Ain Shams University
Faculty of Computer & Information Sciences
Computer Science Department

Developing an Association Rules Mining Technique for Large Updated Databases

This thesis is submitted to Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science.

By

Mohammed Mamdouh Mohammed Fouad
M.Sc. of Computer Science (2009),
Faculty of Computer and Information Sciences,
Ain Shams University

Supervised By

Prof. Dr. Mostafa Gadal-Haqq M. Mostafa
Professor of Computer Science
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Ghada Nasr Ali
Lecturer, Computer Science Department
Faculty of Computer and Information Sciences,
Ain Shams University

Cairo – 2016

To my parents, wife and sons.

My Past, Present and Future.

Acknowledgments

Thanks to **Allah** the Most Merciful and the Most Gracious, who helps me in every moment of my life and in completing this study.

I am indebted to many people for their support and advice to the successful completion of my Ph.D. degree and this thesis.

I would like to thank **Prof. Mostafa Gadal-Haqq** for his continuous supervision, direction, assistance, and guidance during my work in this research. His experience guided me through all the steps of this study beginning from the research proposal to the last words in this thesis.

Many thanks for **Dr. Ghada Nasr** for her continuous support, encouragement and guidance during my work in this research.

I would like to express my great appreciation to **my parents**, and to all **my family** for their love, support and encouragement. My sincere thanks go to **my wife** who always encouraged me and helped me with her support. She does her best and creates a suitable environment for me during the period of this research. Finally, I dedicate this thesis to my kids, **Marwan and Hamza**, with love.

Also, I would like to thank my **Prof. Tarek Gharib** for his support during my academic life and all the steps of this research.

Finally, I would like to thank everyone who supported, helped, or encouraged me to finish and complete this work.

Abstract

Association rules discovery is one of the important research areas in the data mining field. It is interested in finding significant relationships between items in any transactional database with respect to given threshold. With the massive amount of data flow every day, these transactional databases are updated with new transactions that make any prior knowledge invalid. The trivial solution is to run the mining algorithms again without utilizing the previously discovered knowledge. This would be a nightmare process especially with the large database sizes nowadays. The need of new efficient algorithms to deal with the incremental mining of association rules becomes very critical.

In this thesis, the researcher is interested in developing a set of efficient algorithms for mining interesting association rules in large incremental databases mainly with two types of databases; traditional and temporal databases. Recently, there is a new interest in weighted databases. Weights give each item another value that defines its significance in the database rather than the traditional support measure. The researcher presented a new algorithm for mining frequent weighted itemsets as a start point in this hot research topic.

Many experiments were conducted using both real and synthetic large datasets to compare the performance of the proposed algorithms with the recently cited algorithms in each topic. The experiments measure the performance with respect to many aspects such as running time, memory usage analysis with different database characteristics. The results show that the proposed algorithms have great performance optimization over other

algorithms. In addition, the results show that the proposed algorithms achieve linear scalability when dealing with large databases even at very low minimum support thresholds.

Keywords: Data Mining, Association Rules Discovery, Large Databases, Temporal Databases, Frequent Weighted Itemsets, Trie Indexing

Publications

Parts of this thesis are published as original papers in the following references:

1. Mohammed M. Fouad, Mostafa G.M. Mostafa, Abulfatah S. Mashat and Tarek F. Gharib. "***wSWF: A Weighted Sliding Window Filtering Algorithm for Frequent Weighted Itemsets Mining***," The International Journal of Innovative Computing, Information and Control (IJICIC), Vol. 11, no. 4, pp. 1421-1440, 2015. [SJR-IF: 1.156]

2. Mohammed M. Fouad, Mostafa G.M. Mostafa, Abulfatah S. Mashat and Tarek F. Gharib. "***IMIDB: An Algorithm for Indexed Mining of Incremental Databases***," ahead of print, The Journal of Intelligent Systems (JISYS), 2015. [SJR-IF: 0.239]

3. Mohammed M. Fouad and Mostafa G.M. Mostafa, "***IndxTAR: An Efficient Algorithm for Indexed Mining of Incremental Temporal Association Rules***," submitted to the Journal of Intelligent Systems (JISYS), 2015. [SJR-IF: 0.239]

4. Mohammed M. Fouad and Mostafa G.M. Mostafa, "***An Efficient Technique for Indexing Temporal Databases***," The International Journal of Computer Applications (IJCA), Vol. 127, no. 13, pp. 32-37, 2015. [IF: 0.715]

