

Contents

- Abstract..... 6
- 1. Introduction 7
 - 1.1. Data Mining from a Database Perspective..... 7
 - 1.2. Aim and Scope of the Dissertation 11
- 2. Frequent Itemset Mining 14
 - 2.1. Overview, Genesis, Applications, and Importance of the Problem ... 14
 - 2.2. Formulation of the Frequent Itemset Mining Problem 16
 - 2.3. Computational Complexity of the Problem 18
 - 2.4. Overview of Approaches to Frequent Itemset Mining..... 19
 - 2.4.1. Introduction 19
 - 2.4.2. Search Space Traversal Strategies 20
 - 2.4.3. Database Layout 23
 - 2.4.4. Using Memory to Store Mined Data 26
 - 2.4.5. Itemset Support Counting..... 27
 - 2.5. Representative Frequent Itemset Mining Algorithms 28
 - 2.5.1. Introduction 28
 - 2.5.2. Apriori 30
 - 2.5.3. FP-growth..... 36
 - 2.5.4. Partition 39
 - 2.6. Research Trends in Frequent Itemset Mining..... 41
 - 2.6.1. Introduction 41
 - 2.6.2. Taking Advantage of DBMS Functionality in Frequent Itemset Mining..... 42
 - 2.6.3. Sampling for Frequent Itemset Mining..... 44
 - 2.6.4. Concise Representations of Frequent Itemsets 46
 - 2.6.5. Parallel and Distributed Frequent Itemset Mining..... 49
 - 2.6.6. Frequent Itemset Mining over Data Streams 52
 - 2.6.7. Privacy Preserving Frequent Itemset Mining 54
- 3. Data Mining as Advanced Querying 57
 - 3.1. Motivation..... 57
 - 3.2. Prototype Data Mining Query Languages 57
 - 3.3. Data Mining Standards 60
 - 3.4. Data Mining Queries in Contemporary Database Management Systems 67
 - 3.5. Data Mining Queries: Summary of the Current State of the Art and Implications..... 71

4.	Frequent Itemset Query Processing	73
4.1.	Constraint-based Frequent Itemset Mining.....	73
4.2.	Reusing Results of Frequent Itemset Queries	76
4.3.	Reusing Results vs. Pushing Constraints into the Mining Process	80
5.	Processing Batches of Frequent Itemset Queries	82
5.1.	Motivation.....	82
5.2.	General Model of Frequent Itemset Queries.....	83
5.3.	Batches of Frequent Itemset Queries and Problem Formulation	85
5.4.	Model of Query Data Sharing	87
5.5.	Related Work	90
6.	Methods Independent of the Mining Algorithm.....	92
6.1.	Sequential Processing with Result Caching and Reusing	92
6.2.	Result Filtering and Incremental Mining	93
6.3.	Query Scheduling.....	97
6.4.	Query Scheduling with Intermediate Queries	100
6.5.	Mine Merge.....	106
6.6.	Experimental Results	111
6.7.	Summary and Discussion.....	116
7.	Methods for the Apriori Algorithm	118
7.1.	Common Counting.....	118
7.2.	Common Counting with Query Partitioning.....	120
7.2.1.	Motivation	120
7.2.2.	Key Issues.....	121
7.2.3.	Query Partitioning as a Case of Hypergraph Partitioning ...	124
7.2.4.	Computational Complexity of the Problem	128
7.2.5.	Algorithm CCRecursive	131
7.2.6.	Algorithm CCFull.....	133
7.2.7.	Algorithm CCoarsening	136
7.2.8.	Algorithm CCAgglomerative	139
7.2.9.	Algorithm CCAgglomerativeNoise	140
7.2.10.	Algorithm CCGreedy	142
7.2.11.	Algorithm CCSemiGreedy	144
7.3.	Common Candidate Tree	145
7.4.	Experimental Results	148
7.4.1.	Query Partitioning for Common Counting	148
7.4.2.	Common Counting vs. Common Candidate Tree.....	159
7.5.	Summary and Discussion.....	169

8. Methods for the FP-growth Algorithm	171
8.1. Common Building.....	171
8.2. Common FP-tree.....	173
8.3. Experimental Results	176
8.4. Summary and Discussion.....	182
9. Methods for the Partition Algorithm.....	186
9.1. Integration of Dataset Scans for Partition	186
9.2. Partition Mine Merge Improved	187
9.3. Experimental Results	191
9.4. Summary and Discussion.....	195
10. Data Access Methods in Processing Sets of Frequent Itemset Queries	197
10.1. Comparison of Proposed Techniques in Terms of Data Access Schemes	197
10.2. Data Organization and Access Methods in Contemporary DBMSs	199
10.3. Techniques of Processing Sets of Frequent Itemset Queries with Full Table Scans.....	202
10.4. Theoretical Cost Analysis	204
10.5. Experimental Results	207
10.6. Summary and Discussion.....	214
11. Conclusions and Future Work	216
Bibliography	221
Streszczenie.....	238