

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# 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

Copyright is owned by the author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the author.



# 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

<b>Abstract</b>	<b>v</b>
<b>Declaration</b>	<b>vii</b>
<b>Copyright</b>	<b>ix</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>List of Tables</b>	<b>xx</b>
<b>List of Figures</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Data Mining . . . . .	1
1.1.1 Frequent Pattern Mining . . . . .	3
1.1.2 Overview of the Research Area . . . . .	4
1.2 Aims and Objectives . . . . .	6
1.2.1 Aims . . . . .	6
1.2.2 Objectives . . . . .	7
1.3 Main Contributions . . . . .	9
1.4 Thesis Outline . . . . .	11
<b>2 Incremental Mining of Frequent Closed Patterns</b>	<b>15</b>
2.1 Frequent Pattern Mining . . . . .	16
2.1.1 Dealing with Dynamic Datasets . . . . .	17
2.1.2 Preliminary Concepts and Problem Statement . . . . .	19
2.2 Related Work . . . . .	23
2.2.1 Mining Frequent Closed Patterns in Static Transaction Datasets	23
2.2.2 Mining Frequent Patterns in Dynamic Datasets . . . . .	27

2.3	Algorithm Design and Implementation . . . . .	32
2.3.1	Data structures . . . . .	33
2.3.2	The Pre-processing Procedure . . . . .	35
2.3.3	The Incremental Update Procedure . . . . .	38
2.4	Experimental Evaluation . . . . .	47
2.4.1	Experimental Environment and Performance Parameters . . .	47
2.4.2	Measuring Performance with the Minimum Frequency Value .	50
2.4.3	Measuring Performance with the Type and Size of an Update	51
2.4.4	Measuring Scalability with the Number of Transactions in the Original Dataset . . . . .	54
2.5	Summary . . . . .	57
<b>3</b>	<b>Investigating Power-law Relationships in Pattern Support Distri- butions</b>	<b>59</b>
3.1	Introduction . . . . .	61
3.2	Overview of Power-Laws . . . . .	63
3.2.1	Continuous and Discrete Power-law Distributions . . . . .	63
3.2.2	Special Properties of Power-Law Relationships . . . . .	68
3.2.3	Fitting a Discrete Power-law Distribution to a Set of Empirical Data . . . . .	69
3.2.4	Verification of a Power-law Relationship in a Set of Empirical Data . . . . .	73
3.3	Power-law-based Pattern Support Distributions . . . . .	80
3.3.1	Observations on Pattern Support Distributions . . . . .	81
3.3.2	Fitting Discrete Power-law Distributions to Pattern Support Distributions . . . . .	84
3.3.3	Verification of Power-law Relationships in Pattern Support Distributions . . . . .	91
3.3.4	Observations on the Self-similarity Phenomena in Pattern Sup- port Distributions . . . . .	102
3.4	Summary . . . . .	104
<b>4</b>	<b>Efficiently Estimating Power-law-based Pattern Support Distribu- tions</b>	<b>107</b>
4.1	Introduction . . . . .	108
4.2	The Influence of Sampling on Pattern Support Distributions . . . . .	109

4.3	Efficient Estimation of Power-law-based Pattern Support Distributions	125
4.4	Summary . . . . .	136
<b>5</b>	<b>Fuzzy Frequent Pattern Mining</b>	<b>139</b>
5.1	Introduction . . . . .	140
5.2	Fundamental Ideas . . . . .	141
5.3	Overview of Fuzzy Logic Control . . . . .	144
5.4	Related Work . . . . .	148
5.5	Algorithm Design and Implementation . . . . .	149
5.5.1	Design of a New Fuzzy Logic Controller . . . . .	149
5.5.2	Approximate Estimation of Power-law-based Pattern Support Distributions and their cumulative distributions . . . . .	157
5.5.3	Implementation . . . . .	159
5.6	Experimental Evaluation . . . . .	160
5.7	Summary . . . . .	167
<b>6</b>	<b>Summary and Conclusions</b>	<b>169</b>
6.1	An Overview of the Research . . . . .	169
6.2	Mapping Achievements to Aims and Objectives . . . . .	171
6.3	Open Questions and Future Work . . . . .	173
	<b>Bibliography</b>	<b>175</b>



# List of Tables

1.1	A sample transaction dataset for a grocery store . . . . .	5
2.1	A transaction dataset with ordered frequent items . . . . .	25
2.2	A vertical-format-based transaction dataset . . . . .	27
2.3	An updated transaction dataset with ordered items . . . . .	41
2.4	Parameter settings . . . . .	49
2.5	The corresponding standard deviation values of the results shown in Figure 2.16 . . . . .	53
2.6	The corresponding standard deviation values of the results shown in Figure 2.17 . . . . .	53
2.7	The corresponding standard deviation values of the results shown in Figure 2.18 . . . . .	56
3.1	Parameters of the five real transaction datasets used . . . . .	81
3.2	The support threshold values for mining patterns in the five real trans- action datasets . . . . .	82
3.3	Basic parameters of the pattern support value sets used by the fitting procedure, where $n$ is the total number of pattern support values in a dataset, $\text{mean}(sup_n)$ is the mean of the pattern support values in a dataset, $\text{stdev}(sup_n)$ is the standard deviation of the pattern support value in a dataset, and $\text{max}(sup_n)$ is the maximum value of the pattern support values in a dataset . . . . .	86

