

Charu C. Aggarwal
Haixun Wang
Editors

Managing and Mining Graph Data

 Springer

Contents

List of Figures	xv
List of Tables	xxi
Preface	xxiii
1	
An Introduction to Graph Data	1
<i>Charu C. Aggarwal and Haixun Wang</i>	
1. Introduction	1
2. Graph Management and Mining Applications	3
3. Summary	8
References	9
2	
Graph Data Management and Mining: A Survey of Algorithms and Applications	13
<i>Charu C. Aggarwal and Haixun Wang</i>	
1. Introduction	13
2. Graph Data Management Algorithms	16
2.1 Indexing and Query Processing Techniques	16
2.2 Reachability Queries	19
2.3 Graph Matching	21
2.4 Keyword Search	24
2.5 Synopsis Construction of Massive Graphs	27
3. Graph Mining Algorithms	29
3.1 Pattern Mining in Graphs	29
3.2 Clustering Algorithms for Graph Data	32
3.3 Classification Algorithms for Graph Data	37
3.4 The Dynamics of Time-Evolving Graphs	40
4. Graph Applications	43
4.1 Chemical and Biological Applications	43
4.2 Web Applications	45
4.3 Software Bug Localization	51
5. Conclusions and Future Research	55
References	55
3	
Graph Mining: Laws and Generators	69
<i>Deepayan Chakrabarti, Christos Faloutsos and Mary McGlohon</i>	
1. Introduction	70
2. Graph Patterns	71

2.1	Power Laws and Heavy-Tailed Distributions	72
2.2	Small Diameters	77
2.3	Other Static Graph Patterns	79
2.4	Patterns in Evolving Graphs	82
2.5	The Structure of Specific Graphs	84
3.	Graph Generators	86
3.1	Random Graph Models	88
3.2	Preferential Attachment and Variants	92
3.3	Optimization-based generators	101
3.4	Tensor-based	108
3.5	Generators for specific graphs	113
3.6	Graph Generators: A summary	115
4.	Conclusions	115
	References	117
4		
	Query Language and Access Methods for Graph Databases	125
	<i>Huahai He and Ambuj K. Singh</i>	
1.	Introduction	126
1.1	Graphs-at-a-time Queries	126
1.2	Graph Specific Optimizations	127
1.3	GraphQL	128
2.	Operations on Graph Structures	129
2.1	Concatenation	130
2.2	Disjunction	131
2.3	Repetition	131
3.	Graph Query Language	132
3.1	Data Model	132
3.2	Graph Patterns	133
3.3	Graph Algebra	134
3.4	FLWR Expressions	137
3.5	Expressive Power	138
4.	Implementation of the Selection Operator	140
4.1	Graph Pattern Matching	140
4.2	Local Pruning and Retrieval of Feasible Mates	142
4.3	Joint Reduction of Search Space	144
4.4	Optimization of Search Order	146
5.	Experimental Study	148
5.1	Biological Network	148
5.2	Synthetic Graphs	150
6.	Related Work	152
6.1	Graph Query Languages	152
6.2	Graph Indexing	155
7.	Future Research Directions	155
8.	Conclusion	156
	Appendix: Query Syntax of GraphQL	156
	References	157
5		
	Graph Indexing	161
	<i>Xifeng Yan and Jiawei Han</i>	
1.	Introduction	161

