AN EFFICIENT ALGORITHM TO DISCOVER COLOSSAL CLOSED ITEMSETS IN HIGH DIMENSIONAL DATA

BY

FATIMAH AUDAH MD. ZAKI

A thesis submitted in fulfillment of the requirement for the degree of Doctor of Philosophy (Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

MARCH 2020

ABSTRACT

The current trend of data collection involves a small number of observations with a very large number of variables, known as high dimensional data. Mining these data produces an explosive number of smaller itemsets which are less important than colossal (large) ones. As the trend in Frequent Itemset Mining is moving towards mining colossal itemsets, it is important to understand the challenges in order to formulate a better method that is faster in running time, more scalable and able to produce useful and interesting knowledge. For this reason, this thesis has proposed two new algorithms; RARE and RARE II, which mine colossal closed itemsets. Both algorithms apply a minimum cardinality threshold to limit the search space and a closure computation method that does not require storage of previously discovered itemsets for duplicates checking. These approaches improved both memory and time requirement of the algorithms to finish mining tasks. Algorithm RARE searches the rowset lattice in breadth-first manner which resulted to a reduced itemset intersections compare to other state-of-the-art algorithms, CARPENTER and IsTa. Although the different threshold used in CARPENTER and IsTa make direct comparison with RARE difficult, RARE proved to be better. In terms of memory usage, RARE need only one-third of CARPENTER's and one-tenth of IsTa's, while require the least running time to discover 100% of closed itemsets in the dataset. Meanwhile, RARE II further reduced itemset intersections by evaluating only the closed rowsets in order to mine the next closed itemsets, which resulted to an improved run time by more than 50% compare to RARE.

خلاصة البحث

يتضمن الاتجاه الحالي لجمع البيانات عددًا صغيرًا من الملاحظات مع عدد كبير جدًا من المتغيرات ، المعروفة باسم البيانات ذات الأبعاد العالية. ينتج عن استخراج هذه البيانات عددًا كبيرًا من العناصر الأصغر التي تكون أقل أهمية من تلك الضخمة (الكبيرة). نظرًا لأن الاتجاه في التنقيب البياناتي المتكرر لمجموعة العناصر يتجه نحو مجموعات التنقيب الضخمة ، من المهم فهم التحديات من أجل صياغة طريقة أفضل تكون أسرع في التشغيل ، وأكثر قابلية للتطوير وقادرة على إنتاج معرفة مفيدة ومثيرة للاهتمام. لهذا السبب ، اقترحت هذه الرسالة خوارزميتين جديدتين (RARE)، ؛ و (RAREII) ، اللتان تعملان على التنقيب البياناتي لمجموعات العناصر الضخمة المغلقة. تستعمل كلتا الخوارزميتين حدًا أدنى من عدد العناصرفي المجموعة للحد من مساحة البحث وطريقة حساب الإغلاق التي لا تتطلب تخزين مجموعات العناصر المكتشفة سابقًا للتحقق من التكرارات. حسنت هذه الأساليب كلا من متطلبات الذاكرة والوقت للخوارزميات لإنهاء مهام التنقيب. تقوم الخوارزمية (RARE)، بالبحث في □بكة الصفوف بطريقة السعة الأولية ، مما أدى إلى تقاطع عناصر أقل مقارنة بخوارزميات أخرى متطورة مثل (CARPENTER) و(IsTa) . وعلى الرغم من أن العتبات المختلفة المستخدمة في (CARPENTER) و(IsTa) تجعل المقارنة المبا□رة مع (RARE)، صعبة ، فقد أثبتت (RARE) أنها أفضل. من حيث استخدام الذاكرة (RARE)، تحتاج فقط إلى ثلث (CARPENTER) وعشر (IsTa) ، بينما تتطلب أقل وقت تشغيل لاكتشاف 100٪ من العناصر المغلقة في مجموعة البيانات. وفي الوقت نفسه (RAREII)، قلل من تقاطعات مجموعة العناصر من خلال تقييم مجموعات الصفوف المغلقة فقط من أجل استخراج العناصر المغلقة التالية ، مما أدى إلى تحسين وقت التشغيل بأكثر من 50٪ مقارنة بـ (RARE).

