A HYBRID MODEL FOR DISCOVERING SIGNIFICANT PATTERNS IN DATA MINING

ZAILANI BIN ABDULLAH

A thesis submitted in
fulfillment of the requirements for the award of the
Doctor of Philosophy.

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

JULY 2012

ABSTRACT

A significant pattern mining is one of the most important researches and a major concern in data mining. The significant patterns are very useful since it can reveal a new dimension of knowledge in certain domain applications. There are three categories of significant patterns named frequent patterns, least patterns and significant least patterns. Typically, these patterns may derive from the absolute frequent patterns or mixed up with the least patterns. In market-basket analysis, frequent patterns are considered as significant patterns and already make a lot of contribution. Frequent Pattern Tree (FP-Tree) is one of the famous data structure to deal with batched frequent patterns but it must rely on the original database. For detecting the exceptional occurrences or events that have a high implication such as unanticipated substances that cause air pollution, unexpected degree programs selected by students, unpredictable motorcycle models preferred by customers; the least patterns are very meaningful as compared to the frequent one. However, in this category of patterns, the generation of standard tree data structure may trigger the memory overflow due to the requirement of lowering the minimum support threshold. Furthermore, the classical support-confidence measure has many limitations such as tricky in choosing the right support-confidence value, misleading interpretation based on support-confidence combination and not scalable enough to deal with significant least patterns. Therefore, to overcome these drawbacks, in this thesis we proposed a Hybrid Model for Discovering Significant Patterns (Hy-DSP) which consist of the combination of Efficient Frequent Pattern Mining Model (EFP-M2), Efficient Least Pattern Mining Model (ELP-M2) and Significant Least Pattern Mining Model (SLP-M2). The proposed model is developed using the latest .NET framework and C# as a programming language. Experiments with the UCI datasets showed that the Hy-DSP which consist of DOSTrieIT and LP-Growth* outperformed the benchmarked CanTree and FP-Growth up to 4.13 times (75.78%)

and 10.37 times (90.31%), respectively, thus verify its efficiency. In fact, the number of patterns produce by the models is also less than the standard measures.

ABSTRAK

Melombong corak yang signifikan dari pangkalan data adalah merupakan satu perkara yang penting dalam komuniti perlombongan data. Corak yang signifikan adalah sangat berguna kerana ia akan menghasilkan ilmu pengetahuan berdimensi baru dalam sesetengah domain aplikasi. Secara umumnya, corak sebegini boleh dihasilkan melalui corak kerap berkepastian atau gabungannya dengan corak jarang. Dalam analisa pasar-bakul yang umum, corak kerap dikategorikan sebagai corak signifikan dan telah membuat pelbagai sumbangan. Corak kerap berpokok (FP-Tree) adalah merupakan salah satu struktur data yang sangat popular bagi mengendalikan perlombongan corak kerap secara jujukan namun ia memerlukan kebergantungan terhadap pangkalan data asal. Bagi mengesan keberlakuan atau kejadian terkecuali yang berimpak tinggi seperti kehadiran bahan diluar ramalan yang menyebabkan pencemaran udara, pemilihan program-program ijazah diluar jangkaan oleh pelajar, kegemaran terhadap model-model motorsikal diluar dari kebiasaan oleh pelanggan; semestinya corak jarang adalah lebih bermakna berbanding dengan corak kerap. Walau bagaimanapun, penghasilan corak ini secara struktur data pokok yang standard akan melimpahkan ingatan komputer disebabkan oleh penetapan sokongan minima yang sangat rendah. Selain daripada itu, rangka kerja sokongan-keyakinan mempunyai banyak kelemahan seperti kerumitan dalam memilih nilai sokongan-keyakinan yang bersesuaian, pentafsiran maklumat yang kurang tepat bagi kombinasi nilai sokongan-keyakinan dan kurang perluasan bagi mengendalikan corak jarang yang signifikan. Oleh yang demikian, bagi mengatasi masalah ini, tesis ini mencadangkan Model Hybrid bagi Mencari Corak yang Signifikan (Hy-DSP) yang terdiri daripada kombinasi Model Corak Kerap yang Efisien (EFP-M2), Model Corak Jarang yang Efisien (ELP-M2) dan Model Corak Jarang yang Signifikan (SLP-M2). Model yang dicadangkan ini dibangunkan dengan menggunakan Rangkakerja .Net dan bahasa pengaturcaraan C#. Eksperimen dengan set data UCI menunjukkan Hy-DSP yang mengandungi DOSTrieIT dan LP-Growth* dapat mengatasi CanTree dan

FP-Growth masing-masing sebanyak 4.13 kali (75.78%) dan 10.37 kali (90.31%), dan ini mengesahkan keefisiensinya. Malahan, jumlah corak yang dihasilkan oleh model-model ini adalah lebih sedikit berbanding dengan penggunaan ukuran yang piawai.

PUBLICATIONS

A fair amount of the materials presented in this thesis has been published in various refereed conference proceedings and journals.

Journal

1. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2012). Mining Highly-Correlated of Least Association Rules using Scalable Trie-based Algorithm. *Journal of Chinese Institute of Engineers***,** 35(5), pp. 547-554. (impact factor 0.225)

2. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2012). Detecting Critical Least Association Rules in Medical Databases. *International Journal of Modern Physics Conference Series, World Scientific*, 9, pp. 464 – 479.

3. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2012). ECLAR: An Enhanced Critical Least Association Rules Model. *International Journal of Data Warehousing and Mining* (impact factor 0.200) (in press).

4. **Abdullah**, **Z**., Herawan, T., Ahmad, N. & Deris, M.M. (2011). Extracting Highly Positive Association Rules from Students' Enrollment Data. *Elsevier - Procedia Social and Behavioral Sciences, 28,* pp. 107–111.

5. **Abdullah**, **Z**., Herawan, T., Ahmad, N. & Deris, M.M. (2011). Mining Significant Association Rules from Educational Data using Critical Relative Support Approach. *Elsevier - Procedia Social and Behavioral Sciences, 28,* pp. 97–101.

6. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2012). Tracing Significant Information using Critical Least Association Rules Model. *International Journal of Innovative Computing and Applications, Inderscience.* (in press)

7. **Abdullah**, **Z**., Saman, M.Y.M., Herawan, T. & Deris, M.M. (2012). Discovering Interesting Association Rules in University Programs Selected by Student. *Procedia Information Technology and Computer Science, AWER.* (in press)

8. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2012). Fast Algorithm for Constructing Incremental Tree Data Structure. *Procedia Information Technology and Computer Science, AWER.* (in press)

9. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. Mining Interesting Association Rules in Manufacturing Industry A Case Study in MODENAS. *IGI-Global - International Journal of Knowledge-Based Organizations.* (accepted)

10. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. Finding Least Association Rules from Application of Undergraduate Programs. *International Journal of Knowledge Society Research*, 2012. (submitted)

11. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. Weighted Patterns Mining Framework. *Elsevier - Journal of Systems and Software.* (submitted)

12. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. & Abawajy, J. T3: An Efficient Model for Constructing Frequent Pattern Tree. *Elsevier - Information Systems*. (submitted)

13. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. A Fast Technique to Find Items Support from Scalable Trie-based Structure. *IGI-Global - International Journal of Software Innovation.* (submitted)

14. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. A Tool for Visualizing Critical Least Association Rules. *IGI-Global - International Journal of 3-D Information Modeling*. (submitted)

15. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. A Scalable Algorithm for Constructing Frequent Pattern Tree. *IGI-Global - International Journal of Intelligent Information Technologies*. (submitted)

16. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. Detecting Critical Least Association Rules in Manufacturing Industry: A case study in MODENAS. *International Journal of Knowledge-Based Organizations*. (submitted)

17. **Abdullah**, **Z**., Herawan, T., Ahmad, N. & Deris, M.M. WLAR-Viz: Weighted Least Association Rules Visualization. Manuscript recommended to be included in *Special Issues* of ICICA 2012.

18. **Abdullah**, **Z**., Herawan, T., Ahmad, N. & Deris, M.M. Scalable technique to discover items support from trie data structure. WLAR-Viz: Weighted Least Association Rules Visualization. Manuscript recommended to be included in *Special Issues* of ICICA 2012.

19. **Abdullah**, **Z**., Herawan, T., Ahmad, N. & Deris, M.M. DFP-Growth: An Efficient Algorithm for Mining Frequent Patterns in Dynamic Database. Manuscript recommended to be included in *Special Issues* of ICICA 2012.

20. **Abdullah, Z.**, Herawan, T., and Deris, M.M. Mining Least Association Rules of Degree Level Programs Selected by Students. Manuscript recommended to be included in *Special Issues* of SIA 2012.

21. **Abdullah, Z.**, Herawan, T., and Deris, M.M. Fast Determination of Items Support Technique from Enhanced Tree Data Structure. Manuscript recommended to be included in *Special Issues* of SIA 2012.

22. **Abdullah, Z.**, Herawan, T., and Deris, M.M. CLAR-Viz: Critical Least Association Rules Visualization. Manuscript recommended to be included in *Special Issues* of ICHCI 2012.

Book Chapter

23. Noraziah, A., **Abdullah, Z**., Herawan, T. & Deris, M.M. (2012). WLAR-Viz: Weighted Least Association Rules Visualization. *In Liu et al. (Eds.): ICICA 2012, LNCS, 7473*. Chengde, China: Springer Verlag. pp. 592-599

24. **Abdullah, Z**., Noraziah, A., Herawan, T. & Deris, M.M. (2012). An Efficient Algorithm for Mining Frequent Patterns in Dynamic Database. *In Liu et al. (Eds.): ICICA 2012, LNCS, 7473*. Chengde, China: Springer Verlag. pp. 51-58

25. Noraziah, A., **Abdullah, Z**., Herawan, T. & Deris, M.M. (2012). Scalable Technique to Discover Items Support from Trie Data Structure. *In Liu et al. (Eds.): ICICA 2012, LNCS, 7473*. Chengde, China: Springer Verlag. pp. 500-507

26. **Abdullah, Z.**, Herawan, T., Ibrahim-Fakharaldien, M.A., and Deris, M.M. Mining Least Association Rules of Degree Level Programs Selected by Students. SIA 2012, LNCS. (accepted)

27. **Abdullah, Z.**, Herawan, T., Ibrahim-Fakharaldien, M.A., and Deris, M.M. Fast Determination of Items Support Technique from Enhanced Tree Data Structure. SIA 2012, LNCS. (accepted)

28. **Abdullah, Z.**, Herawan, T., Ibrahim-Fakharaldien, M.A., and Deris, M.M. CLAR-Viz: Critical Least Association Rules Visualization. ICHCI 2012, LNCS. (accepted)

29. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2011). Visualizing the Construction of Incremental Disorder Trie Itemset Data Structure (DOSTrieIT) for Frequent Pattern Tree (FP-Tree). *In H.B. Zaman et al. (Eds.): IVIC 2011, LNCS, 7066.* Bangi, Malaysia: Springer Verlag. pp. 183–195.

30. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2011). An Alternative Measure for Mining Weighted Least Association Rule and Its Framework. *In J.M. Zain et al. (Eds.): ICSECS 2011, CCIS LNCS, 188(2)*. Kuantan, Malaysia: Springer Verlag. pp. 475–485.

31. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2010). Mining Significant Least Association Rules using Fast SLP-Growth Algorithm. *In T.H. Kim and H. Adeli (Eds.): AST/UCMA/ISA/ACN 2010, LNCS, 6059*. Miyazaki, Japan: Springer-Verlag. pp. 324–336.

32. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. (2010). Scalable Model for Mining CriticalLeast Association Rules. *In Rongbo Zhu et al. (Eds.): ICICA 2010, LNCS, 6377*. Tangshan, China: Springer-Verlag. pp. 509-516.

33. **Abdullah**, **Z**., Herawan, T. & Deris, M.M. CIAR: A Model for Mining Critical Indirect Association Rules. ACIIDS 2012, LNCS. (submitted)

Proceeding

34. **Abdullah, Z.** and Deris, M.M. (2009). Association Rules Mining with Relative Weighted Support. *In proceeding of IIWAS2009*. KL, Malaysia: ACM Press. pp. 505–509.

35. **Abdullah, Z.**, Zulaikha, S. and Deris, M.M. (2009). An Efficient Algorithm for Mining Causality Least Pattern. *In proceeding of ICACTE2009*, 2. Cairo, Egypt: ASME Press. pp. 1103–1110.

Tools (Software)

Abdullah, **Z**., Herawan, T. & Deris, M.M. ECLAR: A Data Mining Tool for Capturing Critical Least Association Rules (C#).

Abdullah, **Z**., Herawan, T. & Deris, M.M. WELAR: A Data Mining Tool for Capturing Weighted Least Association Rules (C#).

