Contents

1 **Introduction** 1
- 1.1 Introduction 1
- 1.2 Data Mining Process 2
  - 1.2.1 Data Preparation 2
  - 1.2.2 Application of Data Mining Algorithms 3
  - 1.2.3 Pattern Evaluation 3
  - 1.2.4 Knowledge Representation 3
- 1.3 Different Types of Data Representations 4
  - 1.3.1 Relational Data 4
  - 1.3.2 Sequential Data 4
  - 1.3.3 Semi-structured Data 4
  - 1.3.4 Unstructured Data 5
- 1.4 Different Types of Knowledge Mined 5
  - 1.4.1 Type of Information Mined 5
  - 1.4.2 Representing the Mined Knowledge 7
- 1.5 Common Data Mining Tasks 9
  - 1.5.1 Association Mining 10
  - 1.5.2 Classification and Prediction 10
  - 1.5.3 Cluster Analysis 11
  - 1.5.4 Outlier Detection 11
- 1.6 Sources of Data with Complex Structures 12
  - 1.6.1 Online Information 12
  - 1.6.2 Chemical Databases 13
  - 1.6.3 Bioinformatics 13
  - 1.6.4 Ontologies 13
- 1.7 Complex Structures 14
- 1.8 Emergence of Semi-structured Data Sources 14
- 1.9 Challenges of Mining Data with Complex Structures 16
- 1.10 Conclusion 17
- References 17

Bibliografische Informationen
http://d-nb.info/1007904178

digitalisiert durch DEUTSCHE NATIONAL BIBLIOTHEK

2 **Tree Mining Problem** 23
2.1 Introduction 23
2.2 Problem of Association Rule Mining 23
2.2.1 Association Rule Framework 24
2.2.2 Support 24
2.2.3 Confidence 24
2.3 Emerging Field of Tree Mining 26
2.3.1 XML and Association Mining 27
2.3.2 The Parallel between XML and Tree Structure 29
2.3.3 Problem of XML Document Association Mining 29
2.4 General Tree Concepts and Definitions 30
2.5 Frequent Subtree Mining Problem 31
2.5.1 Subtree Types 31
2.5.2 Support Definitions 33
2.6 Illustrative Example 34
2.6.1 Issue with Pseudo Frequent Subtrees 36
2.7 Canonical Form of a Subtree 36
2.8 Conclusion 37
References 37

3 **Algorithm Development Issues** 41
3.1 Introduction 41
3.2 Tree Representation 42
3.2.1 Efficient Representation for Processing XML Documents 43
3.3 Data Structure Issues 46
3.4 Enumeration Techniques 47
3.4.1 Enumeration by Join 47
3.4.2 Enumeration by Extension 48
3.4.3 Structure Guided Enumeration 48
3.4.4 Horizontal vs. Vertical Enumeration 49
3.5 Frequency Counting 50
3.6 Canonical Form Ordering Schemes for Unordered Subtrees 50
3.6.1 Depth-First Canonical Ordering Form 51
3.6.2 Breadth-First Canonical Form Ordering 52
3.7 Overview of Existing Tree Mining Algorithms 53
3.7.1 Algorithm Using the Join Candidate Enumeration Approach 56
3.8 Conclusion 62
References 63

4 Tree Model Guided Framework 67
4.1 Introduction 67
4.2 Tree Model Guided Candidate Subtree Enumeration 69
4.3 Efficient Representations and Data Structure for Trees 72
4.3.1 Dictionary 73
4.3.2 Embedding List 75
4.3.3 Recursive List 76
4.4 Frequency Counting 78
4.4.1 Vertical Occurrence List 78
4.4.2 RMP Coordinate List 79
4.5 Constraints 81
4.5.1 Feasible Computation through Level of Embedding Constraint 82
4.5.2 Splitting Embedded Subtree through Distance Constraint 83
4.6 Conclusion 84
References 85

5 TMG Framework for Mining Ordered Subtrees 87
5.1 Introduction 87
5.2 Overview of the Framework for Mining Ordered Induced/Embedded Subtrees 89
5.3 Detailed Description of the Framework 90
5.3.1 Tree Representation 90
5.3.2 Data Pre-processing 90
5.3.3 Generating the Recursive List (RL), F_1 and F_2 91
5.3.4 Enumerating Ordered Induced/Embedded k-Subtrees Using the TMG Enumeration Technique 93
5.3.5 Frequency Counting 95
5.3.6 Pseudo Code 98
5.4 Mathematical Analysis 98
5.4.1 Mathematical Model of TMG 98
5.4.2 Complexity Analysis of Candidate Generation of an Embedded/Induced Subtree 105
5.5 Experimental Evaluation and Comparisons 112
5.5.1 The Rationale of the Experimental Comparison 113
5.5.2 Implementation Issues 114
5.6 Experimental Results and Discussion 117
5.6.1 Experiment Set I 118
5.6.2 Experiment Set II 127
5.7 Summary 136
References 137