APPROVAL PAGE

The thesis of Fatimah Audah Md. Zaki has been approved by the following:

____

Dr. Nurul Fariza Zulkurnain
Supervisor

____

Prof. Dr. Teddy Surya Gunawan
Co-Supervisor

____

Prof. Dr. Shihab A. Hameed
Internal Examiner

____

Prof. Dr. Azuraliza Abu Bakar
External Examiner

____

Prof. Dr. Muhammad Naqib S/O Ihsan Jan
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Fatimah Audah Md. Zaki

Signature .... Date ....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

AN EFFICIENT ALGORITHM TO DISCOVER COLOSSAL CLOSED ITEMSETS IN HIGH DIMENSIONAL DATA

I declare that the copyright holders of this dissertation are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Fatimah Audah Md. Zaki

……..…………………….. ……………………….

Signature Date

ACKNOWLEDGEMENTS

Firstly, it is my utmost pleasure to dedicate this work to my dear parents and my family, who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

I wish to express my appreciation and thanks to those who provided their time, effort and support for this project. To the members of my dissertation committee, thank you for sticking with me.

Finally, a special thanks to Dr. Nurul Fariza Zulkurnain and Prof. Dr. Teddy Surya Gunawan for their continuous support, encouragement and leadership, and for that, I will be forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA   Deoxyribonucleic Acid

HDSSS   High Dimensional Small Sample Size

ARM   Association Rule Mining

FIM   Frequent Itemset Mining

SNP   Single Nucleotide Polymorphisms

CFI   Closed Frequent Itemset

RA   Rheumatoid Arthritis

SLE   Systemic Lupus Erythematosus

SSc   Systemic Sclerosis

GRN   Gene Regulatory Network

LIMMA   Linear Models for Microarray Data

DE   Differential Gene Expression

GO   Gene Ontology

GE   Grammatical Evolution

SAM   Significant Analysis of Microarrays

FIMI   Frequent Itemset Mining Implementations

LC   Lung Cancer

ALL   Acute Lymphoblastic Leukemia

OC   Ovarian Cancer

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Currently, we are in the age of massive automatic data collection and the existing trends are likely to accelerate in the near future. Various types of data are constantly being collected; i.e. from traditional supermarket transactions and credit card records, to scientific kinds of data such as astronomical images, molecular structures, Deoxyribonucleic acid (DNA) sequences and human genome data. These massive automatic data collections have produced another kind of dataset, known as high-dimensional data, which is characterized by a small number of observations (rows), *n*, compared to a large number of features (columns or dimensions), *m* (Jain et. al., 2018). The term High Dimensional Small Sample Size (HDSSS) is referred to data with $n < 100$ and $m > 10^3$ (Li et. al., 2017) (Mafarja et. al., 2018) (Barddal et. al., 2019), and ultra-high dimensional when the dimension is extremely large, i.e. $m > 10^6$ (Yamada et. al., 2018).

In order to extract insights from high dimensional data; mining algorithm, software tools, and computing specifications should also be improved. Though there are numerous data mining tasks, this thesis will focus on the task of association rule mining (ARM), which has been extensively used with the aim of describing interesting relationships between variables in a dataset. Agrawal et al. (1993) introduced Frequent Itemset Mining (FIM) problem as part of association of rule discovery. An *itemset* is a collection of related items that occur together in a given dataset. The research was initially motivated by the so-called market basket data, in which the goal is to find regularities in the customer behaviour in terms of combinations of products that are often purchased together. In addition to market basket analysis, association rules have been employed in many other areas including bioinformatics, which involve high dimensional data. In bioinformatics, association rules can be used to identify differentially expressed genes (Chen et. al, 2015) (Mallik et. al., 2015), candidate genotype variants related to pharmacogenomics of drugs (i.e. the study of how genes affect a person's response to drugs) (Agapito et. al., 2019), and extracting rules from Single Nucleotide Polymorphisms (SNPs) data (Boutorh and Guessoum, 2016).

However, in the case of high dimensional datasets, several issues have been identified. First, many rules generated from frequent itemsets are irrelevant and redundant, thus are not useful. To overcome this problem, the concept of closed frequent itemset (CFI) has been introduced by Pasquier et al., (1999), which resulted to less number of mined frequent itemsets and more compact association rules. Since then, closed itemsets have been opted in several mining task including colossal closed itemsets (Nguyen et. al., 2017) (Vanahalli and Patil, 2019), frequent closed sequences (Tran et. al., 2015), and closed high-utility-itemset (Dam et. al., 2019).