Abdullah, **Z**., Herawan, T. & Deris, M.M. SPM: A Hybrid Data Mining Tool for Capturing Significant Patterns (C#)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AFOPF	-	Ascending Frequency Ordered Prefix-Tree
AFPIM	-	Adjusting FP-Tree for Incremental Mining
ARs	-	Association Rules
BIT	-	Batch Incremented Tree
BSM	-	Branch Sorting Method
BtB	-	Break the Barrier Algorithm
CATS	-	Compressed and Arranged Transaction Sequence
ChildItem	-	Child Item
CMC	-	Collective Minimum Count
CRS	-	Critical Relative Support
CurPP	-	Current Prefix Path of FP-Tree
DBMS	-	Database Management System
DCP	-	Downward Closure Property
DCS-Apriori	-	Dynamic Collective Support Apriori
DMS	-	Dynamic Minimum Support Count
DOSTrieIT	-	Disorder Support Trie Itemset
DSIP	-	Disorder-Support Itemset Path
EFP-M2	-	Efficient Frequent Pattern Mining Model
ELP-M2	-	Efficient Least Pattern Mining Model
EPFIM	-	Extending FP-Tree for Incremental Mining
ExtPP	-	Extension Prefix Path of FP-Tree
FAR	-	Frequent Association Rules
FIMI	-	Frequent Itemset Mining Dataset Repository
FOLTA	-	Fast Online Trie Algorithm
FP-Growth*	-	Enhanced FP-Growth Algorithm

FP-Tree	-	Frequent Pattern Tree
FRIT	-	Frequent Itemset
FUFPT	-	Fast Updated FP-Tree
HFIT	-	Highly Frequent Itemset
Hy-DSP	-	Hybrid Model for Discovering Significant Patterns
Hyf-DSP	-	Hybrid Framework for Discovering Significant Patterns
ICFP-growth	-	Improved Conditional Pattern-growth
IR	-	Information Retrieval
KDD	-	Knowledge Discovery in Database
LAR	-	Least Association Rules
LEIT	-	Least Itemset
LFIT	-	Least Frequent Itemset
LP-Growth*	-	Enhanced LP-Growth Algorithm
LP-TA	-	LP-Tree algorithm
LP-Tree	-	Least Pattern Tree
LT3A	-	Least Trie Transformation Technique Algorithm
MG	-	Minimal Generators
MinConf	-	Minimum Confidence
MinSupp	-	Minimum Support
MIS	-	Minimum Item Support
MIS-tree	-	Multiple Item Support tree
MISupp	-	Minimum Itemset Support
mRIs	-	minimal Rare Itemsets
MRR	-	Minimal Rare Generators
MSAA	-	Multiple Support Apriori Algorithm
MSApriori	-	Multiple Support Apriori
MZG	-	Minimal Zero Generators
NewPP	-	New prefix path of FP-Tree
OSIP	-	Ordered-Support Itemset Path
ParentItem	-	Parent Item

RSAA	-	Relative Support Apriori Algorithm
SIWE	-	Single Item without Extension
SLAR	-	Significant Least Association Rules
SLP-M2	-	Significant Least Pattern Mining Model
SLP-M2	-	Significant Least Pattern Mining Model
T3	-	Trie Transformation Technique
T3A	-	T3 Algorithm
Tcp	-	FP-Tree based Correlation Mining Algorithm
USIP	-	Unordered-Support Itemset Path
WAR	-	Weighted Association Rules
WARM	-	Weighted Association Rule Mining Algorithm
WSAR*	-	Weighted Support ARs
WSP	-	Weighted Support

LIST OF APPENDICES

REFERENCES

Adda, M., Wu, L. & Feng, Y. (2007). Rare Itemset Mining. *Proc. of the 6th Conference on Machine Learning and Applications.* Ohio, USA: IEEE Computer Society. pp. 73-80.

Aggarwal, C.C. & Yu, P.S. (1998). A New Framework for Itemset Generation. *Proc. of the 17th Symposium on Principles of Database Systems.* Seattle, WA: ACM Press. pp. 18–24.

Agrawal, R. Imielinski, T. & Swami, A. (1993). Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering, 5(6),* pp. 914 – 925.

Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proc. of the 20th VLDB Conference.* Santiago, Chile: Morgan Kaufmann. pp. 487–499.

Amir, A., Feidman, R. & Kashi, R. (1997). A New and Versatile Method for Association Generation. *Information Systems,* 2, pp. 333–347.

Anad, R., Agrawal, R. & Dhar, J. (2009). Variable Support Based Association Rules Mining. *Proc. of the 33rd Annual IEEE International Computer Software and Application Conference*. Washington, USA: IEEE Computer Society. pp. 25 – 30.

Ashrafi, M.Z., Taniar, D. & Smith, K.A. (2004). ODAM: An Optimized Distributed Association Rule Mining Algorithm, *IEEE Distributed Systems Online, 5(3),* pp. 1-18.

Ashrafi, M.Z., Taniar, D. & Smith, K.A. (2007). Redundant Association Rules Reduction Techniques. *International Journal Of Business Intelligence And Data Mining, 2(1),* 29-63.

Borgelt, C. & Kruse, R. (2002). Induction of Association Rules: Apriori Implementation. *Proc of the 15th Conference on Computational Statistics*. Heidelberg, Germany: Physica-Verlag. pp. 395–400.

Brin, S., Motwani, R. & Silverstein, C. (1997a). Beyond Market Basket: Generalizing Association Rules to Correlations. *Proc. of ACM SIGMOD, International Conference on Management of Data*. New York, USA: ACM Press. pp. 255-264.

Brin, S., Motwani, R., Ullman, J.D. & Tsur, S. (1997b). Dynamic Itemset Counting and Implication Rules for Market Basket Data. *Proc. of ACM SIGMOD, International Conference on Management of Data*. New York, USA: ACM Press. pp. 255-264.

Burdick, D., Calimlim, M. & Gehrke, J. (2001). MAFIA: a maximal frequent item set algorithm for transactional databases. *Proc. of the 17th International Conference on Data Engineering*. Heidelberg, Germany: IEEE Computer Society. pp. 443-452.

Cai, C.H., Fu, A.W.C., Cheng, C.H. & Kwong, W.W. (1998). Mining Association Rules with Weighted Items. *Proc. of the 2nd International Database Engineering and Application Symposium (IDEAS'98)*. Cardiff, UK: IEEE Computer Society. pp. 68–77.

Cheung, W. & Zaïane, O.R. (2003). Incremental Mining of Frequent Patterns without Candidate Generation of Support Constraint. *Proc. of the 7th International Database Engineering and Applications Symposium (IDEAS'03).* Hong Kong, China: IEEE Computer Society. pp. 111-117.

Clark, P. & Boswell, R. (1991). Rule Induction with CN2: Some Recent Improvements. *In EWSL 1991, Lecture Notes in Artificial Intelligent, 482.* Porto, Portugal: Springer-Verlag. pp. 151–163.

Ding, B., and Konig, A.R. (2011). Fast Set Intersection in Memory. *Proc. of the VLDB Endowment, 4(4).* Seattle, USA: VLDB Endowment Inc. pp. 255-266.

Ding, J. (2005). *Efficient Association Rule Mining among Infrequent Items*. University of Illinois at Chicago: Ph.D. Thesis.

Dunham, M.H. (2003). *Data Mining: Introductory and Advanced Topics*. New Jersey, USA: Prentice-Hall.

Fayyad, U., Patesesky-Shapiro, G., Smyth, P. & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* Cambridge, MA: MIT Press.

Geng, L. & Hamilton J.J. (2006). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys, 38(3),* pp. 1-32.

Giannikopoulos, P., Varlamis, I. & Eirinaki, M. (2010). Mining Frequent Generalized Patterns For Web Personalization In The Presence Of Taxonomies. *International Journal of Data Warehousing and Mining, 6(1),* pp. 58-76.

Grahne, G., & Zhu, J. (2003). Efficiently using Prefix-Trees in Mining Frequent Itemsets. *Proc. of the Workshop Frequent Itemset Mining Implementations 2003*. Florida, USA: Citeseer. pp. 123-132.

Han, J., Pei, H. & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *Proc. of the 2000 ACM SIGMOD.* Texas, USA: ACM. pp. 1–12.

Han, J. & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco, USA: Morgan Kaufmann.

Han, J. & Pei, J. (2004). Mining Frequent Pattern without Candidate Itemset Generation: A Frequent Pattern Tree Approach. *Data Mining and Knowledge Discovery, 8,* pp. 53 – 87.

He, Z., Xu, X. & Deng, S. (2005). A FP-Tree Based Approach for Mining Strongly Correlated Pairs without Candidate Generations. *In CIS 2005, Part I, Lecture Note in Artificial Intelligent, 3801*: Springer-Verlag. pp. 735 – 740.

Hipp, J., Guntzer, U. & Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining – A General Survey and Comparison. *Proc. of SIGKDD Explorations, ACM SIGKDD, 2(1).* New York, USA: ACM. pp 58 – 64.

Hong, T-P., Lin, J-W. and We, Y-L. (2008). Incrementally Fast Updated Frequent Pattern Trees. *An International Journal of Expert Systems with Applications, 34(4),* pp. 2424 – 2435.

Hu, Y-H. & Chen, Y-L. (2006). Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Tuning Mechanism. *Decision Support Systems, 42(1),* pp. 1 – 24.

Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J. & Tropsha, A. (2004). Mining Protein Family-Specific Residue Packing Patterns from Protein Structure Graphs. *Proc. of the 8th International Conference on Research in Computational Molecular Biology (RECOMB2004),* California, USA: ACM. pp. 308 – 315.

Huang, X., Wu, J., Zhu S. & Xiong, H. (2010) and COSI-tree: Identification of Cosine Interesting Patterns Based on FP-tree. *Proc. of the International Conference on E-Business Intelligence (ICIBI2010).* Yunan, China: Atlantis Press. pp. 456 – 463.

Huang, Y., Xiong, H., Wu, W., Deng, P. &And Zhang, Z. (2007). Mining Maximal Hyperclique Pattern: A Hybrid Search Strategy. *Information Sciences, 177 (3),* pp. 703-721.

Ibrahim, S.P.S & Chandran, K.R. (2011). Compact Weighted Class Association Rule Mining Using Information Gain. *International Journal of Data Mining & Knowledge Management Process, 1(6),* pp. 1 – 13.

Inokuchi, A., Washio, T. & Motoda, H. (2000). An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data. *Proc. of the Principles of Knowledge Discovery and Data Mining (PKDD2000),* Lyon, France: Springer Berlin Heidelberg. pp. 13 – 23.

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for Clustering Data*. New Jersey, USA: Prentice-Hall.

Ji, L., Zhang, B. and Li., J. (2006). A New Improvement of Apriori Algorithm. *Proc. of International Conference on Computer Intelligence and Security (ICCIS2006)*. Guangzhou, China: Springer-Verlag. pp. 840 – 844.

Jiang, B., Hu, X., Wei, Q., Song, J., Han, C. & Liang, M. (2011). Weak Ratio Rules: A Generalized Boolean Association Rules. *International Journal of Data Warehousing and Mining, 7(3),* pp. 50-87.

Jin, R. & Agrawal, G. (2006). A Systematic Approach for Optimizing Complex Mining Tasks on Multiple Datasets. *Proc. of the 22nd International Conference on Data Engineering.* Boston, USA: IEEE. pp. 1 – 17.

Khan, M.S, Muyeba, M. and Coenen, F. (2008). Weighted Association Rule Mining from Binary and Fuzzy Data. *In ICDM 2008, Lecture Note in Artificial Intelligent 5077,* Leipzig, Germany: Springer-Verlag. pp. 200-212.

Khan, M.S., Muyeba, M.K., Coenen, F., Reid, D. & Tawfik, H. (2011). Finding Associations in Composite Data Sets: The Cfarm Algorithm. *International Journal of Data Warehousing and Mining, 7(3),* pp. 1-29.

Kiran, R.U. & Reddy, P. K. (2009). An Improved Frequent Pattern-Growth Approach to Discover Rare Association Rules. *Proc. of the International Conference on Knowledge Discovery and Information Retrieval (ICKDIR2009)*. Funchal - Madeira, Portugal: INSTICC Press. pp. 43 – 52.

Kiran, R.U. & Reddy, P. K. (2009). An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. *Proc. of IEEE Symposium on Computational Intelligence and Data Mining* (*SCIDM2009*). Nashville, USA: IEEE. pp. 340 – 347.

Koh, J-L. & Shieh, S-F. (2004). An Efficient Approach for Maintaining Association Rules Based on Adjusting FP-Tree Structure. *Proc. of the 2004 Database Systems for Advanced Applications, 2004*, Jeju Island, Korea: Springer-Verlag. pp. 417 – 424.

Koh, Y.S. & Rountree, N. (2005). Finding Sporadic Rules using Apriori-Inverse. *In PAKDD2005, Lecture Notes in Artificial Intelligent 3518*. Hanoi, Vietnam: Springer-Verlag. pp. 97 – 106.

Koh, Y.S., Rountree, N. & O'keefe, R.A (2006). Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse. *International Journal of Data Warehousing and Mining, 2(2),* pp. 38-54.