Table of Contents

ACKNOWLEDGMENTS ..... **III**

ABSTRACT ..... **IV**

PUBLICATIONS ..... **VI**

TABLE OF CONTENTS ..... **VII**

LIST OF TABLES ..... **X**

LIST OF FIGURES ..... **XII**

LIST OF ABBREVIATIONS ..... **XIV**

CHAPTER 1 : INTRODUCTION **1**

1.1 OVERVIEW 1
1.2 OBJECTIVES 2
1.3 MOTIVATION 3
1.4 CONTRIBUTIONS 3
1.5 THESIS STRUCTURE 4

CHAPTER 2 : BACKGROUND AND RELATED WORK **6**

2.1 DATA MINING 6
2.1.1 Types of Data *7*
2.1.1.1 Relational Databases *8*
2.1.1.2 Transactional Databases *8*
2.1.1.3 Spatial Databases *8*
2.1.1.4 Temporal and Time-Series Databases *9*
2.1.1.5 World Wide Web *9*
2.1.1.6 Weighted and Utility Databases *10*
2.2 ASSOCIATION RULES MINING 11
2.2.1 Level-wise Approach *12*
2.2.2 Pattern Growth Approach *15*
2.2.3 Discussion *19*
2.2.4 Generating Association Rules from Frequent Itemsets *20*
2.3 INCREMENTAL ASSOCIATION RULES MINING 21

2.4 INCREMENTAL TEMPORAL ASSOCIATION RULES MINING 24
2.5 FREQUENT WEIGHTED ITEMSETS MINING 28

CHAPTER 3 : INCREMENTAL ASSOCIATION RULES MINING 32

3.1 INDEXED MINING OF INCREMENTAL DATABASES 32
3.1.1 Trie Index Data Structure 32
3.1.2 The IMIDB Algorithm 34
3.1.3 Illustrative Example 37
3.2 RESULTS AND DISCUSSION 42
3.2.1 Datasets 42
3.2.2 Running Time Analysis 43
3.2.3 Memory Usage 51
3.2.4 Scalability 51
3.3 SUMMARY 53

CHAPTER 4 : INCREMENTAL TEMPORAL ASSOCIATION RULES MINING 55

4.1 MINING INCREMENTAL TEMPORAL ASSOCIATION RULES 55
4.1.1 TIndex Data Structure 55
4.1.2 TIndex2 Data Structure 59
4.1.3 The Proposed IndxTAR Algorithm 60
4.1.4 Illustrative Example 63
4.2 RESULTS AND DISCUSSION 66
4.2.1 Datasets 66
4.2.2 Number of Candidates 67
4.2.3 Memory Usage 68
4.2.4 Running Time 70
4.2.5 Scalability 77
4.3 SUMMARY 78

CHAPTER 5 : FREQUENT WEIGHTED ITEMSETS MINING 80

5.1 PRELIMINARIES 80
5.2 WEIGHTED SLIDING WINDOW FILTERING 83
5.2.1 Sliding Window Filtering 83
5.2.2 Diffset Strategy [60] 84
5.2.3 The Proposed wSWF Algorithm 85

5.2.4 Illustrative Example ... *87*
5.3 RESULTS AND DISCUSSION ... 92
5.3.1 Datasets ... *92*
5.3.2 Runtime and Memory Usage Analysis ... *94*
5.3.3 Number of C2 Candidates ... *103*
5.3.4 Scalability ... *107*
5.3.5 Comparison Summary ... *108*
5.4 SUMMARY ... 109
CHAPTER 6 : CONCLUSIONS AND FUTURE WORK ... 111
6.1 SUMMARY ... 111
6.2 CONCLUSIONS ... 112
6.3 FUTURE WORK ... 114
REFERENCES ... 115
ملخص الرسالة ... 127