2.	Feature-Based Graph Index	162
2.1	Paths	163
2.2	Frequent Structures	164
2.3	Discriminative Structures	166
2.4	Closed Frequent Structures	167
2.5	Trees	167
2.6	Hierarchical Indexing	168
3.	Structure Similarity Search	169
3.1	Feature-Based Structural Filtering	170
3.2	Feature Miss Estimation	171
3.3	Frequency Difference	172
3.4	Feature Set Selection	173
3.5	Structures with Gaps	174
4.	Reverse Substructure Search	175
5.	Conclusions	177
	References	178
6		
	Graph Reachability Queries: A Survey	181
	<i>Jeffrey Xu Yu and Jiefeng Cheng</i>	
1.	Introduction	181
2.	Traversal Approaches	186
2.1	Tree+SSPI	187
2.2	GRIPP	187
3.	Dual-Labeling	188
4.	Tree Cover	190
5.	Chain Cover	191
5.1	Computing the Optimal Chain Cover	193
6.	Path-Tree Cover	194
7.	2-HOP Cover	196
7.1	A Heuristic Ranking	197
7.2	A Geometrical-Based Approach	198
7.3	Graph Partitioning Approaches	199
7.4	2-Hop Cover Maintenance	202
8.	3-Hop Cover	204
9.	Distance-Aware 2-Hop Cover	205
10.	Graph Pattern Matching	207
10.1	A Special Case: $A \leftrightarrow D$	208
10.2	The General Cases	211
11.	Conclusions and Summary	212
	References	212
7		
	Exact and Inexact Graph Matching: Methodology and Applications	217
	<i>Kaspar Riesen, Xiaoyi Jiang and Horst Bunke</i>	
1.	Introduction	218
2.	Basic Notations	219
3.	Exact Graph Matching	221
4.	Inexact Graph Matching	226
4.1	Graph Edit Distance	227
4.2	Other Inexact Graph Matching Techniques	229
5.	Graph Matching for Data Mining and Information Retrieval	231

6.	Vector Space Embeddings of Graphs via Graph Matching	235
7.	Conclusions	239
	References	240
8		
	A Survey of Algorithms for Keyword Search on Graph Data	249
	<i>Haixun Wang and Charu C. Aggarwal</i>	
1.	Introduction	250
2.	Keyword Search on XML Data	252
	2.1 Query Semantics	253
	2.2 Answer Ranking	254
	2.3 Algorithms for LCA-based Keyword Search	258
3.	Keyword Search on Relational Data	260
	3.1 Query Semantics	260
	3.2 DBXplorer and DISCOVER	261
4.	Keyword Search on Schema-Free Graphs	263
	4.1 Query Semantics and Answer Ranking	263
	4.2 Graph Exploration by Backward Search	265
	4.3 Graph Exploration by Bidirectional Search	266
	4.4 Index-based Graph Exploration – the BLINKS Algorithm	267
	4.5 The ObjectRank Algorithm	269
5.	Conclusions and Future Research	271
	References	271
9		
	A Survey of Clustering Algorithms for Graph Data	275
	<i>Charu C. Aggarwal and Haixun Wang</i>	
1.	Introduction	275
2.	Node Clustering Algorithms	277
	2.1 The Minimum Cut Problem	277
	2.2 Multi-way Graph Partitioning	281
	2.3 Conventional Generalizations and Network Structure Indices	282
	2.4 The Girvan-Newman Algorithm	284
	2.5 The Spectral Clustering Method	285
	2.6 Determining Quasi-Cliques	288
	2.7 The Case of Massive Graphs	289
3.	Clustering Graphs as Objects	291
	3.1 Extending Classical Algorithms to Structural Data	291
	3.2 The XProj Approach	293
4.	Applications of Graph Clustering Algorithms	295
	4.1 Community Detection in Web Applications and Social Networks	296
	4.2 Telecommunication Networks	297
	4.3 Email Analysis	297
5.	Conclusions and Future Research	297
	References	299
10		
	A Survey of Algorithms for Dense Subgraph Discovery	303
	<i>Victor E. Lee, Ning Ruan, Ruoming Jin and Charu Aggarwal</i>	
1.	Introduction	304