Second, many relevant rules that have high-quality metrics exist at low frequency level. From a medical point of view, rules with high confidence are more reliable, but unfortunately, they are infrequent (Ordonez et. al., 2006) (Zhao et. al., 2015). Therefore, mining frequent itemsets becomes impractical as many interesting rules exist at the lower end of the threshold, which resulted to many patterns with low support and high confidence being filtered out. To overcome this problem, several alternative thresholds have been proposed to replace the support-based method. This includes confidence measure (McIntosh and Chawla, 2007), core ratio (Zhu et al., 2007) (Nguyen et. al., 2017), and minimum cardinality (Zulkurnain et. al., 2012) (Sohrabi et. al., 2012).

Third, frequent itemsets mining algorithm generates a large set of rules due to high-dimensionality of a dataset. This has not only lead to high computational cost, but make the rules harder to be interpreted and implemented. Therefore, several rule summarization methods have been proposed to compress large set of rules such that the original rule sets can be recovered with minimal loss of information (Simon et. al., 2015) (Sorte et. al., 2018). However, the need for rule summarization added to the processing time, where it is possible to summarize the rule during the mining process itself. Consider two rules R_1: $X_1 \Rightarrow Y$ and R_2: $X_2 \Rightarrow Y$, where $X_1 \subseteq X_2$, the first rule is simpler and is likely to have higher support. However, the second rule may be more interesting as it summarizes several rules similar to the first rule. Therefore, R_2 covers R_1 and R_1 can be excluded. It can be seen that a concise summary is produced when several rules have been covered by R_2. Furthermore, in bioinformatics, patterns that are larger in pattern size, e.g. X_2, give more important meaning than shorter ones, e.g. X_1. Large size patterns can be called colossal pattern, to differentiate from patterns that have large support. In order to mine colossal patterns, a minimum cardinality threshold can be used instead of minimum support threshold, which known to have a performance bottleneck to mine colossal itemset (Zhu et al., 2007) (Simon et. al., 2015).

Based on the issues represented above, the strategic approach to mining high dimensional data can be concluded as follows:

1. Mining closed itemsets instead of all itemsets
2. Support threshold will not be used to keep the interesting rare itemsets
3. Minimum cardinality threshold will be used to mine colossal itemsets

With these three criteria, an algorithm can be designed to mine colossal closed itemset that satisfy minimum cardinality threshold. Existing algorithms have either followed the techniques of candidate generation in *Apriori* (Agrawal et al., 1993) or mining from a tree data structure as in *FP-growth* (Han et al., 2000). *Apriori* traverses the column enumeration tree using a bottom-up search in breadth-first order. This means that each level of the tree must be fully explored to discover frequent itemsets before moving on to the next level. However, bottom-up row enumeration tree is more suitable in mining high dimensional data as the number of rows are much less compared to the columns (Pan et al., 2003) (Cong et al., 2004a). In terms of search order, several algorithms have been evaluated against *Apriori* and it is concluded that depth-first order would result in the most aggressive pruning of the search space and requires the least amount of memory (Zaki and Hsiao, 2005; Soulet and Rioult, 2014; Tomovic and Stanisic, 2015). Though there are several breadth-first algorithms with improved performance than *Apriori*, none have compared their performance against depth-first algorithms (Shah, 2016) (Darrab and Ergenc, 2017) (Dou et. al., 2018) (Sinthuja et. al., 2019).

The obvious weakness of *Apriori* is that it requires multiple database scans to discover each level of candidate itemsets. This is improved in *FP-growth* by constructing a compact tree data structure called FP-tree, which is a compressed representation of all the transactions in the database. Thus, mining process is done on the tree, avoiding the need to repeatedly scan the database. Since then, several database representations have been proposed including vertical data format (Zaki and Hsiao, 2005), binary format (Besson et al., 2005), and bitmap compression (Burdick et. al., 2005). ARM algorithms search the closed itemsets by generating candidates based on row/column enumeration tree (or mining the FP-tree for algorithms based on *FP-growth*). For each node of the tree, an internal data structure is used to store its information, which can be itemset, rowset, and frequency or cardinality. This information is used to check whether the itemset is closed and will be kept in the main memory for duplicates checking. Among the data structure used are transposed tables in CARPENTER (Pan et al., 2003), *diffset* in CHARM (Zaki and Hsiao, 2005), and prefix tree in IsTa (Borgelt et. al., 2011), *DisClose* (Zulkurnain et. al., 2012), HDminer (Xu and Ji, 2016), and *skipping FP-tree* (Nishina et.al., 2019). An efficient algorithm is highly dependent on the pruning strategy to limit the search space, the simplicity of closure checking method, and its efficiency to avoid duplicates in the output. Though algorithms with internal data structure have efficiently avoiding duplicates in the output, they are computationally expensive and their performances are severely degraded with high dimensional data.