List of Tables

TABLE 2.1: APRIORI ALGORITHM – EXAMPLE TRANSACTIONAL DATABASE 13
TABLE 2.2: FP-GROWTH ALGORITHM – EXAMPLE TRANSACTIONAL DATABASE 16
TABLE 2.3: FP-GROWTH ALGORITHM – CONDITIONAL FP-TREES 19
TABLE 2.4: FP-GROWTH ALGORITHM – OUTPUT FREQUENT ITEMSETS 19
TABLE 3.1: EXAMPLE TRANSACTIONAL DATABASE 33
TABLE 3.2: THE SYMBOLS USED IN IMIDB ALGORITHM 34
TABLE 3.3: ORIGINAL DATABASE *DB* 37
TABLE 3.4: ORIGINAL FREQUENT ITEMSETS L_{ORG} 38
TABLE 3.5: INCREMENTAL DATABASE *DB* 38
TABLE 3.6: INCREMENTAL FREQUENT ITEMSETS L_{INC} 38
TABLE 3.7: FREQUENT ITEMSETS IN CASE 1 40
TABLE 3.8: FREQUENT ITEMSETS IN CASE 2 40
TABLE 3.9: FREQUENT ITEMSETS IN CASE 3 41
TABLE 3.10: FINAL FREQUENT L_{ALL} FOR THE UPDATED DATABASE DB_{ALL} 41
TABLE 3.11: DATABASES INFORMATION [101] 43
TABLE 3.12: RUNNING TIME COMPARISON – SYNTHETIC DATASETS 50
TABLE 3.13: RUNNING TIME COMPARISON – REAL DATASETS 50
TABLE 3.14: MEMORY USAGE ANALYSIS 51
TABLE 3.15: RUNNING TIME OF THE IMIDB WITH RESPECT TO DATABASE SIZE 52
TABLE 4.1: EXAMPLE TEMPORAL DATABASE 56
TABLE 4.2: THE SYMBOLS USED IN INDXTAR ALGORITHM 60
TABLE 4.3: INCREMENTAL TEMPORAL DATABASE (*DB+DB)* 63
TABLE 4.4: ORIGINAL FREQUENT LEVEL-2 ITEMSETS F_2^{DB} 63
TABLE 4.5: CANDIDATE LEVEL-2 ITEMSETS C_2^{DB} – INCREMENTAL PART 64
TABLE 4.6: GENERATING CANDIDATE LEVEL-2 ITEMSETS C_2^{DB+DB} 65
TABLE 4.7: FREQUENT ITEMSETS IN UPDATED DATABASE F_{DB+DB} 65
TABLE 4.8: DATABASES INFORMATION [101] 67
TABLE 4.9: AVERAGE NUMBER OF GENERATED CANDIDATES 68
TABLE 4.10: TINDEX MEMORY USAGE 69
TABLE 4.11: MEMORY USAGE COMPARISON 70
TABLE 4.12: RUNNING TIME COMPARISON – REAL DATASETS (PART 1) 76
TABLE 4.13: RUNNING TIME COMPARISON – REAL DATASETS (PART 2) 76

TABLE 4.14: RUNNING TIME COMPARISON – SYNTHETIC DATASETS 76
TABLE 4.15: INDXTAR RUNNING TIME WITH RESPECT TO DATABASE SIZE. 77
TABLE 5.1: EXAMPLE OF THE TRANSACTIONAL DATABASE *DB* 80
TABLE 5.2: THE WEIGHTS TABLE OF ITEMS IN *DB* 80
TABLE 5.3: EXAMPLE OF THE TRANSACTIONAL DATABASE DB 87
TABLE 5.4: C1 CANDIDATES AND THEIR WEIGHTED SUPPORT 90
TABLE 5.5: C2 CANDIDATE ITEMSETS 91
TABLE 5.6: EXAMPLE OF THE CANDIDATE GENERATION PHASE 91
TABLE 5.7: FREQUENT ITEMSETS MINING DATASETS [101] 93
TABLE 5.8: NUMBER OF *FWIS* FOR EACH DATABASE 93
TABLE 5.9: RUNNING TIME COMPARISON – REAL DATASETS (PART 1) 102
TABLE 5.10: RUNNING TIME COMPARISON – REAL DATASETS (PART 2) 102
TABLE 5.11: MEMORY USAGE COMPARISON 102
TABLE 5.12: NUMBER OF C2 CANDIDATES COMPARISON 106
TABLE 5.13: COMPARISON SUMMARY 108