Koh, Y.S., Pears, R. & Dobbie, G. (2011). Automatic Item Weight Generation for Pattern Mining and Its Application. *International Journal of Data Warehousing and Mining, 7(3),* pp. 30-49.

Leung, C. K-S, Khan Q.I., Li, Z. & Hoque, T. (2007). CanTree: A Canonical-order Tree for Incremental Frequent-Pattern Mining. *Knowledge and Information Systems, 11(3),* pp. 287–311.

Leleu, M., Regotti., C., Boulicaut, J.F. & Euvrard. (2003). GO-SPADE: Mining Sequential Patterns over Databases with Consecutive Repetitions. *In MLDM2003, Lecture Note in Computer Science 2734*. Leipzig, Germany: Springer-Verlag. pp. 293–306.

Li, X., Deng, X. & Tang, S. (2006). A Fast Algorithm for Maintenance of Association Rules in Incremental Databases. *In ADMA2006, Lecture Note in Computer Science, 4093*. Xi'an, China: Springer-Verlag. pp. 56 – 63.

Liu, B., Hsu, W. & Ma, Y. (1999). Mining Association Rules with Multiple Minimum Support. *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD99).* San Diego, USA: ACM. pp. 337 – 341.

Liu, G., Lu, H., Lou, W., Xu, Y. & Yu, J.X. (2004). Efficient Mining of Frequent Patterns using Ascending Frequency Ordered Prefix-Tree. *Data Mining and Knowledge Discovery, 9*, pp. 249 – 274.

Lu, J., Chen, W. & Keech, M. (2010). Graph-Based Modelling of Concurrent Sequential Patterns. *International Journal of Data Warehousing and Mining, 6(2),* pp. 41-58.

Mustafa, M.D., Nabila, N.F., Evans, D.J., Saman, M.Y. & Mamat, A. (2006). Association Rules on Significant Rare Data using Second Support. *International Journal of Computer Mathematics 88 (1),* pp. 69 – 80.

Muyabe, M., Khan M.S. & Coenen, F. (2010). Effective Mining of Weighted Fuzzy Association Rules. in Koh Y.S and Rountree, N (Eds.). *Rare Association Rule Mining and Knowledge Discovery.* Pennsylvania, United States: IGI-Global. pp.47-64

Omiecinski, E.R. (1997). Alternative Interest Measures for Mining Associations in Databases. *IEEE Transactions on Knowledge and Data Engineering, 15(1),* pp. 57 - 69.

Park, J.S., Chen, M. & Yu, P.S. (1995). An Effective Hash Based Algorithm for Mining Association Rules. *Proc. of the ACM SIGMOD International Conference Management of Data (ICMD95).* California, USA: ACM. pp. 175 – 186.

Pei, J., Han, J., Lu, H., Nishio, S., Tang, S. & Yang, D. (2001). Hmine: Hyper-Structure Mining of Frequent Patterns in Large Databases. *Proc. of the IEEE International Conference on Data Mining (ICDM2001).* California, USA: IEEE. pp. 441 - 448.

Piatetsky-Shapiro, G. (1991). Discovery, Analysis and Presentation of Strong Rules. *Knowledge Discovery in Databases*. Cambridge, MA: MIT Press.

Pietracaprina, A. & Zandolin, D. (2003). Mining Frequent Item sets Using Patricia Tries. *Proc. of the IEEE Workshop on Frequent Itemset Mining Implementations (ICDM 2003).* Florida, USA: IEEE. pp. 3 – 14.

Raman, V., Qiao, L., Han, W., Narang, I., Chen, Y-L., Yang, K-H. & Ling, F-L. (2007). Lazy, Adaptive RID-list Intersection, and Its Application to Index Anding. *Proc. of the ACM SIGMOD International Conference on Management of Data (ICMD2007)*. Beijing, China: ACM. pp. 773 – 784.

Sahar, S. & Mansour, Y. (1999). An Empirical Evaluation of Objective Interestingness Criteria. *Proc. of the SPIE Conference on Data Mining and Knowledge Discovery*. Florida, USA: pp. 63–74.

Selvi, C.S.K. & Tamilarasi, A. (2009). Mining Association Rules with Dynamic and Collective Support Thresholds. *International Journal on Open Problems Computational Mathematics, 2(3),* pp. 427–438.

Shenoy, P., Haritsa, J.R., Sudarshan, S., Bhalotia, G., Bawa, M. &Shah, D. (2000). Turbo-charging Vertical Mining of Large Databases. *Proc. of the ACM SIGMOD International Conference on Management of Data (ICMD2000).* Texas, USA: ACM Press. pp. 22 - 23.

Silverstein, C., Brin, S. & Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data Mining and Knowledge Discovery, 2*, pp. 39 – 68.

Sun, K. & Bai, F. (2008), Mining Weighted Association Rules without Preassingned Weight. *IEEE Transaction on Knowledge and Data Engineering, 20 (4)*, pp. 489 – 495.

Szathmary, L. & Napoli, A. (2007). Towards rare Itemset Mining. *Proc. of the ICTAI (1).* Patras, Greece: IEEE. pp. 305 – 312.

Szathmary, L. Valtchev, P. and Napoli, A. (2010). Generating Rare Association Rules Using the Minimal Rare Itemsets Family. *International Journal of Software and Informatics 4(3),* pp. 219 – 238.

Tan, P-N., Kumar, V. & Srivastava, J. (2002). Selecting the Right Interestingness Measure for Association Patterns. *Proc. of the KDD'2002*. Alberta, Canada: ACM Press. pp. 32 – 41.

Tan, P-N., Steinbach, M. & Kumar, V. (2006). *Introduction in Data Mining*. Boston: Addison Wesley.

Tanbeer, S. K., Ahmed, C. F., Jeong, B. S. & Lee Y. K. (2009). Efficient Single-Pass Frequent Pattern Mining using a Prefix-Tree. *Information Science, 279,* pp. 559 – 583.

Tanbeer, S.K., Chowdhury, F.A., Jeong, B.S. & Lee, Y-K (2008). CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining. *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '08). in T. Washio et al. (Eds.), LNAI 5012.* Osaka, Japan: Springer-Verlag. pp. 1022 – 1027.

Taniar, D., Rahayu, W., Lee, V.C.S & Daly, O. (2008). Exception Rules in Association Rule Mining. *Applied Mathematics and Computation, 205(2),* pp. 735 – 750.

Tao, F., Murtagh, F. & Farid, M. (2003). Weighted Association Rule Mining using Weighted Support and Significant Framework. *Proc. of the ACM SIGKDD '03*. Washington, USA: ACM Press. pp. 661 – 666.

Tjioe, H.C. & Taniar, D. (2005). Mining Association Rules In Data Warehouses. *International Journal of Data Warehousing and Mining, 1(3),* pp. 28-62.

Totad, S, G., Geeta, R, B. and Reddy, P.P. (2010). Batch Processing for Incremental FP-Tree Construction. *International Journal of Computer Applications, 5(5),* pp. 28 – 32.

Troiano, L., Scibelli, G. & Birtolo, C. (2009). A Fast Algorithm for Mining Rare Itemset. *Proc. of the 9th International Conference on Intelligent Systems Design and Applications.* Pisa, Italy:IEEE. pp. 149 – 1155.

Tsang, S., Koh, Y.S. and Dobbie, G. (1991). RP-Tree: Rare Pattern Tree Mining. *In DaWaK 2011, Lecture Notes in Computer Science, 6862*. Toulouse, France: Springer-Verlag. pp. 277 – 288.

Tsirogianni, D., Guha, S. & Koudas, N. (2009). Improving the Performance of List Intersection. *Proc. of the VLDB Endowment - PVLDB* 2(1) Seattle, USA: VLDB Endowment Inc. pp. 838 – 849.

Wang, K. & Su, M.Y. (2002). Item Selection by "Hub-Authority" Profit Ranking. *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2002).* New York, USA: pp. 652 – 657.

Wang, K., He, Y. & Han, J. (2003). Pushing Support Constraints into Association Rules Mining. *IEEE Transactions on Knowledge Data Engineering, 15(3),* pp. 642 – 658.

Webb, G.I (2007). Discovering Significant Patterns. *Machine Learning, 68(1),* pp.1 – 33.

Weiss, G.M. (2004). Mining with Rarity: A Unifying Framework. *ACM SIGKDD Explorer News Letter 2004 6(1)*, pp. 7 – 19.

Woon, Y.K., Ng, W.K. & Lim, E.P. (2004). A Support Order Trie for Fast Frequent Itemset Discovery. *IEEE Transactions on Knowledge and Data Engineering, 16(7),* pp. 875 – 879.

Wu, T., Chen, Y. & Han, J. (2010). Re-Examination of Interestingness Measures in Pattern Mining: A Unified Framework. *Data Mining and Knowledge Discovery, 21,* pp. 371 – 397.

Xiong, H., Tan, P.N. & Kumar, V. (2003). Mining Strong Affinity Association Patterns in Datasets with Skewed Support Distribution. *Proc. of the 3rd IEEE International Conference on Data Mining*. Washington, DC: IEEE Computer Society. pp. 387-394.

Yen, S-J., Wang, C-K. & Ouyang, L-Y. (2012). A Search Space Reduced Algorithm for Mining Frequent Patterns. *Journal of Information Science and Engineering, 28,* pp. 177 - 191.

Yun, H., Ha, D., Hwang, B. & Ryu, K.H. (2003). Mining Association Rules on Significant Rare Data using Relative Support. *The Journal of Systems and Software, 67(3),* pp.181 – 191.

Zaki, M.J. & Gouda, K. (2003). Fast Vertical Mining using Diffsets. *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, USA: ACM Press. pp. 326 – 335.

Zaki, M.J & Aggarwal, C.C. (2003). Xrules: An Effective Structural Classifier for XML Data. *Proc. of the 9th ACM SIGKDD Iconference on Knowledge Discovery and Data Mining.* Washington, USA: ACM Press. pp. 316 – 325.

Zaki, M.J. & Hsiao, C.-J. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. *Proc. of the 2nd SIAM International Conference on Data Mining*. Chicago, USA: SIAM. pp. 457 – 473.

Zaki, M.J., Parthasarathy, S., Ogihara, M. & Li, W. (1997). New Algorithms for Fast Discovery of Association Rules. *Proc. of the Third International Conference on Knowledge Discovery and Data Mining*. California, USA: AAAI Press. pp. 283 – 286.

Zhou, L. & Yau, S. (2010). Association Rule and Quantitative Association Rule Mining among Infrequent Items. *in Koh, Y.S. & Rountree, N. (Eds.): Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection.* Pennsylvania, USA: IFI-Global. pp. 15 – 32.

Frequent Itemset Mining Dataset Repository (FIMI). Retrieved June 1, 2011, from http://fimi.ua.ac.be/data/

CHAPTER 1

INTRODUCTION

In this chapter, the background of the research is outlined, followed by problem statements, the objectives and the scopes of the research, process of methodology, contributions and lastly, the thesis organization.

1.1 Research Background

With the rapid development of data collection and storage technology, most of the organization can be easily to accommodate with large volume of data. However, extracting useful and meaningful information is extremely very challenging and quite subjective. Most of the tradition data analysis tools and algorithms are not able to deal with vast amounts of data efficiently. Indeed, some explanations or interpretation cannot be addressed properly by the existing data analysis techniques. As a result, data mining is one of the alternative technologies that merged the traditional data analysis techniques for processing vast amounts of data. It has emerged as among a rapidly growing research field for over the past decade. Data mining is a part and the most influence fields in Knowledge Discovery in Database (KDD) process. It originally roots from machine learning but nowadays becomes the confluence of machine learning, statistics and databases.

In data mining, discovering significant patterns from large data repositories are quite challenging. Efficient data structures, algorithms and measures are among the fundamental components that need to be incorporated in dealing with this problem. Until this recent and based on the past studies, tree-based data structure is a great solution in keeping the massive data. However, by having a good data structure

alone is still not sufficient. Thus, efficient algorithm to extract and generate the desired patterns in a timely manner is becoming a necessity. In addition, those patterns must be also assigned with some measure in an attempt to rank and identify which patterns that is really significant. In more advanced cases, items may have their own value and they are not limited to hold only “1” and “0” values. Therefore, new measures that can easily handle the item with individual weight become a fundamental and must be further explored.

Until recently, designing a complete model that can integrate together the different types of data structures, algorithms and measures are very complicated and nearly unfocused (Zhou and Yau, 2007). Most of the previous models are not designed for integration. In fact, the integration of different models always required additional adoption and reconciliation of the fundamental functionality (Ziegler and Dittrich, 1997). In data mining, the integrated model is very important since it can help the organization to decide which patterns mining problems that they really want to resolve. Therefore, an efficient model for mining significant patterns with integration capability needs to be developed and well experimented.

