A Study of Frequent Pattern Mining in Transaction Datasets

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science

at Massey University, Palmerston North,
New Zealand.

Luofeng XU

2011

To my family.

Abstract

Within data mining, the efficient discovery of frequent patterns—sets of items that occur together in a dataset—is an important task, particularly in transaction datasets. This thesis develops effective and efficient algorithms for frequent pattern mining, and considers the related problem of how to learn, and utilise, the characteristics of the particular datasets being investigated.

The first problem considered is how to mine frequent closed patterns in dynamic datasets, where updates to the dataset are performed. The standard approach to this problem is to use a standard pattern mining algorithm and simply rerun it on the updated dataset. An alternative method is proposed in this thesis that is significantly more efficient provided that the size of the updates is relatively small.

Following this is an investigation of the pattern support distribution of transaction datasets, which measures the numbers of times each pattern appears within the dataset. The evidence for the pattern support distribution of real retail datasets obeying a power law is investigated using qualitative appraisals and statistical goodness-of-fit tests, and the power law is found to be a good model. Based on this, the thesis demonstrates how to efficiently estimate the pattern support distribution based on sampling techniques, reducing the computational cost of finding this distribution.

The last major contribution of the thesis is to consider novel ways to set the main user-specified parameters of frequent pattern mining, the minimum support, which defines how many times a pattern needs to be seen before it is 'frequent'. This is a critical parameter, and very hard to set without a lot of knowledge of the dataset. A method to enable the user to specify rather looser requirements for what they require from the mining is proposed based on the assumption of a power-law-based pattern support distribution and fuzzy logic techniques.

KEYWORDS: Data mining, Frequent pattern mining, Dynamic transaction dataset, Incremental mining, Pattern support distribution, Power-law relationship, Fuzzy logic

Declaration

This aims to certify that the research carried out for my Doctorial Thesis entitled "A Study of Frequent Pattern Mining in Transaction Datasets" in the School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand is my own work and that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Acknowledgements

If any value of this research can be recognized, much of it is due to the endless support that I received from many people. I would like to take this opportunity to express my appreciations to those people, although here I can only mention a few.

First and foremost, my deep gratitude goes to my supervisors and mentors of Dr. Ruili Wang and A/Prof. Stephen Marsland. I thank them for their continuous encouragement, responsiveness, support, considerations, patience and enthusiasm, and for sharing with me their knowledge and experience. This research comes from the helpful discussions with them. It is needless to say that without them, this research would have been impossible.

My gratitude is also expressed to all academic and general staff in our school for always being helpful since the very beginning of this research. Working with them is always so enjoyable. Many thanks to Massey University for financially supporting my PhD study by offering me Massey University Doctoral Scholarship.

Finally, I gratefully acknowledge that I would not be in my present position without the endless love and constant support of my family. Moreover, I would like to thank all of my friends as well, for their presence at times when I most needed them.

I present this research to all of those people and let my warm and sincere blessings accompany them wherever they are!

Greatest thanks to everyone!

Table of Contents

List of Tables

List of Figures