List of Figures

FIGURE 2.1: KNOWLEDGE DISCOVERY IN DATABASES (KDD) PROCESS 6
FIGURE 2.2: EXAMPLE OF APRIORI ALGORITHM 14
FIGURE 2.3: STEPS OF FP-TREE CONSTRUCTION 17
FIGURE 2.4: THE COMPLETE FP-TREE FOR THE EXAMPLE DATABASE 18
FIGURE 3.1: T-INDEX CONSTRUCTION STEPS FOR DATABASE IN TABLE 3-1. 34
FIGURE 3.2: POSSIBLE CASES IN INCREMENTAL MINING. 35
FIGURE 3.3: THE PROPOSED IMIDB ALGORITHM 37
FIGURE 3.4: T-INDEX FOR DB TRANSACTIONS TI_{ORG} 39
FIGURE 3.5: T-INDEX FOR DB TRANSACTIONS TI_{INC} 39
FIGURE 3.6: RUNNING TIME COMPARISON – CHESS DATASET 44
FIGURE 3.7: RUNNING TIME COMPARISON – MUSHROOM DATASET 45
FIGURE 3.8: RUNNING TIME COMPARISON – ACCIDENTS DATASET 45
FIGURE 3.9: RUNNING TIME COMPARISON – SYN-T100-N1-L10 DATASET 47
FIGURE 3.10: RUNNING TIME COMPARISON – SYN-T300-N1-L10 DATASET 48
FIGURE 3.11: RUNNING TIME COMPARISON – SYN-T500-N1-L10 DATASET 48
FIGURE 3.12: RUNNING TIME COMPARISON – SYN-T1000-N1-L10 DATASET 49
FIGURE 3.13: SCALABILITY OF THE IMIDB ALGORITHM. POINTS ARE COMPUTATION PRESENTED IN TABLE 3.15, AND LINES ARE LINEAR FITTING FOR SUCH COMPUTATIONS 52
FIGURE 4.1: TINDEX FOR PART P_1 IN THE EXAMPLE DATABASE 57
FIGURE 4.2: COMPLETE TINDEX FOR TEMPORAL DATABASE IN TABLE 4.1 58
FIGURE 4.3: COMPLETE TINDEX2 FOR TEMPORAL DATABASE IN TABLE 4.1. 59
FIGURE 4.4: THE PROPOSED INDXTAR ALGORITHM 62
FIGURE 4.5: TINDEX2 FOR THE UPDATED DATABASE ($DB+DB$) 64
FIGURE 4.6: RUNNING TIME COMPARISON – CHESS DATASET 70
FIGURE 4.7: RUNNING TIME COMPARISON – MUSHROOM DATASET 71
FIGURE 4.8: RUNNING TIME COMPARISON – RETAIL DATASET 72
FIGURE 4.9: RUNNING TIME COMPARISON – ACCIDENTS DATASET 73
FIGURE 4.10: RUNNING TIME COMPARISON – SYN-T100-N1-L10 DATASET 74
FIGURE 4.11: RUNNING TIME COMPARISON – SYN-T200-N1-L10 DATASET 74
FIGURE 4.12: RUNNING TIME COMPARISON – SYN-T400-N1-L10 DATASET 75
FIGURE 4.13: SCALABILITY OF THE INDXTAR ALGORITHM. POINTS ARE COMPUTATION PRESENTED IN TABLE 4.15, AND LINES ARE LINEAR FITTING FOR SUCH COMPUTATIONS. ... 78

FIGURE 5.1: THE PROPOSED WSWF ALGORITHM 87
FIGURE 5.2: PROCESSING DB PARTS IN WSWF ALGORITHM 89
FIGURE 5.3: (A) RUNTIME COMPARISON AND (B) MEMORY USAGE – BMS-POS DATASET 96
FIGURE 5.4: (A) RUNTIME COMPARISON AND (B) MEMORY USAGE – ACCIDENTS DATASET 97
FIGURE 5.5: (A) RUNTIME COMPARISON AND (B) MEMORY USAGE – MUSHROOM DATASET 99
FIGURE 5.6: (A) RUNTIME COMPARISON AND (B) MEMORY USAGE – CHESS DATASET 100
FIGURE 5.7: (A) RUNTIME COMPARISON AND (B) MEMORY USAGE – CONNECT DATASET 101
FIGURE 5.8: *C2* CANDIDATES – BMS-POS DATASET 104
FIGURE 5.9: *C2* CANDIDATES – ACCIDENTS DATASET 104
FIGURE 5.10: *C2* CANDIDATES – MUSHROOM DATASET 105
FIGURE 5.11: *C2* CANDIDATES – CHESS DATASET 105
FIGURE 5.12: C2 CANDIDATES – CONNECT DATASET 106
FIGURE 5.13: WSWF ALGORITHM SCALABILITY 108

List of Abbreviations

ARM	Association Rules Mining
FIM	Fast Incremental Mining Algorithm
FP-tree	Frequent Pattern tree data structure
FWI	Frequent Weighted Itemsets
IMIDB	Indexed Mining of Incremental Databases Algorithm
IndxTAR	Indexed Temporal Association Rules Mining Algorithm
ITARM	Incremental Temporal Association Rules Mining Algorithm
KDD	Knowledge Discovery in Databases
minsupp	Minimum Support threshold
minWS	Minimum Weighted Support threshold
SWF	Sliding Window Filtering Algorithm
TID(X)	Transaction ID List for itemset X
T-index	Trie Indexing Scheme
tw(T_p)	Transaction Weight of a Transaction T_p
WAR	Weighted Association Rules
WIT-tree	Weighted Itemset TID tree
wSWF	Weighted Sliding Window Filtering Algorithm

Chapter 1

Introduction