1.2 Problem Statement

Mining patterns or Association Rules (ARs) from any application domains is considered as one of the important research areas in data mining. It is a basic step in deriving a suitable hypothesis and finding associations among items (parameters or values). For example, the retail transaction is aimed at finding the association between the most frequent items that are bought together. By understanding the customers’ behavior, it can help the management to perform promotional strategies, determine potential buyers, increase profit-sales etc. This pattern is also known as frequent pattern. Apriori (Agrawal *et al*., 1993) was the first algorithm to capture sets of frequently bought products at the stores. Since mining frequent pattern is very useful in market-basket analysis, thus frequent pattern can be classified as significant pattern.

In some cases, frequent pattern is not really the patterns that they are looking for. In fact, co-occurrence of the regular items that appear too frequent could be less meaningful in certain application domains. Therefore, detecting exceptional

occurrences or events, that resultant of a decisive implication is very important as compared to regular or well-known pattern. Moreover, least pattern can also provide new insights and exciting knowledge for further exploration. For example, cancer is a dangerous disease and has been identified as a very common cause of death. The least combination of cancer symptoms that resultant of high implication can provide a very useful insight for doctors. Other examples are to detect the least pattern at nuclear plant for hazard processes, banking industries for fraudulent credit-card, networking systems for intruders or viruses, etc. Indeed, the occurrences of 1% (least transactions) from 100,000 transactions are becoming very interesting and reasonable for further analyzing since it is basically equivalent to 1000 cases. Since mining least pattern is very useful in certain application domain, thus least pattern can be classified as significant pattern.

From the above explanation, all items are assumed to have an equal weight or also known as binary weight (1 or 0). However, in certain cases, some items might hold their own weight. In fact, the weight can be used to represent the importance of the item in the transactional databases such as the price, profit margin, quantity, etc. For instance, in market-basket analysis the manager may keen to find out the patterns with the highest profit margins. Let assume that the profit of selling the smart phone is more than selling the cell phone accessories. Hence, the association between SIM card and smart phone is more significant than the association between SIM card and cell phone accessories. However, without considering the profit margin for each individual item, it is impossible to discover the most significant pattern. Thus, the transactional items should be able to hold their own weight rather than the typical binary value. Since the mining significant least pattern (least pattern with weight) is very useful in certain cases, thus significant least pattern can be classified as significant pattern.

In the first case i.e. frequent pattern, several works have been performed in the past decades. Frequent pattern tree (FP-Tree) (Han *et al.*, 2000) has become one of the great alternative data structures to represent the vast amount of transactions of database in a compressed manner. For further improvement, several variations of constructing or updating the FP-Tree have been proposed and it can be categorized into multiple and single database scans. For the first category and including FP-Tree (Han *et al*., 2000), the related studies are Ascending Frequency Ordered Prefix-Tree (AFOPF) (Liu *et al*., 2004), Adjusting FP-Tree for Incremental Mining (AFPIM)

(Koh *et al*., 2004) and Extending FP-tree for Incremental Mining (EPFIM) (Li *et al*., 2006). The related researches in the second category are Compressed and Arranged Transaction Sequence (CATS) tree (Cheung & Zaïane., 2003), Fast Updated FP-Tree (FUFPT) (Hong *et al*., 2008), Branch Sorting Method (BSM) (Tanbeer *et al.*, 2009) and Batch Incremented Tree (BIT) (Totad *et al*., 2010).

However, there are two major drawbacks encountered from the past studies. First, the construction of FP-Tree is still based on extracting the patterns that fulfills the support threshold from the semi structured of its original databases. Second, if the existing databases are suddenly updated, the current FP-Tree must be rebuilt again from the beginning because of the changes in items and patterns supports. In some research extensions, the structure of FP-Tree will be reorganized with extensive modification operations (deletion and insertion) due to the addition of new transactions into databases. Therefore, computationally extensive in constructing FP-Tree is still an outstanding issue and need to be resolved to ensure efficiency in mining frequent patterns.

In the second case i.e. least pattern, several researches have been carried out to overcome this problem. Therefore, various approaches have been suggested in the literature such as Cfarm Algorithm (Khan et al., 2011), Automatic Item Weight Generation (Koh *et al*., 2011), Weak Ratio Rules (Jiang *et al*., 2011), FGP Algorithm (Giannikopoulos *et al*., 2010), ConSP (Lu *et al*., 2010), Multiple Support-based Approach (Kiran & Reddy, 2009), Non-Coincidental Sporadic Rules (Koh *et al*., 2006), ODAM (Ashrafi *et al*., 2004), Fixed Antecedent and Consequent (Ashrafi *et al*., 2007), Exceptionality Measure (Taniar *et al.*, 2008), (Zhou *et al*., 2010), Apriori-inverse (Koh and Rountree, 2005), Relative Support Apriori Algorithm (Yun *et al*., 2003), Multiple Minimum Support (Liu *et al*., 1999), Pushing Support Constraints (Wang *et al*., 2003), Transactional Co-Occurrence Matrix (Ding, 2005).

Even though there are quite a number of improvements that have been achieved, there are still three major drawbacks that have been encountered. The first two non-trivial costs are contributed by the implementation of Apriori-like algorithm. First, the cost of generating a complete set of candidate itemsets. For *k*-itemset, Apriori will produce up to $2^k - 2$ total candidates. Second, cost of repeatedly scanning the database and check all candidates by pattern matching activities. The last drawback is nearly all of the proposed measures to discover the least patterns are

embedded with standard Apriori-like algorithm. The point is this algorithm is undoubtedly may suffer from the “rare item problem”.

In the third case i.e. significant least pattern, there are quite limited studies that have been conducted until this recent as compared to the first and second category. In this category, the item may carry its own individual weight. Among the famous weighted items measures are Minimal rare itemset (Szathmary *et al*., 2010), Weighted ARs (Cai *et al.*, 1998), Auto-counted Minimum Support (Selvi *et al.*, 2009) and Weighted Association Rule Mining (Tao *et al.*, 2003).

However, from the proposed measures, there are two shortcomings detected from the past literature. First, the restoration of least pattern is very computationally extensive and may generate a huge number of unnecessary patterns if their proposed measure is set close to zero. Second, all of these approaches are based on the standard Apriori-like algorithms which will finally trigger the “rare item problem” during discovering the desired patterns.

1.3 Research Objectives

In order to ensure this research can successfully arrive at the destiny, several research objectives have been critically designed and derived. The main objectives are:

(i) To design a hybrid model which consists of the integration of three different models for efficiently mining frequent, least and significant patterns.

(ii) To implement the proposed hybrid model by developing a prototype using .Net Framework and C# as a programming language.

(iii) To evaluate the hybrid model in the developed prototype in term of efficiencies and significances using real and benchmarked datasets.

1.4 Research Scopes

The scopes of the study refer to the types of the datasets that have been employed in the series of experiments. The main scopes are:

(i) Only three (3) benchmarked datasets from the Frequent Itemset Mining Dataset Repository (FIMI) are employed for the performance evaluations called Retail, T10I4D100K and Mushroom datasets.

(ii) Only three (3) real datasets are used for the significant evaluations called Kuala Lumpur Air Pollution, UMT Student Enrolment and MODENAS Motorcycle Production datasets.

(iii) Only two (2) benchmarked tree data structures are employed for the comparison purposes namely FP-Tree and CanTree.

(iv) Only FP-Growth algorithm is employed as a benchmarked algorithm for performance analysis in the experiments.

1.5 Research Methodology Process

Research methodology defines what the activity of research is, how to proceed, how to measure progress, and what constitutes success. Therefore, several organized stages, purposes, processes and outcomes are precisely and clearly outlined. Figure 1.1 depicts the incremental development paradigm of research methodology process.

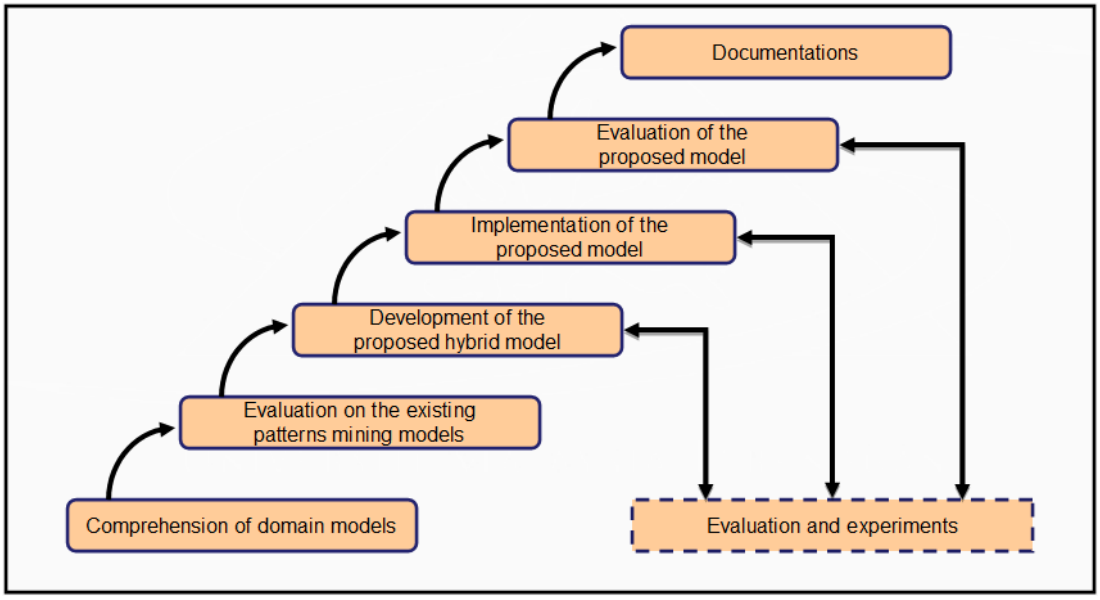

Figure 1.1: Research Methodology Process

1.5.1 Comprehension of Domain Models

(i) *Purpose*: Discovering the related domain models for significant patterns mining models such as Frequent Association Rules (FAR), Least Association Rules (LAR) and Significant Least Association Rules (SLAR).

(ii) *Process*: Reviewing the research papers and relevant documentations from journals, proceeding, thesis and books.

(iii) *Outcome*: Completion of literature review chapter

1.5.2 Evaluation of Existing Patterns Mining Models

(i) *Purpose*: Deep understanding of the current data structures and algorithms for significant pattern mining models.

(ii) *Process*: Implementing the selected data structures and algorithms for comparison purposes in the future.

(iii) *Outcome*: Completion of introduction and literature review chapters, experimental results and a framework for future works.

1.5.3 Development of the purposed hybrid model

(i) *Purpose*: First, designing new models for different categories of significant patterns mining models. Second, designing a hybrid model by integrating all models. Third, fine-tuning and embedding the elements of efficiencies and significances.

(ii) *Process*: Designing new models and a hybrid model which containing the components of efficient algorithms, flexible data structures and novel measures.

(iii) *Outcome*: First, the completion of the proposed prototype which consist of the efficient frequent pattern model, efficient least pattern mining model and significant least pattern mining model. Second, completion of compiling all artifacts in the proposed models into methodology chapter.

1.5.4 Implementation of the proposed hybrid model

(i) *Purpose*: Implementing the proposed hybrid model in the prototype and ensuring workability.

(ii) *Process*: Implementing the proposed hybrid model by converting all algorithms, data structures and measures into C# programming which is running on .Net framework.

(iii) *Outcome*: Completion of the workable prototype to mine different category of the significant patterns.

1.5.5 Evaluation of the proposed hybrid model

(i) *Purpose*: Comparing the performance of the proposed models with other models in term of the efficiencies and significances.

(ii) *Process*: First, evaluating the efficiency of the algorithms and data structures in the proposed models against the algorithms and data structures in the benchmarked models, respectively. Second, analyzing the significances of the generated patterns based on the proposed measures against the other standard measures.

(iii) *Outcome*: First, the completion of the result and discussion chapter based on the experiments. Second, completion of abstract and conclusion chapters which derived from the findings.

1.5.6 Documentation

(i) *Purpose*: Documenting all the research artifacts into proper research articles and thesis.

(ii) *Process*: First, writing conference proceedings and journal papers to prove the novelty of the research efforts. Second, compiling research artifacts into a conclusion chapter.

(iii) *Outcome*: Completion of a thesis, conference proceedings, book chapters and journal papers.

1.6 Thesis Contribution and Organization

Specifically, this thesis makes fives (5) contributions in the field of data mining as follows:

(i) Hybrid Model for Discovering Significant Patterns (Hy-DSP). This model consists of the integration of three different models called Efficient Frequent Pattern Mining Model (EFP-M2), Efficient Least Pattern Mining Model (ELP-M2) and Significant Least Pattern Mining Model (SLP-M2) to efficiently mine frequent pattern, least pattern and significant least pattern, respectively. These patterns are also known as significant pattern. Besides employing the existing new data structures, algorithms and measures from the

current models, Hy-DSP also introduced a new algorithm called Least Trie Transformation Technique Algorithm (LT3A) for integration purposes. In Hy-DSP, three categories of significant patterns based on the specific requirement of the application domains can be easily produced. Experiments with the FIMI datasets showed that Hy-DSP which consist of a new Disorder Support Trie Itemset (DOSTrieIT) data structure and enhanced LP-Growth algorithm (LP-Growth*) outperformed the benchmarked CanTree data structure and FP-Growth algorithm up to 4.13 times (75.78%) and 10.31% times (90.31%), respectively, thus verify its efficiencies.