The discovery processes in this mining task are twofold; mining the itemsets and generating the association rules. Since the later task is very straightforward and computationally inexpensive, most of the research focus has been on improving the itemset mining process. In this study, the method to mine colossal closed itemsets from high dimensional data is explored. The focus is on the traversal of the search tree; i.e. depth-first and breadth-first, pruning strategy, the closure checking method, and the data structure to store intermediate results. As more applications in bioinformatics require knowledge discovery using ARM, it is important to understand the challenges in order to design better algorithms.

1.2 PROBLEM STATEMENT

Mining colossal closed itemsets from high dimensional data requires different approaches from mining frequent itemsets. In terms of search order, most algorithms have implemented depth-first search as it is claimed to allow the most aggressive pruning of the search space and requires the least amount of memory. However, this is not always true as it is highly dependent on the data structure to store intermediate results, the pruning strategies applied, and the devised closure checking method. The use of complex internal data structures to store intermediate results while enumerating the search space has added to the computational complexity. This is worsening when the search is in depth-first order, where many itemsets of the parent nodes need to be stored. Existing algorithms focused on designing scalable data structures to store and compare all closed itemsets to avoid redundancy in the output. However, the memory usage increases as the cardinality of closed itemsets increases. Therefore, a method which does not require the storage of all colossal closed itemsets will result to a more efficient algorithm. In addition, closure checking method that is combined with the right search order and pruning strategies will reduce the number of intermediate results. The huge number of items in high dimensional dataset makes bottom-up row-enumeration search is more suitable than column enumeration. As each closed itemset corresponds to a unique set of rows, they are ensured to be found by enumerating all combinations of rows (nodes) in the search tree, as long as the threshold is satisfied. However, it is known that not all nodes lead to closed itemsets, hence enumerating them added to computational complexity. The pruning strategies proposed in some algorithms has resulted to incomplete set of discovered closed itemsets in the output, risking the possibility of losing important knowledge from the dataset. Therefore, devising a method that can efficiently skip unnecessary nodes of the row enumeration tree, without missing any closed itemsets in the output will significantly reduce the run time and enhances the scalability of an algorithm.

1.3 RESEARCH OBJECTIVE

In this study, two main research objectives are addressed which include:

1. To propose a method of traversing the search space with effective pruning strategy based on the search order.
2. To propose a method of traversing the relevant nodes without missing any closed itemsets in the output.

1.4 RESEARCH PHILOSOPHY

While it is proven that row enumeration search is more suitable than column enumeration in mining colossal closed itemsets, in terms of the search order, there is no performance comparison between depth-first and breadth-first algorithms. This study explored the advantages of searching the itemsets in breadth-first order in reducing the number of intermediate results and the running time. In breadth-first order, the row enumeration tree is visited level-by-level, where the parent node can be removed once the children nodes have been generated. The benefit of removing the nodes that have been visited is a huge memory saving, which contributes to the scalability of the algorithm.

ARM algorithms have rigidly enumerated every nodes of the search tree to discover closed itemsets. While this ensured no missing itemsets in the output, exploring the entire search space is time consuming. Apart from the imposed threshold, many pruning strategies have been proposed. However, based on literature, it is shown that there is a relationship between a closed itemset with a specific sequence order of items, which can minimize the number of closure computation. Therefore, it is important to study its implementation in discovering colossal closed itemsets.

1.5 RESEARCH SCOPE

This study focus on designing an efficient algorithm to mine colossal closed itemsets. The performance and correctness of the proposed algorithm are evaluated against well-known algorithms; i.e. CARPENTER and IsTa. The algorithms used minimum support threshold and enumerate the search space in depth-first order and stored intermediate results in tables and prefix tree, respectively. Low density synthetic dataset is used in order to make a direct comparison, as the benchmark algorithms failed to complete mining task on high density real dataset. Thus, for real dataset, the experimental results of the proposed algorithms are presented with a secondary *y*-axis which represents the maximum support of the discovered colossal closed itemsets. Similarly, a secondary *y*-axis is also added to the results of CARPENTER and IsTa which represents the maximum cardinality of the closed frequent itemsets discovered. The synthetic dataset is generated using the tool available online while real dataset is obtained from a published discretized microarray data. The correctness of the proposed algorithm is validated by comparing the number of mined closed itemsets in the output with well-known algorithms, available online; i.e. CHARM and DCI_CLOSED.