UNIVERSITI TEKNOLOGI MARA

FREQUENT ITEMSET MINING USING GRAPH THEORY

MOHAMMAD ARSYAD BIN MOHD YAKOP

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science

Faculty of Computer and Mathematical Sciences

May 2017

ABSTRACT

There is a number of algorithms focusing on frequent itemsets mining (FIM) field, however, some of the problems still require attention, particularly when the mining process involves a high dimensional dataset. The Directed Acyclic Graph in High Dimensional Dataset Mining (DAGHDDM) is a graph-based mining algorithm that represents itemsets in complete graph before FIM takes place. Nevertheless, the creation of the complete graph creates unnecessary edges and make the search space large and affects the overall performance. This research aims to speed up the searching process by creating relevant edges in the graph to reduce the search space by rearranging the items using the common prefix rowset. We proposed a novel frequent itemset mining using a graph theory called Frequent Row Graph Closed (FRG-Closed). Designing the FRG-Closed involves new data structure creation known as Frequent Row Graph or FR-Graph. The searching process in the FR-Graph involves the construction of two methods: getPath and item-merging. Experiments were performed to compare the performance of FRG-Closed and Directed Acyclic Graph in High Dimensional Dataset Mining (DAGHDDM) algorithm. The result of the experiments revealed the FRG-Closed capability to mine the frequent closed itemset faster than its counterpart, DAGHDDM algorithm. Moreover, the FRG-Closed is also able to handle lower minimum support compared to the DAGHDDM for a larger transaction.

ACKNOWLEDGEMENT

Firstly, I wish to thank Allah for giving me the opportunity to further my study in Master degree and for completing this challenging journey successfully. I would first like to thank my thesis supervisor Dr Shuzlina Abdul Rahman of the Computer Science and Mathematics at Universiti Teknologi MARA. The door to Dr Shuzlina office was always open whenever I ran into a trouble spot or had a question about my research or writing. She consistently allowed this thesis to be my own work, but steered me in the right the direction whenever she thought I needed it.

Special thanks to Puan Sofianita Mutalib who is my discussion partner that always gives ideas and suggestion in this research, Prof Dr Hajah Azlinah Mohamed as the second reader of this thesis, and I am gratefully indebted to her for her very valuable comments on this thesis. Not forget to my colleagues where always support and help with advise and when I needed a boost of morale. Deep appreciation to Universiti Teknologi MARA (UiTM) for the research grants under PSI 600-RMI/DANA 5/3/PSI (65/2013) for their financial support.

Finally, I must express my very profound gratitude to my parents for the vision and determination in educating me and to my spouse, Nurbaiti Abd Rahman for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you. *Alhamdulilah.*

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1 OVERVIEW

This thesis demonstrates a research in the field of association rule mining in particular frequent itemset mining. The aim of this research is to construct and enhance the frequent itemset algorithm for a high-dimensional dataset. This chapter presents the main components of the research, which include the research background, the problem description, the research questions and objectives, the scope, and the research significance. This chapter ends by describing the organization of the thesis.

1.2 RESEARCH BACKGROUND

Nowadays, recorded dataset has expanded dramatically since data was captured in various types and ways. This phenomena introduced the 'Big Data' era involving a colossal data collection based on 5V's, which are Volume, Velocity, Variety, Veracity and Value (Fan, Han, & Liu, 2013; Laney, 2001; Yin & Kaynak, 2015). The volume refer to the vast amount of data generated each second. The velocity is refer to the number of data was generated in a time. Giving example the number of transactions happed in each miniutes or number of status updated in social media every hour. The variety is refer to the different types of data that are available today such as video, sound, image, sensor etc. The veracity refer to the quality of the available data. Lastly the value, where is refer to the how valuable the available data. The massive volume of data as a result posed a problem in analyzing them for understanding. Analysis of large amount of data could be carried out for diverse purposes requiring a variety of tasks. There are two characteristics of large data: high-dimensionality (large number of items) and a large number of records (rows), reflecting the first "V" which is data volume. Based on the previous study, the high dimensional data is relate with the data that have large number of dimension/item and small number of records. Example of the high dimensional data is microarrays, time series data in financial and neuroimaging.