(ii) EFP-M2 introduces the DOSTrieIT data structure. In this model, the construction of FP-Tree totally relies on the flexibility of DOSTrieIT in handling the updatable database. EFP-M2 is different from typical construction of FP-Tree which is wholly depending on the main source of the original database. A new technique called Trie Transformation Technique (T3) is also employed to transform the data from DOSTrieIT into FP-Tree. Moreover, two new algorithms are directly involved in this model namely Fast Online Trie Algorithm (FOLTA), T3 Algorithm (T3A) and enhanced FP-Growth algorithm (FP-Growth*). The frequent pattern that has been extracted from EFP-M2 is known as FAR and classified as significant pattern. Experiments with the FIMI datasets showed that FP-Growth* algorithm is on average faster at 7,947.27 times (99.34%) than FP-Growth in generating the frequent pattern.

(iii) ELP-M2 efficiently constructs and mine least patterns from its original database. It consists of an enhanced trie data structure called Least Pattern Tree (LP-Tree). The construction of the LP-Tree is derived from the original database. A new algorithm called LP-Tree algorithm (LP-TA) is used to construct LP-Tree data structure. To mine the entire least patterns, LP-Growth algorithm is employed. The least pattern that has been extracted from ELP-M2 is also known as LAR and classified as significant pattern. Experiments with the FIMI datasets showed that LP-Growth algorithm is on average faster at 1.38 times (26.37%) than FP-Growth in producing the least pattern. Moreover, the average number of iterations during constructing LP-Tree is 96.94% lesser than FP-Tree.

(iv) SLP-M2 efficiently constructs and mine significant least patterns from its original database. This model also employs the pervious LP-Tree data structure and previous algorithms namely T3A, LP-TA and LP-Growth. Two extra measures called Critical Relative Support (CRS) and Weighted Support ARs (WSAR*) are introduced in this model. These measures are then applied to the least patterns in an attempt to prune and finally mine the rules that are really significant. The least pattern that satisfies the threshold value from the respective measures is known as SLAR and classified as significant pattern. Experiments with the three real datasets showed that, the average number of SLAR produced by our proposed measures is still lowest as compared to the other standard measures.

(v) Hybrid Framework for Discovering Significant Patterns (Hyf-DSP). This framework is based on the simplification of Hy-DSP. In comparative studies, an asterisk (*) is only added at end of LP-Growth (becomes LP-Growth*) if the data source is referred to DOSTrieIT.

The rest of this thesis is organized as follows: Related work is presented in Chapter 2. The basic terminology of ARs is formulated in Chapter 3. The proposed models and framework are explained in Chapter 4. Comparison tests are reported in Chapter 5. The conclusions and future research directions are mentioned in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

In this section, the basic concepts and related works including KDD, data mining, significant patterns, trie data structure, pattern measures, ARs mining, ARs, frequent pattern algorithms, significant pattern algorithms, FP-Tree, CATS-Tree, Can-Tree, correlation analysis, multiple supports, relative Support, weighted ARs and weighted support ARs are discussed in details.

2.1 Knowledge Discovery in Database

Knowledge Discovery in Database (KDD) is a multi-steps process to discover novel, implicit, previously unknown, and potentially useful of information from database repositories. The term KDD is always used interchangeable with data mining. In fact, KDD is the application of the scientific methods for data mining. Data mining is one of the processes in KDD. Until this recent, several variations to describe the phases or steps performed in the KDD process model have been put forward. The total steps involved can be in the range of 4 until 12. Although the total number of steps may differ, most of the descriptions show consistency in their content. One of the broad descriptions of the KDD process model is introduced by Fayyad *et al.* (1996). According to him, the KDD process model contains five fundamental steps which are problem identification, data extraction, data pre-processing, data mining, and pattern interpretation or presentation (Fayyad *et al.*, 1996). The core part of KDD process is shown in Figure 2.1.

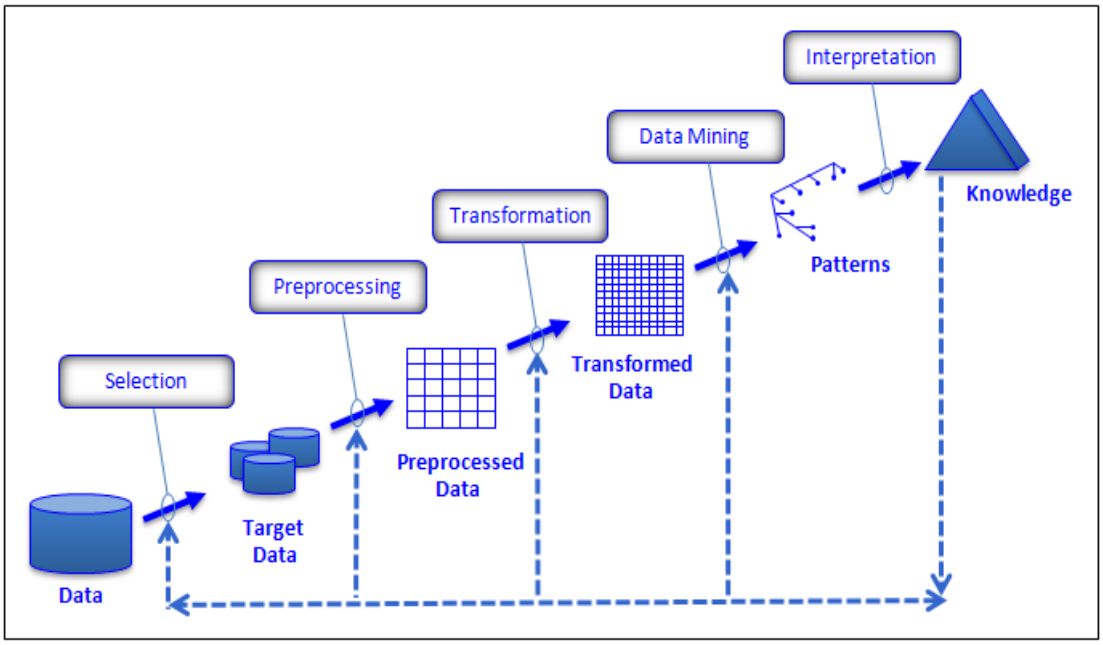

Figure 2.1: An Overview Process of KDD (Fayyad *et al*., 1996)

The focus of the first step in KDD is to understand the domain applications, the prior knowledge and general goal of knowledge discovery. A statement on what to be accomplished is clearly defined. The second step is to select the dataset with appropriate number of attributes for further exploration. It may consist of human experts or KDD tools to help in analyzing the initial data. The third step is to ensure data is valid by removing the outliers or noisy data (data cleaning). It is important to decide the best strategy on how to do about the missing data value. The fourth step is to select the suitable algorithm and technique for data transformation. Some attributes and instance are added or removed from the target data. The fifth step is to present the best model of mined patterns by applying certain data mining algorithms. The sixth step is to examine the model and determine whether it is useful and interesting. It is possible to repeat the previous steps to refine the model.

2.2 Data Mining

Data mining is a process of automatically discover hidden and useful information from large database repositories. It is a core step in KDD process. For more broaden definition, data mining is "*the nontrivial process of identifying valid, novel,*

potentially useful, and ultimately comprehensible knowledge from database" to assist in the decision making process (Fayyad *et al.*, 1996). Data mining techniques are employed to find novel, unknown and useful patterns. It also provides the functionality to predict the future result. Not all of the information discovery tasks are classified as data mining. For example, searching records from the database management system (DBMS) or Internet are the tasks related to information retrieval (IR) area. Undeniable, data mining techniques have been deployed to enhance the ability of information retrieval systems. As mentioned earlier, data mining is part of the KDD process model which transforms the input data into meaning information. These input data can be stored in various formats such as in flat files, spreadsheets, XML, relational tables, etc. Because of the multi-formats and outliers usually occurred in raw data, data pre-processing steps is perhaps the most tedious and time-consuming in the overall KDD process model.

Data mining task can be divided into 2 general categories. The first one is a predictive task. The main objective of this category is to predict the value of the desired attribute by giving the values of others attributes. The desired attribute to be predicted is normally known as targeted (dependent) variable and attributes used in the prediction are commonly known as explanatory (independent) variables. The second part is a descriptive task. The main objective here is to generate patterns based on the relationship between the values of attributes. Usually, this part is always required further exploration and additional post-processing techniques in explaining the obtained results.

Specifically, the cores of data mining tasks are broken up into association analysis, cluster analysis, predictive modelling and anomaly detection. Association analysis refers to the task of discovering the patterns or ARs in the data. Apriori algorithm (Agrawal *et al.*, 1993) was the first widely accepted solution in association analysis. Cluster analysis focuses on clustering the data based on their relatedness. It seeks for identifying groups of similar data points in a multidimensional dataset (Jain & Dubes, 1998). Predictive modelling seeks for forecasting the targeted variables by providing the explanatory (independent) variables. The most two popular types in predictive modelling are classification (discrete target variables) and regression (continuous target variables). Anomaly detection emphasizes more on tracing the abnormal occurrences from the data. It is also known as noisy or outlier data.

2.3 Significant Patterns

Significant patterns (Webb, 2007) can be defined as a set of extracting patterns from data repositories that are potentially very useful and meaningful for certain applications. During the mining processes, certain threshold values will be employed as a filtering mechanism to extract these patterns. Any pattern that fails to satisfy the user-defined thresholds are pruned out and considered as not important or insignificant. Typically, the significances of these patterns highly depend on the problems that need to be resolved. In other words, it is closely related to domain-specific applications. Therefore, for more specific classification, the significant patterns can be divided into three core categories; frequent pattern, least pattern and significant least pattern.

Frequent pattern was first introduced by Agrawal (1993) to mine the ARs between items and also known as market basket analysis. Besides ARs, it also reveals the strong rules, correlation, sequential rules, causality, and many other important discoveries. There are two important reasons of finding frequent patterns from data repositories. First, frequent patterns can effectively summarize the underlying datasets, and provide new information about the data. These patterns can help the domain experts to discover new knowledge hiding in the data. Second, frequent pattern serves as the basic input for others data mining tasks including association rule mining, classification, clustering, and change detection, etc. (Huan *et al.*, 2004; Inokuchi *et al.*, 2000; Jin & Agrawal, 2006; Zaki *et al.*, 2003). In the real world, mining the frequent itemset may involve with the massive datasets and highly pattern dimensions. Therefore, minimizing the computational cost and ensuring the high efficiency in mining activities is very important. Hence, numerous strategies and improvement of data structures have been put forward until this recent.

In some situations, the frequent pattern is not very useful as compared to least pattern. In fact, the least pattern might produce something that is very interesting or meaningful in certain domain applications. However, the process of extracting significant patterns is not straight forward. Usually, special algorithms and measures are required to extract the respective patterns. These rules are very important in discovering rarely occurring but significantly important, such as air pollution detection, critical fault detections, network intrusions etc. and their possible causes.

At the moment, many series of ARs mining algorithms are using the minimum supports-confidence measure to limit the number of ARs. As a result, by increasing or decreasing the minimum support (*MinSupp*) or confidence (*MinConf*) values, the interesting rules might be missing or untraceable. The extraction of least patterns is very challenging for the data mining algorithm (Weiss, 2004). In fact, the standard frequent pattern mining algorithms are inefficient at capturing the least patterns (Koh *et al.*, 2007; Adda *et al.*, 2007; Selvi *et al.*, 2009; Szathmary *et al*., 2010). Since the complexity of the study, difficulties in the algorithms (Yun *et al*., 2003) and it may require excessive computational cost, there are very limited attentions have been paid to discover the significant least patterns. Therefore, designing the new and specific algorithm is undeniable very important to specifically deal with least patterns (Szathmary *et al*., 2010).

In some cases, the least and frequent patterns are not really the interested patterns. The main limitation of the previous two categories of patterns is all items in the transaction are assumed to hold an equal weight (1 or 0). However, in certain domain applications, the items might hold their own weight to represent the importance of the items such as the price, profit margin, quantity etc. Therefore, least patterns with weighted items or also known as significant least pattern is considered as more interesting and useful. The use of weighted items can lead into prioritize the patterns according to their importance rather than typical support and confidence measures (Ibrahim & Chandran, 2011). Among the popular studies in this area are Multiple Support Apriori Algorithm (Liu *et al*., 1999), Relative Support Apriori Algorithm (Yun *et al.*, 2003), Weighted Association Rules (Cai *et al.*, 1998) and Weighted Association Rule Mining (Tao *et al.*, 2003). However, the process of mining significant least pattern is also very challenging (Khan *et al.*, 2008) and facing the similar difficulties as the least pattern.

