

Contents

1. Introduction	1
1.1 High-Dimensional Applications	1
1.2 Motivations	4
1.3 The Objectives and Contributions	7
1.4 Organization of the Monograph	8
2. High-Dimensional Indexing	9
2.1 Introduction	9
2.2 Hierarchical Multi-dimensional Indexes	11
2.2.1 The R-tree	11
2.2.2 Use of Larger Fanouts	14
2.2.3 Use of Bounding Spheres	15
2.2.4 The <i>kd</i> -tree	16
2.3 Dimensionality Reduction	17
2.3.1 Indexing Based on Important Attributes	18
2.3.2 Dimensionality Reduction Based on Clustering	18
2.3.3 Mapping from Higher to Lower Dimension	20
2.3.4 Indexing Based on Single Attribute Values	22
2.4 Filtering and Refining	26
2.4.1 Multi-step Processing	26
2.4.2 Quantization	27
2.5 Indexing Based on Metric Distance	29
2.6 Approximate Nearest Neighbor Search	32
2.7 Summary	33
3. Indexing the Edges – A Simple and Yet Efficient Approach to High-Dimensional Range Search	37
3.1 Introduction	37
3.2 Basic Concept of iMinMax	38
3.2.1 Sequential Scan	41
3.2.2 Indexing Based on Max/Min	41
3.2.3 Indexing Based on iMax	42
3.2.4 Preliminary Empirical Study	45
3.3 The iMinMax Method	46

3.4	Indexing Based on iMinMax	47
3.5	The iMinMax(θ)	49
3.6	Processing of Range Queries	52
3.7	iMinMax(θ) Search Algorithms	57
3.7.1	Point Search Algorithm	57
3.7.2	Range Search Algorithm	57
3.7.3	Discussion on Update Algorithms	58
3.8	The iMinMax(θ_i)	58
3.8.1	Determining θ_i	59
3.8.2	Refining θ_i	62
3.8.3	Generating the Index Key	63
3.9	Summary	64
4.	Performance Study of Window Queries	65
4.1	Introduction	65
4.2	Implementation	65
4.3	Generation of Data Sets and Window Queries	66
4.4	Experiment Setup	66
4.5	Effect of the Number of Dimensions	67
4.6	Effect of Data Size	69
4.7	Effect of Skewed Data Distributions	70
4.8	Effect of Buffer Space	76
4.9	CPU Cost	77
4.10	Effect of θ_i	78
4.11	Effect of Quantization on Feature Vectors	80
4.12	Summary	83
5.	Indexing the Relative Distance – An Efficient Approach to KNN Search	85
5.1	Introduction	85
5.2	Background and Notations	86
5.3	The iDistance	87
5.3.1	The Big Picture	88
5.3.2	The Data Structure	90
5.3.3	KNN Search in iDistance	91
5.4	Selection of Reference Points and Data Space Partitioning ...	95
5.4.1	Space-Based Partitioning	96
5.4.2	Data-Based Partitioning	99
5.5	Exploiting iDistance in Similarity Joins	102
5.5.1	Join Strategies	102
5.5.2	Similarity Join Strategies Based on iDistance	103
5.6	Summary	107

6. Similarity Range and Approximate KNN Searches with iMinMax	109
6.1 Introduction	109
6.2 A Quick Review of iMinMax(θ)	109
6.3 Approximate KNN Processing with iMinMax	110
6.4 Quality of KNN Answers Using iMinMax	115
6.4.1 Accuracy of KNN Search	118
6.4.2 Bounding Box Vs. Bounding Sphere	118
6.4.3 Effect of Search Radius	118
6.5 Summary	120
7. Performance Study of Similarity Queries	123
7.1 Introduction	123
7.2 Experiment Setup	123
7.3 Effect of Search Radius on Query Accuracy	123
7.4 Effect of Reference Points on Space-Based Partitioning Schemes	126
7.5 Effect of Reference Points on Cluster-Based Partitioning Schemes	127
7.6 CPU Cost	131
7.7 Comparative Study of iDistance and iMinMax	133
7.8 Comparative Study of iDistance and A-tree	134
7.9 Comparative Study of the iDistance and M-tree	136
7.10 iDistance – A Good Candidate for Main Memory Indexing?	137
7.11 Summary	139
8. Conclusions	141
8.1 Contributions	141
8.2 Single-Dimensional Attribute Value Based Indexing	141
8.3 Metric-Based Indexing	142
8.4 Discussion on Future Work	143
References	145

High-Dimensional Indexing
Transformational Approaches to High-Dimensional
Range and Similarity Searches

Yu, C.

2002, XII, 156 p., Softcover

ISBN: 978-3-540-44199-1