UNIVERSITI PUTRA MALAYSIA

PRIME-BASED METHOD FOR INTERACTIVE MINING OF FREQUENT PATTERNS

MOHAMMAD-HOSSEIN NADIMI-SHAHRAKI

FSKTM 2010 10

PRIME-BASED METHOD FOR INTERACTIVE MINING OF FREQUENT PATTERNS

MOHAMMAD-HOSSEIN NADIMI-SHAHRAKI

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2010

PRIME-BASED METHOD FOR INTERACTIVE MINING OF FREQUENT PATTERNS

By

MOHAMMAD-HOSSEIN NADIMI-SHAHRAKI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,
in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

December 2010

Dedicated to my Wife, my Daughter and
my Parents for their Love and Affection

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

PRIME-BASED METHOD FOR INTERACTIVE MINING OF FREQUENT PATTERNS

By

MOHAMMAD-HOSSEIN NADIMI-SHAHRAKI

December 2010

Chairman: Norwati Mustapha, PhD

Faculty: Computer Science and Information Technology

Over the past decade, an increasing number of efficient mining algorithms have been proposed to mine the frequent patterns by satisfying a user specified threshold called minimum support (*minsup*). However, determining an appropriate value for *minsup* to find proper frequent patterns in different applications is extremely difficult. Since rerunning the mining algorithms from scratch can be very time consuming, researchers have introduced interactive mining to find proper patterns by using the current mining model with various *minsup*. Thus far, a few efficient interactive mining algorithms have been proposed. However, their runtime do not fulfill the need of short runtime in real time applications especially where data is sparse and proper frequent patterns are mined with very low values of *minsup*.

As response to the above-mentioned challenges, this study is devoted towards developing an interactive mining method based on prime number and its special characteristic “uniqueness” by which the content of the relevant data is transformed into a compact layout. At first, a general architecture for interactive mining is proposed consisting of two isolated components: mining model and mining process. Then, the proposed method is developed based on the architecture such that the mining model is constructed once, and it can be frequently mined by various *minsup*.

In the mining model construction, the content of relevant data is captured by a novel tree structure called *PC-tree* with one database scan and mining materials are consequently formed. The *PC-tree* is a well-organized tree structure, which is systematically built based on descendant making introduced in this study. Moreover, this study introduces a mining algorithm called *PC-miner* to mine the mining model frequently with various values of *minsup*. It grows an effective candidate head set introduced in this study starting from the longest candidate patterns by using the Apriori principle. Meanwhile, during the growing of the candidate head set in each round, the longest candidate patterns are used to find maximal frequent patterns from which the frequent patterns can be derived. Moreover, the *PC-miner* reduces the number of candidate patterns and comparisons by using several pruning techniques.

A comprehensive experimental analysis is conducted by several experiments and scenarios to evaluate the correctness and effectiveness of the proposed method especially for interactive mining. The experimental results verify that the proposed method constructs the mining model independent of *minsup* once and this enable the model to be frequently mined. The results also show that the proposed method mines frequent patterns correctly and efficiently. Moreover, the results verify that the proposed method speeds up interactive mining of frequent patterns over both sparse and dense datasets with more scalable total runtime for very low values of *minsup* over sparse datasets as compared to results from the previous work.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

KAEDAH BERASASKAN NOMBOR PERDANA UNTUK PERLOMBONGAN SECARA INTERAKTIF BAGI CORAK YANG KERAP

Oleh

MOHAMMAD-HOSSEIN NADIMI-SHAHRAKI

Disember 2010

Pengerusi: **Norwati Mustapha, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Selama berdekad yang lalu, bilangan cadangan algoritma perlombongan yang efisyen telah bertambah dalam melombong corak kerap yang memenuhi nilai ambang ketetapan pengguna iaitu sokongan minimum (*minsup*). Walau bagaimanapun, penetapan sesuatu nilai *minsup* yang sesuai bagi mendapatkan corak kerap yang wajar dalam aplikasi yang berbeza adalah sangat sukar. Oleh kerana melarikan semula algoritma perlombongan dari awal adalah sangat memakan masa, para penyelidik telah memperkenalkan perlombongan interaktif bagi mendapatkan corak kerap yang wajar dengan menggunakan model perlombongan semasa dengan pelbagai *minsup*. Setakat ini, sebilangan algoritma perlombongan interaktif yang efisien telah dicadangkan. Namun, masa lariannya tidak memenuhi keperluan masa larian yang pendek dalam aplikasi masa nyata terutamanya apabila pangkalan data adalah jarang dan corak kerap yang wajar dilombong dengan nilai *minsup* yang sangat rendah.