2.	Types of Dense Components	305
2.1	Absolute vs. Relative Density	305
2.2	Graph Terminology	306
2.3	Definitions of Dense Components	307
2.4	Dense Component Selection	308
2.5	Relationship between Clusters and Dense Components	309
3.	Algorithms for Detecting Dense Components in a Single Graph	311
3.1	Exact Enumeration Approach	311
3.2	Heuristic Approach	314
3.3	Exact and Approximation Algorithms for Discovering Densest Components	322
4.	Frequent Dense Components	327
4.1	Frequent Patterns with Density Constraints	327
4.2	Dense Components with Frequency Constraint	328
4.3	Enumerating Cross-Graph Quasi-Cliques	328
5.	Applications of Dense Component Analysis	329
6.	Conclusions and Future Research	331
	References	333
11		
	Graph Classification	337
	<i>Koji Tsuda and Hiroto Saigo</i>	
1.	Introduction	337
2.	Graph Kernels	340
2.1	Random Walks on Graphs	341
2.2	Label Sequence Kernel	342
2.3	Efficient Computation of Label Sequence Kernels	343
2.4	Extensions	349
3.	Graph Boosting	349
3.1	Formulation of Graph Boosting	351
3.2	Optimal Pattern Search	353
3.3	Computational Experiments	354
3.4	Related Work	355
4.	Applications of Graph Classification	358
5.	Label Propagation	358
6.	Concluding Remarks	359
	References	359
12		
	Mining Graph Patterns	365
	<i>Hong Cheng, Xifeng Yan and Jiawei Han</i>	
1.	Introduction	366
2.	Frequent Subgraph Mining	366
2.1	Problem Definition	366
2.2	Apriori-based Approach	367
2.3	Pattern-Growth Approach	368
2.4	Closed and Maximal Subgraphs	369
2.5	Mining Subgraphs in a Single Graph	370
2.6	The Computational Bottleneck	371
3.	Mining Significant Graph Patterns	372
3.1	Problem Definition	372
3.2	gboost: A Branch-and-Bound Approach	373

3.3	gPLS: A Partial Least Squares Regression Approach	375
3.4	LEAP: A Structural Leap Search Approach	378
3.5	GraphSig: A Feature Representation Approach	382
4.	Mining Representative Orthogonal Graphs	385
4.1	Problem Definition	386
4.2	Randomized Maximal Subgraph Mining	387
4.3	Orthogonal Representative Set Generation	388
5.	Conclusions	389
	References	389
13		
	A Survey on Streaming Algorithms for Massive Graphs	393
	<i>Jian Zhang</i>	
1.	Introduction	393
2.	Streaming Model for Massive Graphs	395
3.	Statistics and Counting Triangles	397
4.	Graph Matching	400
4.1	Unweighted Matching	400
4.2	Weighted Matching	403
5.	Graph Distance	405
5.1	Distance Approximation using Multiple Passes	406
5.2	Distance Approximation in One Pass	411
6.	Random Walks on Graphs	412
7.	Conclusions	416
	References	417
14		
	A Survey of Privacy-Preservation of Graphs and Social Networks	421
	<i>Xintao Wu, Xiaowei Ying, Kun Liu and Lei Chen</i>	
1.	Introduction	422
1.1	Privacy in Publishing Social Networks	422
1.2	Background Knowledge	423
1.3	Utility Preservation	424
1.4	Anonymization Approaches	424
1.5	Notations	425
2.	Privacy Attacks on Naive Anonymized Networks	426
2.1	Active Attacks and Passive Attacks	426
2.2	Structural Queries	427
2.3	Other Attacks	428
3.	K -Anonymity Privacy Preservation via Edge Modification	428
3.1	K -Degree Generalization	429
3.2	K -Neighborhood Anonymity	430
3.3	K -Automorphism Anonymity	431
4.	Privacy Preservation via Randomization	433
4.1	Resilience to Structural Attacks	434
4.2	Link Disclosure Analysis	435
4.3	Reconstruction	437
4.4	Feature Preserving Randomization	438
5.	Privacy Preservation via Generalization	440
6.	Anonymizing Rich Graphs	441