2.4 Trie Data Structure

Trie data structure (Amir *et al*., 1997; Borgelt & Kruse, 2002; Brin *et al*., 1997; Haans & Tiwary, 1998) is a popular organization structure for keeping data. It emulates a hierarchical data structure with a set of linked nodes. The trie consists of root node, a set of nodes (or vertices) and a set of arcs (or edges). In each node of

prefix sub-trees, it has three fields: item-name, support and node-link. Item-names is the name of items which represented by this node, support is the frequency of this item in the portion of the path that reaching this node, and node-link links to the next node carrying the same item-name or null if there is no node. Besides the root node, other nodes have exactly only one parent. It is possible to reach any node by following a unique path of arcs from the root. If arcs are considered as bidirectional, there is a unique path between any two nodes.

2.5 Pattern Measures

The support-confidence framework is the most commonly used measure in identifying, and consequently defining the strength of ARs. Support is the number of occurrences of some set of attributes (itemsets) in a dataset. It also refers to support count. Confidence is to show the support for an AR in a rule set, which is the level of "confident" of a rule. However, it has many limitations and not really good in determining the desired patterns (Brin, 1997). Hence, several interestingness measures have been proposed (Tan *et al.*, 2006; Wu *et al*., 2010) in the literature to mine the preferred patterns. The challenges are, selecting an appropriate measure is quite troublesome and not straight forward. As a result, Tan (2006) proposed the best approaches for selecting the desired measures based on their list of key properties. These mappings exercise help in identifying which measure that can be the most appropriate for the respective domain applications.

There are two common categories of pattern measures; objective interestingness measure and subjective interestingness measure. The objective interestingness measure used the items or itemsets support in generating the interesting patterns. Examples of the popular objective interestingness measures are support, confidence (Apriori *et al*., 1993), correlation (Brin *et al.*, 1997), IS-measure (Tan, 2006) etc. The subjective interestingness measures employed the values captured from the data and domain users. Examples of subjective interestingness measures are Weighted ARs (Cai *et al.*, 1998), Weighted Association Rule Mining (Tao *et al*., 2003), etc.

2.6 Association Rules Mining

Association Rules (ARs) mining is one of the most important and well researched techniques of data mining. It was first introduced in (Agrawal, *et al*., 1993). Until today, mining of ARs has been extensively studied in the literature (Hipp *et al*., 2000). It aims at discovering correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. (Zhao *et al.*, 2010). Association is a rule, which implies certain association relationships among a set of objects such as occur together or one implies the other (Tan *et al*., 2006). Its main goal is to find associations among items from transactional database.

ARs mining can find an interesting association or correlation relationships among a large set of data items (Han & Kamber, 2001). Usually, ARs are considered to be interesting if they satisfy both a *MinSupp* threshold and a *MinConf* threshold. The most common approach to finding ARs is to break up the problem into two parts (Dunham, 2003):

(i) Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-determined *MinSupp* count (Han and Kamber, 2001).

(ii) Generate strong ARs from the frequent itemsets: By definition, these rules must satisfy *MinSupp* and *MinConf* (Han & Kamber, 2001).

2.7 Association Rules

ARs mining techniques are employed to extract the interesting associations among attributes (items or entities) in a database. ARs are dissimilar to traditional production rules. It can have multiple attributes of input (antecedent) and output (consequence). Moreover, the output attributes for one rule can be an input for another rule. In market-basket analysis, ARs are very popular technique since all possible combinations of product can be generated. Therefore, with a limited number of attributes, it is possible to generate until thousand numbers of ARs.

2.8 Frequent Pattern Algorithms

Apriori (Agrawal *et al.,* 1994) is the first technique to generate the frequent pattern based on generate-and-test strategy. It employs a level-wise searching, where k-itemsets (an itemset that contains *k* items) are used to produce -itemsets. These *k*-itemsets are also known as candidate itemsets. There are two main principles adopted by Apriori. First, every subset of a frequent itemset is also frequent (also called downward closure property). Second, every superset of a non-frequent itemset is also non-frequent. Implementation of these principles enables Apriori to reduce the searching space by pruning out any non-frequent itemsets at early levels. Transactions in the database are repeatedly scanned to match with the patterns appear in candidate itemsets. If there is no further extension of itemsets, the algorithm will terminate immediately.

Apriori is considered as one of the most influential algorithm for mining frequent itemsets for Boolean ARs. However, it suffers from two nontrivial costs (Agrawal *et al.,* 1996). First, the cost of generating the huge number of candidate itemsets. Second, the cost of repeatedly scanned the database and check the vast number of candidates itemset by pattern matching exercise. As an attempt to optimize and increase the Apriori efficiencies, several variations based on Apriori have been proposed such as AprioriTid and Apriori-Hybrid (Agrawal *et al.,* 1994), Dynamic Itemset Counting (Brin *et al.,* 1997), Direct Hashing & Pruning (Park *et al.,* 1995), Partition Algorithm (Hipp *et al.,* 2000), High-Dimension Oriented Apriori (Ji *et al.,* 2006), Variable Support-based Association Rule Mining (Anad *et al.,* 2009), etc.

Due to the limitation in Apriori algorithms, frequent pattern based algorithms without candidate itemsets have been proposed. FP-Growth algorithm uses a combination of the vertical and horizontal database layout to store the database in main memory. This method constructs a compact data structure known as FP-Tree from the original transaction database. The main focus is to avoid cost generation of candidate itemsets, resulting in greater efficiency. All the transactions in the FP-tree are stored in support descending order. By this implementation, the representation of the database in FP-Tree is kept smaller because of the more frequently occurring items are arranged closer to the root, the more likely it to be shared.

However, the main challenge in FP-Growth algorithm is the vast number of conditional pattern trees are recursively generated during the process of mining frequent itemsets. In addition, the algorithm also used different type of traversing during generating and mining the FP-Tree. These processes have definitely increased the computational cost. As a result, several variations of FP-Growth algorithm have been proposed such as FP-Growth algorithms are H-Mine Algorithms (Pei *et al.*, 2001), PatriciaMine (Pietracaprina & Zandolin, 2003), FPgrowth* (Grahne *et al.*, 2003), SOTrieIT (Woon *et al.*, 2004), Tcp (He *et al*., 2005), CFP-Growth (Hu & Chen, 2006), ICFP-Growth (Kiran & Reddy, 2009a), etc.

Besides Apriori-like and frequent pattern based approaches, vertical data format is considered as a new dimension of improving the frequent itemsets mining. Eclat (Zaki *et al.*, 1997) is the first algorithm to find frequent patterns by a depth-first search. It is a method of mining frequent itemsets by transforming the transaction database in the horizontal data format into the vertical data format. Apriori and FP-growth algorithms mine the frequent patterns from typical horizontal data format (i.e., {TID: itemset}), where TID is a transaction-id and itemset is a set of items in the transaction TID. However for ECLAT, frequent patterns can also be mined to data displayed in vertical data format (i.e., {item: TID_set}).

Eclat generates candidate itemsets using only the join step from Apriori, since the itemsets necessary for the prune step are not available. In comparison with Apriori, counting the supports of all itemsets is performed much more efficiently. A few years later, Zaki, Hsiao and Gouda (Zaki & Hsiao, 2002; Zaki & Gouda, 2003) proposed a new approach to efficiently compute the support of an itemset using the vertical database layout. Since the introduction of Eclat, numerous variation of Eclat algorithm have been proposed such as VIPER (Shenoy *et al.*, 2000), MAFIA (Burdic *et al.*, 2001), dEclat (Zaki *et al.*, 2003), etc.

2.9 Significant Pattern Algorithms

Detecting rare and low-rank (also known as infrequent, non-frequent, unusual, exceptional or sporadic) patterns, patterns with low support but high confidence with highly efficient is a difficult task in data mining. This type of patterns cannot be revealed easily using traditional patterns mining algorithms. Usually, to capture these

patterns via traditional approaches, such as the Apriori algorithm, *MinSupp* has to be set very low, which resultants the generation of a large amount of redundant patterns. The problem of discovering rare items has recently captured the interest of the data mining community (Adda *et al*., 2007).

Liu *et al.* (1999) proposed Multiple Support Apriori (MSApriori) based on level wise-search to discover the rare patterns. Here, the user can specify multiple *MinSupp* to reflect different natures and/or frequencies of items. Thus, each item will be associated with a similar of different minimum item support (MIS) value. This model enables users to produce rare item rules without causing frequent items to generate too many meaningless rules. However, the MSApriori algorithm still adopts an Apriori-like candidate set test-and-generate approach and it is always too costly and time consuming. It becomes more critical when there are existed long patterns in datasets.

Selvi *et al.* (2009) introduced Dynamic Collective Support Apriori (DCS-Apriori) to produce an interesting rare ARs by using two auto-counted *MinSupp*. Dynamic Minimum Support Count (DMS) and Collective Minimum Count (CMC) are calculated at every level. In each level onwards, different DMS and CMS values are employed to generate candidate and frequent itemsets, respectively. However, the model is not yet tested using the real dataset and still suffers from candidate itemset generations.

Szathmary *et al.* (2007) proposed two different algorithms to mine the rare itemsets. The first algorithm is a naive one that relies on an Apriori-style enumeration, and the second one is an optimized method that limits the exploration to frequent generators only. As part of algorithms, three types of itemsets are also defined. First, Minimal Generators (MG) are itemsets with a lower support than its subset. Second, Minimal Rare Generators (MRR) are itemset with non-zero support and subsets of all frequent items. Third, Minimal Zero Generators (MZG) are itemset with zero supports and subsets of all non-zero support of items.

Szathmary *et al.* (2010) suggested Break the Barrier (BtB) algorithm to extract highly confidence rare ARs with below the barrier. Three main steps are involved in BtB. First, it finds the minimal rare itemsets (mRIs). mRIs are also a rare generator. Second, to find the closure from the previous mRIs as an attempt to obtain their equivalence classes. Third, to generate the rare ARs from the rare equivalence classes. These rules are also called mRG because their antecedents are

minimal rare generator.

Adda *et al.* (2007) proposed AfRIM that uses a top-down approach which is similar to Rarity. Rare itemset search om AfRIM begins with the itemset that contains all items found in the database. Candidate generation occurs by finding common k-itemset subsets between all combinations of rare k+1-itemset pairs in the previous level. Candidates are pruned in a similar way to Rarity Algorithm. AfRIM examines itemset that have zero support, which may be inefficient.

Koh *et al.* (2005) proposed a novel Apriori-Inverse algorithm to mine the least itemset without generating any frequent rules. Apriori-Inverse uses maximum support instead of typical *MinSupp* to generate candidate itemset. Classification of interested candidate itemset is for those itemset that fall below a maximum support value and above a minimum absolute support value. Two classes of rules are produced from this algorithm known as perfectly sporadic rules and imperfectly sporadic rules. However, the main challenges are it still suffers from too many candidate itemset generations and computational times during generating the least ARs.

Wang *et al.* (2003) proposed Adaptive Apriori to capture the required itemset. Several support constraints are used to each itemset. These constraints exploit the dependency chains between items and determine the “best” *MinSupp* to be pushed to each itemset. In case of more than one constraint is applicable to an itemset, the lowest *MinSupp* is chosen. Adaptive Apriori solved the problem of uniform *MinSupp* as faced by Apriori algorithm while generating the candidate itemset. However, this algorithm still suffers from necessity of scanning multiple times of database for generating the required itemset.

He *et al.* (2005) proposed FP-Tree based Correlation Mining (Tcp) algorithm to mine complete set of significant patterns called jumping emerging pattern (JEP). JEP is a special type of emerging pattern. It is an itemset whose support increases abruptly from zero in one dataset, to non-zero in another dataset. Tcp algorithm involved with constructing FP-Tree and mining the correlation JEP. It constructs a FP-Tree from the transactional database without using the support threshold. In the correlation mining, FP-Growth algorithm is utilized to generate all item pairs and compute their correlation values.

Hu and Chen (2006) introduced CFP-growth algorithm, for mining the complete set of frequent patterns with multiple *MinSupp*. As part of the algorithm,

multiple item support tree (MIS-tree) is also introduced. MIS-tree is an extension of the FP-tree structure (Han, 2004) for storing compressed and crucial information about frequent patterns. As similar to MIS model, each item (node) in MIS-Tree will equip with different MIS value. However, this model can only mine the knowledge under the constraint of single MIS values rather than setting the multiples MIS values.

Kiran & Reddy (2009a) suggested Improved Conditional Pattern-growth (ICFP-growth) which is an extension of FP-growth-like approach to mine rare patterns. ICFP-growth is better than CFP-Growth because this approach can prune the items that are not contributed to produce the desired pattern. ICFP-growth is equipped with various heuristics to efficiently minimize the search space for finding the complete set of rare frequent patterns. The notion of "support different" is also proposed as a mechanism to ensure efficiency in mining rare frequent patterns. Moreover, this approach skips the construction of conditional pattern bases for the suffix items (or patterns) which are infrequent.

