UNIVERSITI PUTRA MALAYSIA

FREQUENT LEXICOGRAPHIC ALGORITHM FOR MINING ASSOCIATION RULES

NORWATI MUSTAPHA.

FSKTM 2005 9

FREQUENT LEXICOGRAPHIC ALGORITHM FOR MINING ASSOCIATION RULES

By

NORWATI MUSTAPHA

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

June 2005

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirements for the degree of Doctor of Philosophy

FREQUENT LEXICOGRAPHIC ALGORITHM FOR MINING ASSOCIATION RULES

By

NORWATI MUSTAPHA

June 2005

Chairman : **Associate Professor Md. Nasir Sulaiman, PhD**

Faculty : **Computer Science and Information Technology**

The recent progress in computer storage technology have enable many organisations to collect and store a huge amount of data which is lead to growing demand for new techniques that can intelligently transform massive data into useful information and knowledge. The concept of data mining has brought the attention of business community in finding techniques that can extract nontrivial, implicit, previously unknown and potentially useful information from databases. Association rule mining is one of the data mining techniques which discovers strong association or correlation relationships among data. The primary concept of association rule algorithms consist of two phase procedure. In the first phase, all frequent patterns are found and the second phase uses these frequent patterns in order to generate all strong rules. The common precision measures used to complete these phases are support and confidence. Having been investigated intensively during the past few years, it has been shown that the first phase involves a

major computational task. Although the second phase seems to be more straightforward, it can be costly because the size of the generated rules are normally large and in contrast only a small fraction of these rules are typically useful and important. As response to these challenges, this study is devoted towards finding faster methods for searching frequent patterns and discovery of association rules in concise form.

An algorithm called Flex (**F**requent **lex**icographic patterns) has been proposed in obtaining a good performance of searching frequent patterns. The algorithm involved the construction of the nodes of a lexicographic tree that represent frequent patterns. Depth first strategy and vertical counting strategy are used in mining frequent patterns and computing the support of the patterns respectively.

The mined frequent patterns are then used in generating association rules. Three models were applied in this task which consist of traditional model, constraint model and representative model which produce three kinds of rules respectively; all association rules, association rules with 1-consequence and representative rules. As an additional utility in the representative model, this study proposed a set-theoretical intersection to assist users in finding duplicated rules.

Four datasets from UCI machine learning repositories and domain theories except the pumsb dataset were experimented. The Flex algorithm and the other two existing algorithms Apriori and DIC under the same specification are tested toward these datasets and their extraction times for mining frequent patterns were recorded and compared. The experimental results showed that the proposed algorithm outperformed both existing algorithms especially for the case of long patterns. It also gave promising results in the case of short patterns. Two of the datasets were then chosen for further experiment on the scalability of the algorithms by increasing their size of transactions up to six times. The scale-up experiment showed that the proposed algorithm is more scalable than the other existing algorithms.

The implementation of an adopted theory of representative model proved that this model is more concise than the other two models. It is shown by number of rules generated from the chosen models. Besides a small set of rules obtained, the representative model also having the lossless information and soundness properties meaning that it covers all interesting association rules and forbid derivation of weak rules. It is theoretically proven that the proposed set-theoretical intersection is able to assist users in knowing the duplication rules exist in representative model.

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

**ALGORITMA FREQUENT LEXICOGRAPHIC
BAGI MELOMBONG PETUA-PETUA SEKUTUAN**

Oleh

NORWATI MUSTAPHA

Jun 2005

Pengerusi : Profesor Madya Md. Nasir Sulaiman, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Sebagaimana perkembangan semasa di dalam teknologi storan komputer telah membuatkan banyak organisasi mampu untuk mengumpul dan menyimpan sejumlah data yang besar, terdapat pertambahan permintaan bagi teknik-teknik baru yang mampu menukar secara pintar data yang besar itu kepada maklumat dan pengetahuan yang berguna. Konsep perlombongan data telah menarik perhatian komuniti perniagaan sebagai satu teknik yang memetik maklumat penting, tersirat, tidak diketahui pada awalnya dan berpotensi penggunaanya daripada data di dalam pangkalan data. Melombongi petua kesatuan adalah salah satu teknik perlombongan data yang mencari kesatuan yang kuat atau hubungan perkaitan di antara data. Konsep utama disebalik kebanyakan algoritma-algoritma petua kesatuan ialah satu tatacara yang mempunyai dua fasa. Di dalam fasa yang pertama, semua corak yang kerap ditemui dan fasa yang kedua menggunakan corak yang kerap ini bagi tujuan untuk menjana semua petua-petua yang