6.1	Link Protection in Rich Graphs	442
6.2	Anonymizing Bipartite Graphs	443
6.3	Anonymizing Rich Interaction Graphs	444
6.4	Anonymizing Edge-Weighted Graphs	445
7.	Other Privacy Issues in Online Social Networks	446
7.1	Deriving Link Structure of the Entire Network	446
7.2	Deriving Personal Identifying Information from Social Networking Sites	448
8.	Conclusion and Future Work	448
Acknowledgments		449
References		449
15		
A Survey of Graph Mining for Web Applications		455
<i>Debora Donato and Aristides Gionis</i>		
1.	Introduction	456
2.	Preliminaries	457
2.1	Link Analysis Ranking Algorithms	459
3.	Mining High-Quality Items	461
3.1	Prediction of Successful Items in a Co-citation Network	463
3.2	Finding High-Quality Content in Question-Answering Portals	465
4.	Mining Query Logs	469
4.1	Description of Query Logs	470
4.2	Query Log Graphs	470
4.3	Query Recommendations	477
5.	Conclusions	480
References		481
16		
Graph Mining Applications to Social Network Analysis		487
<i>Lei Tang and Huan Liu</i>		
1.	Introduction	487
2.	Graph Patterns in Large-Scale Networks	489
2.1	Scale-Free Networks	489
2.2	Small-World Effect	491
2.3	Community Structures	492
2.4	Graph Generators	494
3.	Community Detection	494
3.1	Node-Centric Community Detection	495
3.2	Group-Centric Community Detection	498
3.3	Network-Centric Community Detection	499
3.4	Hierarchy-Centric Community Detection	504
4.	Community Structure Evaluation	505
5.	Research Issues	507
References		508
17		
Software-Bug Localization with Graph Mining		515
<i>Frank Eichinger and Klemens Böhm</i>		
1.	Introduction	516
2.	Basics of Call Graph Based Bug Localization	517

2.1	Dynamic Call Graphs	517
2.2	Bugs in Software	518
2.3	Bug Localization with Call Graphs	519
2.4	Graph and Tree Mining	520
3.	Related Work	521
4.	Call-Graph Reduction	525
4.1	Total Reduction	525
4.2	Iterations	526
4.3	Temporal Order	528
4.4	Recursion	529
4.5	Comparison	531
5.	Call Graph Based Bug Localization	532
5.1	Structural Approaches	532
5.2	Frequency-based Approach	535
5.3	Combined Approaches	538
5.4	Comparison	539
6.	Conclusions and Future Directions	542
	Acknowledgments	543
	References	543
18		
	A Survey of Graph Mining Techniques for Biological Datasets	547
	<i>S. Parthasarathy, S. Tatikonda and D. Ucar</i>	
1.	Introduction	548
2.	Mining Trees	549
2.1	Frequent Subtree Mining	550
2.2	Tree Alignment and Comparison	552
2.3	Statistical Models	554
3.	Mining Graphs for the Discovery of Frequent Substructures	555
3.1	Frequent Subgraph Mining	555
3.2	Motif Discovery in Biological Networks	560
4.	Mining Graphs for the Discovery of Modules	562
4.1	Extracting Communities	564
4.2	Clustering	566
5.	Discussion	569
	References	571
19		
	Trends in Chemical Graph Data Mining	581
	<i>Nikil Wale, Xia Ning and George Karypis</i>	
1.	Introduction	582
2.	Topological Descriptors for Chemical Compounds	583
2.1	Hashed Fingerprints (FP)	584
2.2	Maccs Keys (MK)	584
2.3	Extended Connectivity Fingerprints (ECFP)	584
2.4	Frequent Subgraphs (FS)	585
2.5	Bounded-Size Graph Fragments (GF)	585
2.6	Comparison of Descriptors	585
3.	Classification Algorithms for Chemical Compounds	588
3.1	Approaches based on Descriptors	588
3.2	Approaches based on Graph Kernels	589
4.	Searching Compound Libraries	590

<i>Contents</i>	xiii
4.1 Methods Based on Direct Similarity	591
4.2 Methods Based on Indirect Similarity	592
4.3 Performance of Indirect Similarity Methods	594
5. Identifying Potential Targets for Compounds	595
5.1 Model-based Methods For Target Fishing	596
5.2 Performance of Target Fishing Strategies	600
6. Future Research Directions	600
References	602
Index	607