6 TMG Framework for Mining Unordered Subtrees 139
6.1 Introduction 139
6.2 Canonical Form Used in TMG Framework 141
6.3 Overview of the TMG Framework for Mining Unordered Induced/Embedded Subtrees 142
6.4 Detailed Description of TMG Framework for Unordered Subtree Mining 143
6.4.1 Canonical Form Ordering 145
6.4.2 Enumerating Unordered Subtrees Using TMG Framework 149
6.4.3 Frequency Counting of Candidate Subtrees 152
6.4.4 Pseudo Code of the Algorithm 153
6.5 Experimental Comparison with Existing Approaches for Unordered Subtree Mining 154
6.5.1 Testing the Approach for Mining of Unordered Induced Subtrees 155
6.5.2 Testing the Approach for Mining of Unordered Embedded Subtrees 159
6.5.3 Additional Tests 167
6.6 Conclusion 171
References 173

7 Mining Distance-Constrained Embedded Subtrees 175
7.1 Introduction 175
7.2 Distance-Constrained Embedded Subtrees 178
7.3 Motivation for Integrating the Distance Constraint 180
7.4 Extending TMG Framework to Extract Distance-Constrained Embedded Subtrees 181
7.4.1 Tree Representation 181
7.4.2 Mining Ordered Distance-Constrained Subtrees 182
7.4.3 Mining Unordered Distance-Constrained Subtrees 182
7.5 Experimental Results and Discussion 183
7.5.1 Ordered Distance-Constrained Embedded Subtrees 184
7.5.2 Unordered Distance-Constrained Embedded Subtrees 186
7.6 Conclusion 189
References 189

8 Mining Maximal and Closed Frequent Subtrees 191
8.1 Introduction 191
8.2 Problem of Closed/Maximal Subtree Mining.................... 193
8.3 Methods for Mining Closed/Maximal Subtrees 194
8.4 CMTreeMiner Algorithm 196
8.5 Conclusion 198
References 198

9 Tree Mining Applications ... 201
9.1 Introduction ... 201
9.2 Types of Knowledge Representations Considered ... 202
9.3 Application for General Knowledge Analysis Tasks ... 204
9.3.1 Implications of Using Different Support Definitions ... 204
9.3.2 Implications for Mining Different Subtree Types ... 207
9.3.3 Implications for Mining Constrained Embedded Subtrees ... 213
9.4 Mining of Healthcare Data ... 215
9.4.1 Mining of Patients' Records ... 216
9.4.2 Experiment ... 219
9.5 Web Log Mining ... 221
9.5.1 Transforming Web Usage Patterns to Trees ... 223
9.5.2 Experiments ... 228
9.6 Application for the Knowledge Matching Task ... 237
9.6.1 Method Description ... 238
9.6.2 Experiments ... 240
9.7 Mining Substructures in Protein Data ... 243
9.8 Conclusion ... 244
References ... 245

10 Extension of TMG Framework for Mining Frequent Subsequences ... 249
10.1 Introduction ... 249
10.2 General Sequence Concepts and Definitions ... 250
10.2.1 Problem of Mining Frequent Sequences from a Database of Sequences ... 251
10.3 Overview of Some Existing Techniques for Mining Sequential Data ... 251
10.3.1 Apriori-Like Approaches ... 252
10.3.2 Pattern Growth Based Approaches ... 253
10.3.3 Other Types of Sequential Pattern Mining ... 256
10.3.4 Constraint-Based Sequential Mining ... 260
10.4 WAP-Mine Algorithm ... 263
10.4.1 WAP-Tree Construction ... 264
10.4.2 Mining Frequent Subsequences from WAP-Tree ... 267
10.4.3 Other WAP-Tree Based Algorithms ... 270
10.5 Overview of the Proposed Solution ... 270
10.6 SEQUEST: Mining Frequent Subsequences from a Database of Sequences ... 271
10.6.1 Database Scanning ... 272
10.6.2 Constructing DMA-Strips ... 272
10.6.3 Enumeration of Subsequences ... 273
10.6.4 Frequency Counting ... 274

10.6.5 Pruning 275
10.6.6 SEQUEST Pseudo-Code 275
10.7 Experimental Results and Discussions 276
10.7.1 Performance Test 277
10.7.2 Scalability Test 278
10.7.3 Frequency Distribution Test 279
10.7.4 Large Database Test 279
10.7.5 Overall Conclusions 280
10.8 Conclusion 280
References 281

11 Graph Mining 287
11.1 Introduction 287
11.2 General Graph Concepts and Definitions 288
11.3 Graph Isomorphism Problem 288
11.4 Existing Graph Mining Methods 289
11.4.1 Apriori-Like Methods 290
11.4.2 Pattern-Growth Methods 291
11.4.3 Inductive Logic Programming (ILP) Methods 292
11.4.4 Greedy Search Methods 293
11.4.5 Other Methods 294
11.4.6 Mining Closed/Maximal Subgraph Patterns 297
11.5 Conclusion 298
References 298

12 New Research Directions 301
12.1 Introduction 301
12.2 Frequent Pattern Reduction 302
12.2.1 Frequent Pattern Reduction and Rule Evaluation 303
12.2.2 Reducing Frequent Subtrees 306
12.3 Top-Down Approach for Frequent Subtree Mining 308
12.4 Model Guided Approach for Graph Mining 311
12.5 Conjoint Mining of Different Data Types 311
12.5.1 A Framework for Conjoint Mining of Relational and Semi-structured Data 312
12.6 Ontology Learning 313
12.6.1 Ontology Definition and Formulations 314
12.6.2 Concept Term Matching Problem 316
12.6.3 Structural Representation Matching Problem 318
12.6.4 Ontology Learning Method Aims 319
12.7 Conclusion 322
References 322