
Similarity Search

The Metric Space Approach

Contents

Dedication	v
Foreword	xiii
Preface	xv
Acknowledgments	xvii
Part I Metric Searching in a Nutshell	
Overview	3
1. FOUNDATIONS OF METRIC SPACE SEARCHING	5
1 The Distance Searching Problem	6
2 The Metric Space	8
3 Distance Measures	9
3.1 Minkowski Distances	10
3.2 Quadratic Form Distance	11
3.3 Edit Distance	12
3.4 Tree Edit Distance	13
3.5 Jaccard's Coefficient	13
3.6 Hausdorff Distance	14
3.7 Time Complexity	14
4 Similarity Queries	15
4.1 Range Query	15
4.2 Nearest Neighbor Query	16
4.3 Reverse Nearest Neighbor Query	17
4.4 Similarity Join	17
4.5 Combinations of Queries	18
4.6 Complex Similarity Queries	18

5	Basic Partitioning Principles	20
5.1	Ball Partitioning	20
5.2	Generalized Hyperplane Partitioning	21
5.3	Excluded Middle Partitioning	21
5.4	Extensions	21
6	Principles of Similarity Query Execution	22
6.1	Basic Strategies	22
6.2	Incremental Similarity Search	25
7	Policies for Avoiding Distance Computations	26
7.1	Explanatory Example	27
7.2	Object-Pivot Distance Constraint	28
7.3	Range-Pivot Distance Constraint	30
7.4	Pivot-Pivot Distance Constraint	31
7.5	Double-Pivot Distance Constraint	33
7.6	Pivot Filtering	34
8	Metric Space Transformations	35
8.1	Metric Hierarchies	36
8.1.1	Lower-Bounding Functions	36
8.2	User-Defined Metric Functions	38
8.2.1	Searching Using Lower-Bounding Functions	38
8.3	Embedding Metric Space	39
8.3.1	Embedding Examples	39
8.3.2	Reducing Dimensionality	40
9	Approximate Similarity Search	41
9.1	Principles	41
9.2	Generic Algorithms	44
9.3	Measures of Performance	46
9.3.1	Improvement in Efficiency	46
9.3.2	Precision and Recall	46
9.3.3	Relative Error on Distances	48
9.3.4	Position Error	49
10	Advanced Issues	50
10.1	Statistics on Metric Datasets	51
10.1.1	Distribution and Density Functions	51
10.1.2	Distance Distribution and Density	52
10.1.3	Homogeneity of Viewpoints	54
10.2	Proximity of Ball Regions	55
10.3	Performance Prediction	58

<i>Contents</i>	ix
10.4 Tree Quality Measures	60
10.5 Choosing Reference Points	63
2. SURVEY OF EXISTING APPROACHES	67
1 Ball Partitioning Methods	67
1.1 Burkhard-Keller Tree	68
1.2 Fixed Queries Tree	69
1.3 Fixed Queries Array	70
1.4 Vantage Point Tree	72
1.4.1 Multi-Way Vantage Point Tree	74
1.5 Excluded Middle Vantage Point Forest	75
2 Generalized Hyperplane Partitioning Approaches	76
2.1 Bisector Tree	76
2.2 Generalized Hyperplane Tree	77
3 Exploiting Pre-Computed Distances	78
3.1 AESA	78
3.2 Linear AESA	79
3.3 Other Methods	80
4 Hybrid Indexing Approaches	81
4.1 Multi Vantage Point Tree	81
4.2 Geometric Near-neighbor Access Tree	82
4.3 Spatial Approximation Tree	85
4.4 M-tree	87
4.5 Similarity Hashing	88
5 Approximate Similarity Search	89
5.1 Exploiting Space Transformations	89
5.2 Approximate Nearest Neighbors with BBD Trees	90
5.3 Angle Property Technique	92
5.4 Clustering for Indexing	94
5.5 Vector Quantization Index	95
5.6 Buoy Indexing	97
5.7 Hierarchical Decomposition of Metric Spaces	97
5.7.1 Relative Error Approximation	98
5.7.2 Good Fraction Approximation	98
5.7.3 Small Chance Improvement Approximation	98
5.7.4 Proximity-Based Approximation	99
5.7.5 PAC Nearest Neighbor Search	99

Part II Metric Searching in Large Collections of Data

Overview	103
3. CENTRALIZED INDEX STRUCTURES	105
1 M-tree Family	105
1.1 The M-tree	105
1.2 Bulk-Loading Algorithm of M-tree	109
1.3 Multi-Way Insertion Algorithm	112
1.4 The Slim Tree	113
1.4.1 Slim-Down Algorithm	114
1.4.2 Generalized Slim-Down Algorithm	116
1.5 Pivoting M-tree	118
1.6 The M^+ -tree	121
1.7 The M^2 -tree	124
2 Hash-based metric indexing	125
2.1 The D-index	126
2.1.1 Insertion and Search Strategies	129
2.2 The eD-index	131
2.2.1 Similarity Self-Join Algorithm with eD-index	133
3 Performance Trials	136
3.1 Datasets and Distance Measures	137
3.2 Performance Comparison	138
3.3 Different Query Types	140
3.4 Scalability	141
4. APPROXIMATE SIMILARITY SEARCH	145
1 Relative Error Approximation	145
2 Good Fraction Approximation	148
3 Small Chance Improvement Approximation	150
4 Proximity-Based Approximation	152
5 PAC Nearest Neighbor Searching	153
6 Performance Trials	154
6.1 Range Queries	155
6.2 Nearest Neighbors Queries	156
6.3 Global Considerations	159

<i>Contents</i>	xi
5. PARALLEL AND DISTRIBUTED INDEXES	161
1 Preliminaries	161
1.1 Parallel Computing	162
1.2 Distributed Computing	163
1.2.1 Scalable and Distributed Data Structures	163
1.2.2 Peer-to-Peer Data Networks	164
2 Processing M-trees with Parallel Resources	164
2.1 CPU Parallelism	165
2.2 I/O Parallelism	165
2.3 Object Declustering in M-trees	167
3 Scalable Distributed Similarity Search Structure	167
3.1 Architecture	168
3.2 Address Search Tree	169
3.3 Storage Management	169
3.3.1 Bucket Splitting	170
3.3.2 Choosing Pivots	171
3.4 Insertion of Objects	171
3.5 Range Search	172
3.6 Nearest Neighbor Search	173
3.7 Deletions and Updates of Objects	174
3.8 Image Adjustment	175
3.9 Logarithmic Replication Strategy	177
3.10 Joining the Peer-to-Peer Network	178
3.11 Leaving the Peer-to-Peer Network	178
4 Performance Trials	179
4.1 Datasets and Computing Infrastructure	180
4.2 Performance of Similarity Queries	180
4.2.1 Global Costs	181
4.2.2 Parallel Costs	183
4.2.3 Comparison of Search Algorithms	188
4.3 Data Volume Scalability	189
Concluding Summary	193
References	197
Author Index	211
Index	215
Abbreviations	219