2.10 Frequent Pattern Tree

The main bottleneck of the Apriori-like methods is at the candidate itemset generation and test. This problem was overcame by introducing a compact data structure, called frequent pattern tree, or FP-Tree then based on this structure, an FP-Tree-based pattern fragment growth method was developed, FP-growth. FP-Growth is currently one of the benchmarked and the fastest algorithms for frequent pattern mining (Woon *et al.*, 2004). This algorithm is based on a prefix subtree (or paths) representation of the transaction database called FP-Tree. FP-Growth requires two times of scanning the transaction database. First, it scans the database to compute a list of frequent items sorted by descending order and eliminates infrequent items. Second, it scans to compress the database into a FP-Tree structure. Then, the algorithm mines the FP-Tree by recursively building its conditional FP-Tree. A simple example (Tao *et al.*, 2003) of implementing the FP-Tree algorithm is shown in Table 2.2 (Transaction and Ordered Items) and Figure 2.2 (FP-Tree Construction), respectively.

First, the algorithm sorts the frequent items in transactional database and all infrequent items are removed. Let say a *MinSupp* is set to 3, therefore alphabets *f, c, a, b, m, p* are only kept. The algorithm scans the entire transactions start from T1 to T5. In T1, it prunes from {*f, a, c, d, g, i, m, p*} to {*f, c, a, m, p, g*}. Then, the algorithm compresses this transaction into a prefix path tree which f becomes the root. Each path on the tree represents a set of transaction with the same prefix. This process will execute recursively until the end of the transaction. Once the tree has been completely built, then the next pattern mining will be performed.

However, FP-Tree is not suitable for incremental frequent pattern mining. It requires twice database scanning for tree construction and thus very computational excessive. Due to this constraint, several works based on FP-Tree extensions have been established and one of them was CATS-Tree (Cheung & Zaïane, 2003) data structure.

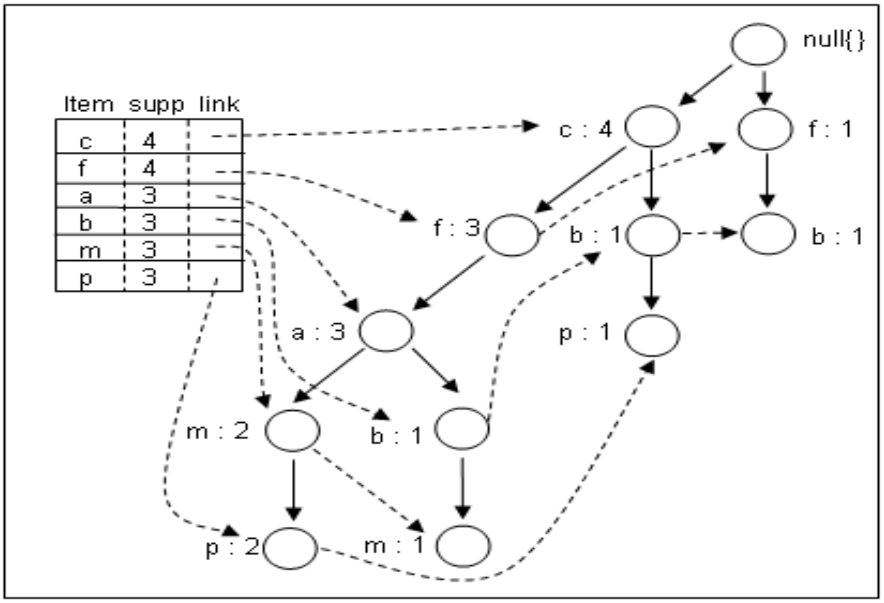

Figure 2.2: FP-Tree construction

Table 2.1: Transaction and ordered items

TID	Items	Items (frequent ordering)
T1	a c f m p	c f a m p
T2	a b c f l m o	c f a b m
T3	b f h j o	f b
T4	b c k s p	c b p
T5	a c e f l p m n	c f a m p

2.11 CATS-Tree

CATS-Tree (Cheung & Zaïane, 2003) is an extension model of FP-Tree to improve the storage compression and allow frequent pattern mining without generation of candidate itemsets. CATS-Tree is a prefix tree and contains all elements of FP-Tree including the header, the item links etc. It requires single database scan to build the complete tree. The construction of CATS-Tree is illustrated from Figure 2.3 until Figure 2.7 based on the sample transactions in Table 2.1.

For the first time, a transaction is inserted at the root with the original items or nodes order. For the next transactions, if the items of the new transactions are similar to the existing nodes, they will be merged together. In certain cases, the position of the existing nodes will be adjusted due to the merging process.

Frequent itemset mining plays a fundamental role in data mining and has been received many attentions in the past decade. More than hundreds of papers have been published in an attempt to increase its efficiencies via enhancement or new algorithm developments. It was first introduced by Agrawal *et al.* (1993) to mine the ARs between items and also known as market basket analysis. Besides ARs, it also reveals the strong rules, correlation, sequential rules, causality, and many other important discoveries. In the real world, mining the frequent itemset may involve with the massive datasets and highly pattern dimensions. Therefore, minimizing the computational cost and ensuring the high efficiency in mining activities is very important. Hence, numerous strategies and improvement of data structures have been put forward until this recent.

In Figure 2.3, the first transaction (T1: *a c f m p*) will be inserted at the root of empty tree. The order of the nodes in the tree is similar to the original order of items in the transaction. At this stage, all nodes have an equal support count, which is 1.

REFERENCES

Adda, M., Wu, L. & Feng, Y. (2007). Rare Itemset Mining. *Proc. of the 6th Conference on Machine Learning and Applications.* Ohio, USA: IEEE Computer Society. pp. 73-80.

Aggarwal, C.C. & Yu, P.S. (1998). A New Framework for Itemset Generation. *Proc. of the 17th Symposium on Principles of Database Systems.* Seattle, WA: ACM Press. pp. 18–24.

Agrawal, R. Imielinski, T. & Swami, A. (1993). Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering, 5(6),* pp. 914 – 925.

Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proc. of the 20th VLDB Conference.* Santiago, Chile: Morgan Kaufmann. pp. 487–499.

Amir, A., Feidman, R. & Kashi, R. (1997). A New and Versatile Method for Association Generation. *Information Systems,* 2, pp. 333–347.

Anad, R., Agrawal, R. & Dhar, J. (2009). Variable Support Based Association Rules Mining. *Proc. of the 33rd Annual IEEE International Computer Software and Application Conference*. Washington, USA: IEEE Computer Society. pp. 25 – 30.

Ashrafi, M.Z., Taniar, D. & Smith, K.A. (2004). ODAM: An Optimized Distributed Association Rule Mining Algorithm, *IEEE Distributed Systems Online, 5(3),* pp. 1-18.

Ashrafi, M.Z., Taniar, D. & Smith, K.A. (2007). Redundant Association Rules Reduction Techniques. *International Journal Of Business Intelligence And Data Mining, 2(1),* 29-63.

Borgelt, C. & Kruse, R. (2002). Induction of Association Rules: Apriori Implementation. *Proc of the 15th Conference on Computational Statistics*. Heidelberg, Germany: Physica-Verlag. pp. 395–400.

Brin, S., Motwani, R. & Silverstein, C. (1997a). Beyond Market Basket: Generalizing Association Rules to Correlations. *Proc. of ACM SIGMOD, International Conference on Management of Data*. New York, USA: ACM Press. pp. 255-264.

Brin, S., Motwani, R., Ullman, J.D. & Tsur, S. (1997b). Dynamic Itemset Counting and Implication Rules for Market Basket Data. *Proc. of ACM SIGMOD, International Conference on Management of Data*. New York, USA: ACM Press. pp. 255-264.

Burdick, D., Calimlim, M. & Gehrke, J. (2001). MAFIA: a maximal frequent item set algorithm for transactional databases. *Proc. of the 17th International Conference on Data Engineering*. Heidelberg, Germany: IEEE Computer Society. pp. 443-452.

Cai, C.H., Fu, A.W.C., Cheng, C.H. & Kwong, W.W. (1998). Mining Association Rules with Weighted Items. *Proc. of the 2nd International Database Engineering and Application Symposium (IDEAS'98)*. Cardiff, UK: IEEE Computer Society. pp. 68–77.

Cheung, W. & Zaïane, O.R. (2003). Incremental Mining of Frequent Patterns without Candidate Generation of Support Constraint. *Proc. of the 7th International Database Engineering and Applications Symposium (IDEAS'03).* Hong Kong, China: IEEE Computer Society. pp. 111-117.

Clark, P. & Boswell, R. (1991). Rule Induction with CN2: Some Recent Improvements. *In EWSL 1991, Lecture Notes in Artificial Intelligent, 482.* Porto, Portugal: Springer-Verlag. pp. 151–163.

Ding, B., and Konig, A.R. (2011). Fast Set Intersection in Memory. *Proc. of the VLDB Endowment, 4(4).* Seattle, USA: VLDB Endowment Inc. pp. 255-266.

Ding, J. (2005). *Efficient Association Rule Mining among Infrequent Items*. University of Illinois at Chicago: Ph.D. Thesis.

Dunham, M.H. (2003). *Data Mining: Introductory and Advanced Topics*. New Jersey, USA: Prentice-Hall.

Fayyad, U., Patesesky-Shapiro, G., Smyth, P. & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* Cambridge, MA: MIT Press.

Geng, L. & Hamilton J.J. (2006). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys, 38(3),* pp. 1-32.

Giannikopoulos, P., Varlamis, I. & Eirinaki, M. (2010). Mining Frequent Generalized Patterns For Web Personalization In The Presence Of Taxonomies. *International Journal of Data Warehousing and Mining, 6(1),* pp. 58-76.

Grahne, G., & Zhu, J. (2003). Efficiently using Prefix-Trees in Mining Frequent Itemsets. *Proc. of the Workshop Frequent Itemset Mining Implementations 2003*. Florida, USA: Citeseer. pp. 123-132.

Han, J., Pei, H. & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *Proc. of the 2000 ACM SIGMOD.* Texas, USA: ACM. pp. 1–12.

Han, J. & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco, USA: Morgan Kaufmann.

Han, J. & Pei, J. (2004). Mining Frequent Pattern without Candidate Itemset Generation: A Frequent Pattern Tree Approach. *Data Mining and Knowledge Discovery, 8,* pp. 53 – 87.

He, Z., Xu, X. & Deng, S. (2005). A FP-Tree Based Approach for Mining Strongly Correlated Pairs without Candidate Generations. *In CIS 2005, Part I, Lecture Note in Artificial Intelligent, 3801*: Springer-Verlag. pp. 735 – 740.

Hipp, J., Guntzer, U. & Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining – A General Survey and Comparison. *Proc. of SIGKDD Explorations, ACM SIGKDD, 2(1).* New York, USA: ACM. pp 58 – 64.

Hong, T-P., Lin, J-W. and We, Y-L. (2008). Incrementally Fast Updated Frequent Pattern Trees. *An International Journal of Expert Systems with Applications, 34(4),* pp. 2424 – 2435.

Hu, Y-H. & Chen, Y-L. (2006). Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Tuning Mechanism. *Decision Support Systems, 42(1),* pp. 1 – 24.

Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J. & Tropsha, A. (2004). Mining Protein Family-Specific Residue Packing Patterns from Protein Structure Graphs. *Proc. of the 8th International Conference on Research in Computational Molecular Biology (RECOMB2004),* California, USA: ACM. pp. 308 – 315.

Huang, X., Wu, J., Zhu S. & Xiong, H. (2010) and COSI-tree: Identification of Cosine Interesting Patterns Based on FP-tree. *Proc. of the International Conference on E-Business Intelligence (ICIBI2010).* Yunan, China: Atlantis Press. pp. 456 – 463.

Huang, Y., Xiong, H., Wu, W., Deng, P. &And Zhang, Z. (2007). Mining Maximal Hyperclique Pattern: A Hybrid Search Strategy. *Information Sciences, 177 (3),* pp. 703-721.

Ibrahim, S.P.S & Chandran, K.R. (2011). Compact Weighted Class Association Rule Mining Using Information Gain. *International Journal of Data Mining & Knowledge Management Process, 1(6),* pp. 1 – 13.

Inokuchi, A., Washio, T. & Motoda, H. (2000). An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data. *Proc. of the Principles of Knowledge Discovery and Data Mining (PKDD2000),* Lyon, France: Springer Berlin Heidelberg. pp. 13 – 23.

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for Clustering Data*. New Jersey, USA: Prentice-Hall.

Ji, L., Zhang, B. and Li., J. (2006). A New Improvement of Apriori Algorithm. *Proc. of International Conference on Computer Intelligence and Security (ICCIS2006)*. Guangzhou, China: Springer-Verlag. pp. 840 – 844.