kuat. Ukuran ketepatan yang biasa digunakan bagi melengkapkan fasa-fasa ini adalah sokongan dan keyakinan. Setelah disiasat secara intensif selama beberapa tahun yang lalu, ianya menunjukkan bahawa fasa yang pertama adalah merupakan tugas pengiraan utama. Walaupun fasa yang kedua adalah sejajar, ianya mungkin mahal kerana petua-petua yang dijana biasanya besar tetapi sebaliknya peratusan bagi petua-petua yang sangat berguna biasanya hanya satu pecahan yang sangat kecil. Sebagai tindakbalas kepada cabaran-cabaran ini, kajian ini menumpukan kepada mencari kaedah-kaedah yang lebih cepat bagi mencari corak-corak yang kerap dan mendapatkan petua-petua sekutuan dalam bentuk yang ringkas dan padat.

Satu algoritma yang dipanggil Flex (Frequent lexicographic patterns) telah dicadangkan dalam memperolehi satu prestasi yang baik bagi mencari corak-corak yang kerap. Algoritma ini melibatkan pembentukan nod-nod bagi satu pepohon leksikografi yang mewakili corak-corak yang kerap itu. Strategi dalam dahulu telah digunakan dalam melombongi corak-corak yang kerap bersama-sama dengan strategi membilang secara menegak bagi membantu dalam pengiraan sokongan untuk setiap corak.

Corak-corak yang kerap yang telah dilombongi kemudiannya digunakan dalam pengiraan petua-petua. Tiga model telah digunakan dalam tugas ini yang terdiri daripada model tradisional, model kekangan dan model perwakilan yang akan mengeluarkan tiga jenis petua; semua petua sekutuan, petua sekutuan dengan 1-keputusan dan petua perwakilan. Sebagai utiliti tambahan di dalam model perwakilan, kajian ini telah mencadangkan satu tindanan set-teori untuk membantu pengguna-pengguna dalam mencari petua-petua yang berulang.

Empat data set daripada *UCI machine learning repositories and domain theories* kecuali pumsb data set telah diuji. Dengan melarikan algoritma Flex dan dua algoritma yang sedia ada iaitu Apriori dan DIC di bawah spesifikasi yang sama, masa melombongi corak yang kerap telah dibandingkan. Hasil eksperimen menunjukkan algoritma yang dicadangkan telah melebihi tahap kedua-dua algoritma sedia ada terutamanya untuk kes bagi corak-corak yang panjang. Ia juga memberikan hasil yang setanding untuk kes bagi corak-corak yang pendek. Dua data set kemudiannya telah dipilih untuk eksperimen seterusnya ke atas penskalaan algoritma-algoritma berkenaan dengan meningkatkan saiz transaksi sehingga enam kali ganda. Eksperimen penskalaan telah menunjukkan algoritma yang dicadangkan adalah lebih berskala daripada algoritma-algoritma sedia ada.

Perlaksanaan satu teori yang diadaptasi bagi model perwakilan telah membuktikan bahawa model ini lebih ringkas dan padat daripada model-model yang lain. Ini ditunjukkan oleh bilangan petua-petua yang dikeluarkan daripada model-model yang dipilih. Disamping set petua yang sedikit disediakan, model perwakilan juga mempunyai ciri-cirinya iaitu maklumat yang tidak hilang dan kukuh bermaksud ia merangkumi semua petua kesatuan yang menarik dan menghalang terbitan petua-petua yang lemah. Terdapat juga pembuktian secara teori iaitu tindanan set-teori yang dicadangkan mampu membantu pengguna-pengguna dalam mengetahui petua-petua berulang yang wujud di dalam model pewakilan.

ACKNOWLEDGEMENTS

First of all, thank to God, the most Gracious and most Merciful.

Many people have made contributions in completing this work. My deepest appreciation and gratitude to the supervisory committee leads by Prof. Madya Dr. Md. Nasir Sulaiman and committee members, Prof. Madya Dr. Mohamed Othman and Prof. Madya Hj. Mohd. Hasan Selamat for their virtuous, ideas, intellectual experiences and support that led the way in so many aspects of research work. Gratitude also go to my beloved husband, Prof. Madya Dr. Husaini Omar, my lovely daughters, Syafrina and Syafiqa and my son, Aswan for their supporting, understanding and love. I wish to thank my parents who are constantly praying for my success.

