jarang. Keputusan implementasi telah dibandingkan dengan teknik yang telah sedia ada iaitu Algoritma Apriori Sokongan Relatif (RSAA). Hasil daripada pemerhatian didapati bahawa teknik ERP bukan hanya berupaya mengekstrak data item yang jarang, tetapi juga ia berjaya melaksanakan fungsinya dalam masa yang lebih singkat berbanding dengan teknik RSAA. Hasil daripada implimentasi yang telah dilakukan keatas suatu contoh pangkalan data hubungan, didapati bahawa algoritma yang diperkenalkan ini boleh diaplikasikan bagi melombong item yang jarang untuk jadual hubungan berbilang. Penghasilan corak daripada jadual hubungan berbilang ini menyediakan pengetahuan yang penting dan berguna bagi menyokong membuat keputusan yang berkesan bagi suatu organisasi.