Jiang, B., Hu, X., Wei, Q., Song, J., Han, C. & Liang, M. (2011). Weak Ratio Rules: A Generalized Boolean Association Rules. *International Journal of Data Warehousing and Mining, 7(3),* pp. 50-87.

Jin, R. & Agrawal, G. (2006). A Systematic Approach for Optimizing Complex Mining Tasks on Multiple Datasets. *Proc. of the 22nd International Conference on Data Engineering.* Boston, USA: IEEE. pp. 1 – 17.

Khan, M.S, Muyeba, M. and Coenen, F. (2008). Weighted Association Rule Mining from Binary and Fuzzy Data. *In ICDM 2008, Lecture Note in Artificial Intelligent 5077,* Leipzig, Germany: Springer-Verlag. pp. 200-212.

Khan, M.S., Muyeba, M.K., Coenen, F., Reid, D. & Tawfik, H. (2011). Finding Associations in Composite Data Sets: The Cfarm Algorithm. *International Journal of Data Warehousing and Mining, 7(3),* pp. 1-29.

Kiran, R.U. & Reddy, P. K. (2009). An Improved Frequent Pattern-Growth Approach to Discover Rare Association Rules. *Proc. of the International Conference on Knowledge Discovery and Information Retrieval (ICKDIR2009)*. Funchal - Madeira, Portugal: INSTICC Press. pp. 43 – 52.

Kiran, R.U. & Reddy, P. K. (2009). An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. *Proc. of IEEE Symposium on Computational Intelligence and Data Mining* (*SCIDM2009*). Nashville, USA: IEEE. pp. 340 – 347.

Koh, J-L. & Shieh, S-F. (2004). An Efficient Approach for Maintaining Association Rules Based on Adjusting FP-Tree Structure. *Proc. of the 2004 Database Systems for Advanced Applications, 2004*, Jeju Island, Korea: Springer-Verlag. pp. 417 – 424.

Koh, Y.S. & Rountree, N. (2005). Finding Sporadic Rules using Apriori-Inverse. *In PAKDD2005, Lecture Notes in Artificial Intelligent 3518*. Hanoi, Vietnam: Springer-Verlag. pp. 97 – 106.

Koh, Y.S., Rountree, N. & O'keefe, R.A (2006). Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse. *International Journal of Data Warehousing and Mining, 2(2),* pp. 38-54.

Koh, Y.S., Pears, R. & Dobbie, G. (2011). Automatic Item Weight Generation for Pattern Mining and Its Application. *International Journal of Data Warehousing and Mining, 7(3),* pp. 30-49.

Leung, C. K-S, Khan Q.I., Li, Z. & Hoque, T. (2007). CanTree: A Canonical-order Tree for Incremental Frequent-Pattern Mining. *Knowledge and Information Systems, 11(3),* pp. 287–311.

Leleu, M., Regotti., C., Boulicaut, J.F. & Euvrard. (2003). GO-SPADE: Mining Sequential Patterns over Databases with Consecutive Repetitions. *In MLDM2003, Lecture Note in Computer Science 2734*. Leipzig, Germany: Springer-Verlag. pp. 293–306.

Li, X., Deng, X. & Tang, S. (2006). A Fast Algorithm for Maintenance of Association Rules in Incremental Databases. *In ADMA2006, Lecture Note in Computer Science, 4093*. Xi'an, China: Springer-Verlag. pp. 56 – 63.

Liu, B., Hsu, W. & Ma, Y. (1999). Mining Association Rules with Multiple Minimum Support. *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD99).* San Diego, USA: ACM. pp. 337 – 341.

Liu, G., Lu, H., Lou, W., Xu, Y. & Yu, J.X. (2004). Efficient Mining of Frequent Patterns using Ascending Frequency Ordered Prefix-Tree. *Data Mining and Knowledge Discovery, 9*, pp. 249 – 274.

Lu, J., Chen, W. & Keech, M. (2010). Graph-Based Modelling of Concurrent Sequential Patterns. *International Journal of Data Warehousing and Mining, 6(2),* pp. 41-58.

Mustafa, M.D., Nabila, N.F., Evans, D.J., Saman, M.Y. & Mamat, A. (2006). Association Rules on Significant Rare Data using Second Support. *International Journal of Computer Mathematics 88 (1),* pp. 69 – 80.

Muyabe, M., Khan M.S. & Coenen, F. (2010). Effective Mining of Weighted Fuzzy Association Rules. in Koh Y.S and Rountree, N (Eds.). *Rare Association Rule Mining and Knowledge Discovery.* Pennsylvania, United States: IGI-Global. pp.47-64

Omiecinski, E.R. (1997). Alternative Interest Measures for Mining Associations in Databases. *IEEE Transactions on Knowledge and Data Engineering, 15(1),* pp. 57 - 69.

Park, J.S., Chen, M. & Yu, P.S. (1995). An Effective Hash Based Algorithm for Mining Association Rules. *Proc. of the ACM SIGMOD International Conference Management of Data (ICMD95).* California, USA: ACM. pp. 175 – 186.

Pei, J., Han, J., Lu, H., Nishio, S., Tang, S. & Yang, D. (2001). Hmine: Hyper-Structure Mining of Frequent Patterns in Large Databases. *Proc. of the IEEE International Conference on Data Mining (ICDM2001).* California, USA: IEEE. pp. 441 - 448.

Piatetsky-Shapiro, G. (1991). Discovery, Analysis and Presentation of Strong Rules. *Knowledge Discovery in Databases*. Cambridge, MA: MIT Press.

Pietracaprina, A. & Zandolin, D. (2003). Mining Frequent Item sets Using Patricia Tries. *Proc. of the IEEE Workshop on Frequent Itemset Mining Implementations (ICDM 2003).* Florida, USA: IEEE. pp. 3 – 14.

Raman, V., Qiao, L., Han, W., Narang, I., Chen, Y-L., Yang, K-H. & Ling, F-L. (2007). Lazy, Adaptive RID-list Intersection, and Its Application to Index Anding. *Proc. of the ACM SIGMOD International Conference on Management of Data (ICMD2007)*. Beijing, China: ACM. pp. 773 – 784.

Sahar, S. & Mansour, Y. (1999). An Empirical Evaluation of Objective Interestingness Criteria. *Proc. of the SPIE Conference on Data Mining and Knowledge Discovery*. Florida, USA: pp. 63–74.

Selvi, C.S.K. & Tamilarasi, A. (2009). Mining Association Rules with Dynamic and Collective Support Thresholds. *International Journal on Open Problems Computational Mathematics, 2(3),* pp. 427–438.

Shenoy, P., Haritsa, J.R., Sudarshan, S., Bhalotia, G., Bawa, M. &Shah, D. (2000). Turbo-charging Vertical Mining of Large Databases. *Proc. of the ACM SIGMOD International Conference on Management of Data (ICMD2000).* Texas, USA: ACM Press. pp. 22 - 23.

Silverstein, C., Brin, S. & Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data Mining and Knowledge Discovery, 2*, pp. 39 – 68.

Sun, K. & Bai, F. (2008), Mining Weighted Association Rules without Preassingned Weight. *IEEE Transaction on Knowledge and Data Engineering, 20 (4)*, pp. 489 – 495.

Szathmary, L. & Napoli, A. (2007). Towards rare Itemset Mining. *Proc. of the ICTAI (1).* Patras, Greece: IEEE. pp. 305 – 312.

Szathmary, L. Valtchev, P. and Napoli, A. (2010). Generating Rare Association Rules Using the Minimal Rare Itemsets Family. *International Journal of Software and Informatics 4(3),* pp. 219 – 238.

Tan, P-N., Kumar, V. & Srivastava, J. (2002). Selecting the Right Interestingness Measure for Association Patterns. *Proc. of the KDD'2002*. Alberta, Canada: ACM Press. pp. 32 – 41.

Tan, P-N., Steinbach, M. & Kumar, V. (2006). *Introduction in Data Mining*. Boston: Addison Wesley.

Tanbeer, S. K., Ahmed, C. F., Jeong, B. S. & Lee Y. K. (2009). Efficient Single-Pass Frequent Pattern Mining using a Prefix-Tree. *Information Science, 279,* pp. 559 – 583.

Tanbeer, S.K., Chowdhury, F.A., Jeong, B.S. & Lee, Y-K (2008). CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining. *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '08). in T. Washio et al. (Eds.), LNAI 5012.* Osaka, Japan: Springer-Verlag. pp. 1022 – 1027.

Taniar, D., Rahayu, W., Lee, V.C.S & Daly, O. (2008). Exception Rules in Association Rule Mining. *Applied Mathematics and Computation, 205(2),* pp. 735 – 750.

Tao, F., Murtagh, F. & Farid, M. (2003). Weighted Association Rule Mining using Weighted Support and Significant Framework. *Proc. of the ACM SIGKDD '03*. Washington, USA: ACM Press. pp. 661 – 666.

Tjioe, H.C. & Taniar, D. (2005). Mining Association Rules In Data Warehouses. *International Journal of Data Warehousing and Mining, 1(3),* pp. 28-62.

Totad, S, G., Geeta, R, B. and Reddy, P.P. (2010). Batch Processing for Incremental FP-Tree Construction. *International Journal of Computer Applications, 5(5),* pp. 28 – 32.

Troiano, L., Scibelli, G. & Birtolo, C. (2009). A Fast Algorithm for Mining Rare Itemset. *Proc. of the 9th International Conference on Intelligent Systems Design and Applications.* Pisa, Italy:IEEE. pp. 149 – 1155.

Tsang, S., Koh, Y.S. and Dobbie, G. (1991). RP-Tree: Rare Pattern Tree Mining. *In DaWaK 2011, Lecture Notes in Computer Science, 6862*. Toulouse, France: Springer-Verlag. pp. 277 – 288.

Tsirogianni, D., Guha, S. & Koudas, N. (2009). Improving the Performance of List Intersection. *Proc. of the VLDB Endowment - PVLDB* 2(1) Seattle, USA: VLDB Endowment Inc. pp. 838 – 849.

Wang, K. & Su, M.Y. (2002). Item Selection by "Hub-Authority" Profit Ranking. *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2002).* New York, USA: pp. 652 – 657.

Wang, K., He, Y. & Han, J. (2003). Pushing Support Constraints into Association Rules Mining. *IEEE Transactions on Knowledge Data Engineering, 15(3),* pp. 642 – 658.

Webb, G.I (2007). Discovering Significant Patterns. *Machine Learning, 68(1),* pp.1 – 33.

Weiss, G.M. (2004). Mining with Rarity: A Unifying Framework. *ACM SIGKDD Explorer News Letter 2004 6(1)*, pp. 7 – 19.

Woon, Y.K., Ng, W.K. & Lim, E.P. (2004). A Support Order Trie for Fast Frequent Itemset Discovery. *IEEE Transactions on Knowledge and Data Engineering, 16(7),* pp. 875 – 879.

Wu, T., Chen, Y. & Han, J. (2010). Re-Examination of Interestingness Measures in Pattern Mining: A Unified Framework. *Data Mining and Knowledge Discovery, 21,* pp. 371 – 397.

Xiong, H., Tan, P.N. & Kumar, V. (2003). Mining Strong Affinity Association Patterns in Datasets with Skewed Support Distribution. *Proc. of the 3rd IEEE International Conference on Data Mining*. Washington, DC: IEEE Computer Society. pp. 387-394.

Yen, S-J., Wang, C-K. & Ouyang, L-Y. (2012). A Search Space Reduced Algorithm for Mining Frequent Patterns. *Journal of Information Science and Engineering, 28,* pp. 177 - 191.

Yun, H., Ha, D., Hwang, B. & Ryu, K.H. (2003). Mining Association Rules on Significant Rare Data using Relative Support. *The Journal of Systems and Software, 67(3),* pp.181 – 191.

Zaki, M.J. & Gouda, K. (2003). Fast Vertical Mining using Diffsets. *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, USA: ACM Press. pp. 326 – 335.

Zaki, M.J & Aggarwal, C.C. (2003). Xrules: An Effective Structural Classifier for XML Data. *Proc. of the 9th ACM SIGKDD Iconference on Knowledge Discovery and Data Mining.* Washington, USA: ACM Press. pp. 316 – 325.

Zaki, M.J. & Hsiao, C.-J. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. *Proc. of the 2nd SIAM International Conference on Data Mining*. Chicago, USA: SIAM. pp. 457 – 473.

Zaki, M.J., Parthasarathy, S., Ogihara, M. & Li, W. (1997). New Algorithms for Fast Discovery of Association Rules. *Proc. of the Third International Conference on Knowledge Discovery and Data Mining*. California, USA: AAAI Press. pp. 283 – 286.

Zhou, L. & Yau, S. (2010). Association Rule and Quantitative Association Rule Mining among Infrequent Items. *in Koh, Y.S. & Rountree, N. (Eds.): Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection.* Pennsylvania, USA: IFI-Global. pp. 15 – 32.

Frequent Itemset Mining Dataset Repository (FIMI). Retrieved June 1, 2011, from http://fimi.ua.ac.be/data/