I am also indebted to the following: Ms Azuraliza, Kryszkiewicz, M., Zaki, M.J. and Bayardo, R.J. for their stimulating discussions and ideas. Sincere thanks towards everyone in Faculty of Computer Science and Information Technology, Universiti Putra Malaysia for their helps and making me have an enjoyable period of time.

Norwati Mustapha

June 2005

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AR	Association Rules
AR1	Association Rules with 1-consequence
ARM	Association Rule Mining
DIC	Dynamic Itemset Counting
FC	Frequent Closed
FP	Frequent Patterns
KDD	Knowledge Discovery in Databases
RR	Representative Rules
SAR	Strong Association Rules

CHAPTER I

INTRODUCTION

Background

Data Mining aims at the discovery of useful knowledge in large data collections. The rapidly growing interest in the field is stimulated by the large amounts of computerized data available in business and also in science. For instance, supermarkets store electronic copies of millions of receipts, while banks and credit card companies maintain extensive collections of transactions histories. It is no longer possible to analyse it manually using traditional methods or even a well-known technologies in statistics and computer science. Therefore, the concept of *Knowledge Discovery in Databases* (KDD) has been brought as an effort to analyse the huge volume of data and to find useful knowledge that provide new insight into business (Piatetsky-Shapiro and Fawley, 1991; Fayyad *et. al.,* 1996).

Knowledge Discovery in Databases is defined as the non-trivial extraction of valid, implicit, potentially useful and ultimately understandable patterns (knowledge) in large databases (Cabena *et. al.*, 1998). In general, there are many kinds of patterns (knowledge) that can be extracted from data. For example, association rules can be mined for market basket analysis, classification rules can be found for accurate classifiers, clusters and outliers can be identified for customer relation management.

There are several tasks in data mining and one of the important tasks is association rule mining. Since its introduction in 1993 by Agrawal *et. al.,* mining of such rules is still one of the most popular pattern discovery in KDD (Hipp *et. al.*, 2000). Association rule mining is a task of data mining to extract interesting relationship among data attributes in large dataset. An example of such rule might be that *98% of customers that purchase bread and cheese also purchase milk*. The problem of discovering all association rules can be decomposed into two subproblems (Agrawal *et. al.,* 1993a). First, find all sets of items (patterns) that have transaction support above minimum support called frequent patterns. Second, use the frequent patterns to generate the desired rules.

In the literature, there are several algorithms have been proposed and implemented by researchers to find faster methods for generating frequent patterns. The most popular algorithm is Apriori (Agrawal *et. al.,* 1994) where the downward closure property of itemset support was introduced. Apriori makes additional use of this property by pruning those candidates that have an infrequent subset before counting their supports. This optimization becomes possible because breadth first search ensures that the support values of all subsets of a candidate are known in advance. The critical part of Apriori is counting all candidates in each of the transactions and involved repetitive passing over the database. The performance of Apriori degrades when mining long patterns and it is not suitable for low values of minimum support.

The Partition algorithm was proposed by Savasere *et. al.* (1995) takes a different approach. It splits the database into several chunks that it can be accommodated in main-memory and they are treated independently. Whereas this optimization helps to cope with large databases, it adds the additional overhead of an extra pass to determine the globally frequent patterns. For lower values of minimum support, Partition suffers strongly because of the increasing number of locally frequent patterns that finally turn out to be globally infrequent.

The method of random sampling was introduced by Toivonen (1996) to generate frequent patterns may save considerable expense in terms of the I/O costs. The weakness of using this method is that it may often result in inaccuracies because of the presence of data skew. Data which are located on the same page may often be highly correlated and may not represent the over all distribution of patterns through the entire database.

DIC algorithm (Brin *et. al.*, 1997b) is further variation of the Apriori. DIC soften the strict separation between counting and generation candidates. It employed a prefix-tree instead of hash tree used in Apriori. Interlocking support determination and candidate generation result in decreasing the number of database scans. Experimental result shows that DIC is better than Apriori for low minimum support values.

Anti-skew algorithms for mining frequent patterns has been discusses by Lin and Dunham (1998). The techniques proposed in this paper reduce the maximum number of scans. The algorithm uses a sampling process in order to collect knowledge about the data and reduce the number of passes. The problems created by data skewness also arise in the context of parallel methods which divide the load among processors by partitioning the transaction data among the different processors This is because each