3.4 The best-fit discrete power-law distributions to the pattern support value sets in Table 3.3, where  $\hat{x}_{\min}$  is the estimate of the lower bound parameter  $x_{\min}$  of the best-fit power-law distribution,  $\hat{\alpha}$  is the estimate of the scaling exponent parameter  $\alpha$  of the best-fit power-law distribution,  $\hat{\sigma}_1$  and  $\hat{\sigma}_2$  are the standard deviations of the estimate of  $\alpha$  of the best-fit discrete power-law distribution based on Equations 3.24 and 3.25 respectively with an assumption that  $\hat{x}_{\min}$  is equal to  $x_{\min}$  required by Equations 3.24 and 3.25,  $m$  is the corresponding number of the mined pattern support values that are in the interval  $[\hat{x}_{\min}, +\infty)$ , and  $\Lambda$  is the natural logarithm of the corresponding maximum likelihood computed based on  $\hat{x}_{\min}$  and  $\hat{\alpha}$  . . . . . 87

3.5 The computed uncertainty on the estimates of the parameters of the best-fit discrete power-law distributions to the pattern support value sets in Table 3.3 with 1,000 repetitions, except the Accidents dataset with 100 repetitions and the Pumsb dataset with 500 repetitions due to the very expensive computational costs . . . . . 90

3.6 Quantitative goodness-of-fit test results of power-law behaviour in the five studied pattern support distributions, where  $p_1$  is the empirical  $p$ -value for the power-law distribution best fitted to a pattern support distribution,  $D$ -statistic is the maximum difference between the cumulative distribution of a pattern support distribution and the one of the best-fit power-law distribution to the pattern support distribution, LR is the log likelihood ratio of the best-fit power distribution to a corresponding competing distribution,  $p_2$  is the statistically significant  $p$ -value for a log likelihood ratio test. Positive values of LR indicate that the power-law distribution is a better fit than a competing distribution when the corresponding  $p_2$  is less than 0.1. . . . . 94

3.7 The best-fit discrete power-law distributions to the pattern support distributions found from the five real transaction datasets when  $\hat{x}_{\min} = \text{min\_sup}_1$  . . . . . 97

3.8 The computed uncertainty on the best-fit discrete power-law distributions to the pattern support distributions in the corresponding population . . . . . 98

3.9	Quantitative goodness-of-fit test results of power-law behaviours in the pattern support distributions with the maximum $D$ -statistic value over thirty bootstrapped datasets of the Retail and BMS-POS datasets	99
3.10	Parameters of the 3C_chain and Book datasets and the scaling exponents of their best-fit power-law distributions based on linear regression	100
3.11	The similarity of the values of the scaling exponents of the best-fit power-law distributions for the 4 real retail transaction datasets	101
4.1	The best-fit discrete power-law distributions to the pattern support distributions of the corresponding thirty samples drawn from the Retail/BMS-POS dataset with the sampling ratios 70% and 30% respectively	122
4.2	Parameters of the synthetic transaction dataset	132
4.3	Parameters of the best-fit power-law distribution to the full pattern support distribution in the synthetic transaction dataset based on MLE	132
4.4	Parameters of the best-fit power-law distribution to the full item support distribution in the synthetic transaction dataset based on MLE	132
4.5	Parameters of the sample dataset drawn from the synthetic transaction dataset	133
4.6	Parameters of the best-fit power-law distribution to the full pattern support distribution in the sample dataset of the synthetic transaction dataset based on MLE	133
4.7	Parameters of the best-fit power-law distribution to the full item support distribution in the sample dataset of the synthetic transaction dataset based on MLE	134
5.1	A table of fuzzy rules	155
5.2	Some experimental results with respect to the Retail and BMS-POS datasets to show the inconvenience when specifying minimum support values without knowing enough about the target datasets. Note that the value of $min\_sup$ linearly increases in these experiments.	161
5.3	Other experimental results with respect to the Retail and BMS-POS datasets to show the inconvenience when specifying minimum support values without knowing enough about the target datasets. Note that the value of $min\_sup$ exponentially increases in these experiments.	161

5.4	The approximately-estimated values of the parameters of the best-fit power-law-based pattern support distribution and cumulative pattern support distribution of the Retail dataset from its sample datasets by using the methods in Section 5.5.2 . . . . .	162
5.5	The approximately-estimated values of the parameters of the best-fit power-law-based pattern support distribution and cumulative pattern support distribution of the BMS-POS dataset from its sample datasets by using the methods in Section 5.5.2 . . . . .	162
5.6	The experimental results with respect to the Retail dataset . . . . .	164
5.7	The experimental results with respect to the BMS-POS dataset . . .	165