Bagi menyahut cabaran tersebut, kajian ini memberi tumpuan kepada pembangunan kaedah perlombongan interaktif berdasarkan nombor perdana dan ciri khususnya "keunikan" yang mana kandungan data yang relevan ditransfomasikan ke dalam satu susun atur yang padat. Pada awalnya, satu senibina umum bagi perlombongan interaktif yang terdiri daripada dua komponen terasing: model perlombongan dan proses perlombongan telah dicadangkan. Kemudian, kaedah cadangan tersebut dibangunkan berdasarkan senibina yang mana sebaik sahaja model perlombongan dibina, ianya boleh dilombong secara kerap dengan menggunakan pelbagai *minsup*.

Dalam pembinaan model perlombongan, kandungan data yang relevan diperoleh daripada struktur pepohon baharu yang dipanggil *PC-tree* dengan satu imbasan pangkalan data dan seterusnya bahan-bahan perlombongan akan dibentuk. *PC-tree* adalah satu struktur pepohon yang terancang yang dibina secara sistematik berdasarkan pembentukan turunan yang diperkenalkan dalam kajian ini. Tambahan lagi, kajian ini memperkenalkan satu algoritma perlombongan yang dipanggil *PC-miner* bagi melombong model perlombongan secara kerap dengan pelbagai nilai *minsup*. Algoritma ini mengembangkan satu set induk calon efektif yang diperkenalkan dalam kajian ini bermula daripada calon corak yang paling panjang dengan menggunakan prinsip Apriori. Sementara itu, semasa set induk calon dikembangkan dalam setiap pusingan, calon corak yang paling panjang digunakan untuk mendapatkan corak kerap maksima yang membolehkan corak kerap diterbitkan. Ini akan mengurangkan jumlah calon corak dan perbandingan melalui beberapa teknik pemangkasan.

Analisis eksperimental yang menyeluruh telah dibuat melalui beberapa eksperimen dan senario untuk menilai ketepatan dan keberkesanan kaedah yang dicadangkan terutamanya bagi perlombongan interaktif. Keputusan eksperimen membuktikan bahawa kaedah yang dicadangkan iaitu membina model perlombongan bebas *minsup* hanya sekali dan ini membolehkan model dilombong dengan lebih kerap. Keputusan juga menunjukkan kaedah yang dicadangkan melombong corak kerap dengan betul dan cekap. Tambahan lagi, keputusan mengesahkan bahawa kaedah perlombongan yang dicadangkan mempercepatkan perlombongan interaktif bagi corak kerap ke atas kedua-dua dataset jarang dan padat dengan jumlah keseluruhan masa larian lebih berskala untuk nilai-nilai *minsup* yang sangat rendah ke atas dataset jarang berbanding dengan hasil kajian-kajian terdahulu.

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my supervisor Dr. Norwati Mustapha for giving me an opportunity to start this study. Through the course of my study, I have had the great fortune to get to know and interact with her. Her comments and suggestions for further development as well as her assistance during writing this thesis are invaluable to me. Her specific background on frequent pattern mining, interest, teaching and research style has provided for me an exceptional opportunity to learn more.

I would like to express my sincere thanks and appreciation to the supervisory committee members Associate Professor Dr. Md Nasir Sulaiman and Associate Professor Dr. Ali Mamat for their guidance, valuable suggestions and advice throughout this work in making this a success.

My deepest appreciation to my wife Ms. Zahra Naseri and my daughter Nazanin who have been supportive and patiently waiting for me to complete my study. Finally, I owe my sincere thanks to my parents for their encouragement and affirmation, which made it possible for me to achieve this work.

For the others who have directly or indirectly helped me in the completion of my work, I thank you all.

I certify that an Examination Committee met on 16/ 12 / 2010 to conduct the final examination of **Mohammad-Hossein Nadimi-Shahraki** on his Doctor of Philosophy thesis entitled **"Prime-based method for interactive mining of frequent patterns"** in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the Doctor of Philosophy degree. Members of the Examination Committee were as follows:

Hamidah Ibrahim, PhD
Associate Professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Chairman)

Azmi Jaafar, PhD
Associate Professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

Masrah Azrifah Azmi Murad, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

David Taniar, PhD
Associate Professor
Clayton School of Information Technology
Monash University of Australia
(External Examiner)

SHAMSUDDIN SULAIMAN, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of philosophy. The members of the Supervisory Committee were as follows:

Norwati Mustapha, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Md. Nasir B Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Ali Mamat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

HASANAH MOHD. GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at University Putra Malaysia or other institution.

MOHAMMAD-HOSSEIN NADIMI-SHAHRAKI

Date: 16-12-2010

TABLE OF CONTENTS

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7