# List of Figures

2.1	The rerunning method . . . . .	18
2.2	The incremental method . . . . .	18
2.3	The simple cases of dataset update . . . . .	22
2.4	The FP-tree based on $TD$ in Table 2.1 when $min\_freq = 40\%$ . . . . .	26
2.5	The DB-tree based on $TD$ in Table 2.1 . . . . .	37
2.6	The two-level hash-indexed result tree based on $TD$ in Table 2.1 when $min\_freq = 40\%$ . . . . .	38
2.7	The paths associated with $a$ in the DB-tree given in Figure 2.5 . . . . .	42
2.8	A checked sub-path and its sub-tree in Figure 2.7 . . . . .	42
2.9	Extracting a sub-path and its sub-tree from the paths in Figure 2.7 . . . . .	42
2.10	Inserting the reordered sub-path and its sub-tree back into the paths in Figure 2.9 . . . . .	43
2.11	The paths associated with $f$ in the result tree given in Figure 2.6 . . . . .	43
2.12	The reordered paths associated with $f$ in the result tree given in Figure 2.6 . . . . .	43
2.13	The updated DB-tree based on the updated dataset shown in Table 2.3 . . . . .	48
2.14	The updated result tree based on the updated dataset shown in Ta- ble 2.3 when $min\_freq = 40\%$ . . . . .	48
2.15	The relationship between the run time and the minimum frequency value . . . . .	51
2.16	Scalability with the size of an update adding new transactions in the original dataset . . . . .	52
2.17	Scalability with the size of an update deleting existing transactions from the original dataset . . . . .	54
2.18	Scalability with the number of transactions in the original dataset, when the number of added transactions is fixed at 0.1% of the original dataset . . . . .	55

3.1	The behaviour of Hurwitz zeta function $\zeta(\alpha, x)$ , where $\alpha = 2.5$ and $25,000 \leq x \leq 250,000$ . . . . .	67
3.2	The partial pattern support distributions of five real transaction datasets	83
3.2	The partial pattern support distributions of five real transaction datasets (con't) . . . . .	84
3.3	The normalized partial pattern support distributions of five real transaction datasets and their maximum likelihood discrete power-law fits in natural log-log plots . . . . .	88
3.3	The normalized partial pattern support distributions of five real transaction datasets and their maximum likelihood discrete power-law fits in natural log-log plots (con't) . . . . .	89
3.4	The self-similarity phenomena in the pattern support distributions between the Retail/BMS-POS dataset and its samples . . . . .	103
3.5	The self-similarity phenomena in the pattern support distributions of the Retail/BMS-POS dataset with different scales on pattern length .	104
3.6	The similarity phenomena between the pattern support distributions with a certain different length in the Retail dataset . . . . .	105
4.1	The influence of sampling on the PSD in a dataset where the support of each pattern is continually uniformly distributed . . . . .	112
4.2	The influence of sampling on the PSD in a dataset where the support of each pattern is discretely uniformly distributed . . . . .	114
4.3	The influence of sampling on the number of the patterns with a certain absolute support value in a sample dataset . . . . .	115
4.4	Observations on the pattern support distributions of the samples of the Retail and BMS-POS datasets. Note that the axes in these figures are different lengths. . . . .	121
4.5	$\frac{\zeta(\alpha-1)}{\zeta(\alpha)}$ when $\alpha > 2$ . . . . .	128
4.6	The pattern support distribution and item support distribution (ISD) of the synthetic transaction dataset and their corresponding maximum likelihood discrete best power-law fits in the natural log-log plot	133
4.7	The pattern support distribution and item support distribution (ISD) of the sample dataset drawn from the synthetic transaction dataset and their corresponding maximum likelihood discrete best power-law fits in the natural log-log plot . . . . .	134

4.8	The curve of the local minimum values of $f$ in the demonstrating example . . . . .	135
5.1	Basic structure of a fuzzy logic controller . . . . .	146
5.2	Triangular and trapezoidal shaped fuzzy membership functions . . . .	151
5.3	The corresponding membership functions of the fuzzy sets defined in the range of $ref_f$ . . . . .	152
5.4	The corresponding membership functions of the fuzzy sets defined in the range of $ref_s$ , where $m$ is the number of patterns whose support values are expected to be not less than the mean support value $mean(sup_{all})$ of all patterns in a target dataset . . . . .	153
5.5	The corresponding membership functions of the fuzzy sets defined in the range of $min\_sup$ . . . . .	154
5.6	The corresponding membership functions of the fuzzy sets defined in the range of $ref_s$ in the example . . . . .	156
5.7	The corresponding membership functions of the fuzzy sets defined in the range of $min\_sup$ in the example . . . . .	157
5.8	The corresponding membership functions of the fuzzy sets defined in the range of $ref_s$ in the Retail dataset . . . . .	162
5.9	The corresponding membership functions of the fuzzy sets defined in the range of $min\_sup$ in the Retail dataset . . . . .	163
5.10	The corresponding membership functions of the fuzzy sets defined in the range of $ref_s$ in the BMS-POS dataset . . . . .	163
5.11	The corresponding membership functions of the fuzzy sets defined in the range of $min\_sup$ in the BMS-POS dataset . . . . .	164

