CHAPTER 24

Advanced Indexing Techniques

We studied the concept of indexing, as well as a number of different index structures in Chapter 14. While some index structures, such as B^+-trees, were covered in detail, others such as hashing, write-optimized indices, bitmap indices, and spatial indices were only briefly outlined in Chapter 14. In this chapter we provide further details of these index structures. We provide detailed coverage of the LSM tree and its variants. We then provide a detailed description of bitmap indices. Next, we provide more detailed coverage of spatial indexing, covering quad trees and R-trees in more detail. Finally, we cover hashing, with detailed coverage of dynamic hashing techniques.

Bibliographical Notes

The Log-Structured Merge (LSM) tree is presented in [O'Neil et al. (1996)], while the Stepped Merge tree is presented in [Jagadish et al. (1997)]. [Vitter (2001)] provides an extensive survey of external-memory data structures and algorithms.

Bitmap indices, and variants called **bit-sliced indices** and **projection indices**, are described in [O'Neil and Quass (1997)]. They were first introduced in the IBM Model 204 file manager on the AS 400 platform. They provide very large speedups on certain types of queries, and are today implemented on most database systems. Research on bitmap indices includes [Wu and Buchmann (1998), Chan and Ioannidis (1998), Chan and Ioannidis (1999)], and [Johnson (1999)].

[Samet (2006)] provides a textbook coverage of spatial data structures. [Samet (1995)] provides an overview of the large amount of work on spatial index structures. An early description of the quad tree is provided by [Finkel and Bentley (1974)]. [Samet (1990)] and [Samet (1995)] describe numerous variants of quad trees. [Bentley (1975)] describes the k-d tree, and [Robinson (1981)] describes the k-d-B tree. The R-tree was originally presented in [Guttman (1984)]. Extensions of the R-tree are presented by [Sellis et al. (1987)], which describes the R^+ tree, and [Beckmann et al. (1990)], which describes the R^* tree. These structures provide better worst case com-

plexity guarantees for search than R-trees, but at a higher space cost. [Roussopoulos et al. (1995)] describe algorithms for nearest neighbor search on R-trees.

Discussions of the basic data structures in hashing can be found in [Cormen et al. (2009)]. [Knuth (1973)] analyzes a large number of different hashing techniques. Several dynamic hashing schemes exist. Extendable hashing was introduced by [Fagin et al. (1979)]. Linear hashing was introduced by [Litwin (1978)] and [Litwin (1980)]. A performance comparison with extendable hashing is given by [Rathi et al. (1990)]. An alternative given by [Ramakrishna and Larson (1989)] allows retrieval in a single disk access at the price of a high overhead for a small fraction of database modifications. Partitioned hashing is an extension of hashing to multiple attributes, and is covered in [Rivest (1976), Burkhard (1976)], and [Burkhard (1979)].

Bibliography

[Beckmann et al. (1990)] N. Beckmann, H. P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1990), pages 322–331.

[Bentley (1975)] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching", *Communications of the ACM*, Volume 18, Number 9 (1975), pages 509–517.

[Burkhard (1976)] W. A. Burkhard, "Hashing and Trie Algorithms for Partial Match Retrieval", *ACM Transactions on Database Systems*, Volume 1, Number 2 (1976), pages 175–187.

[Burkhard (1979)] W. A. Burkhard, "Partial-match Hash Coding: Benefits of Redundancy", *ACM Transactions on Database Systems*, Volume 4, Number 2 (1979), pages 228–239.

[Chan and Ioannidis (1998)] C.-Y. Chan and Y. E. Ioannidis, "Bitmap Index Design and Evaluation", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1998), pages 355–366.

[Chan and Ioannidis (1999)] C.-Y. Chan and Y. E. Ioannidis, "An Efficient Bitmap Encoding Scheme for Selection Queries", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1999), pages 215–226.

[Cormen et al. (2009)] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd edition, MIT Press (2009).

[Fagin et al. (1979)] R. Fagin, J. Nievergelt, N. Pippenger, and H. R. Strong, "Extendible Hashing – A Fast Access Method for Dynamic Files", *ACM Transactions on Database Systems*, Volume 4, Number 3 (1979), pages 315–344.

[Finkel and Bentley (1974)] R. A. Finkel and J. L. Bentley, "Quad Trees: A Data Structure for Retrieval on Composite Keys", *Acta Informatica*, Volume 4, (1974), pages 1–9.

[Guttman (1984)] A. Guttman, "R-Trees: A Dynamic Index Structure for Spatial Searching", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1984), pages 47–57.

[Jagadish et al. (1997)] H. V. Jagadish, P. P. S. Narayan, S. Seshadri, S. Sudarshan, and R. Kanneganti, "Incremental Organization for Data Recording and Warehousing", In *Pro-*

ceedings of the 23rd International Conference on Very Large Data Bases, VLDB '97 (1997), pages 16–25.

[Johnson (1999)] T. Johnson, "Performance Measurements of Compressed Bitmap Indices", In *Proc. of the International Conf. on Very Large Databases* (1999), pages 278–289.

[Kim (1995)] W. Kim, editor, *Modern Database Systems*, ACM Press (1995).

[Knuth (1973)] D. E. Knuth, *The Art of Computer Programming, Volume 3*, Addison Wesley, Sorting and Searching (1973).

[Litwin (1978)] W. Litwin, "Virtual Hashing: A Dynamically Changing Hashing", In *Proc. of the International Conf. on Very Large Databases* (1978), pages 517–523.

[Litwin (1980)] W. Litwin, "Linear Hashing: A New Tool for File and Table Addressing", In *Proc. of the International Conf. on Very Large Databases* (1980), pages 212–223.

[O'Neil and Quass (1997)] P. O'Neil and D. Quass, "Improved Query Performance with Variant Indexes", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1997), pages 38–49.

[O'Neil et al. (1996)] P. O'Neil, E. Cheng, D. Gawlick, and E. O'Neil, "The Log-structured Merge-tree (LSM-tree)", *Acta Inf.*, Volume 33, Number 4 (1996), pages 351–385.

[Ramakrishna and Larson (1989)] M. V. Ramakrishna and P. Larson, "File Organization Using Composite Perfect Hashing", *ACM Transactions on Database Systems*, Volume 14, Number 2 (1989), pages 231–263.

[Rathi et al. (1990)] A. Rathi, H. Lu, and G. E. Hedrick, "Performance Comparison of Extendable Hashing and Linear Hashing Techniques", In *Proc. ACM SIGSmall/PC Symposium on Small Systems* (1990), pages 178–185.

[Rivest (1976)] R. L. Rivest, "Partial Match Retrieval Via the Method of Superimposed Codes", *SIAM Journal of Computing*, Volume 5, Number 1 (1976), pages 19–50.

[Robinson (1981)] J. Robinson, "The k-d-B Tree: A Search Structure for Large Multidimensional Indexes", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1981), pages 10–18.

[Roussopoulos et al. (1995)] N. Roussopoulos, S. Kelley, and F. Vincent, "Nearest Neighbor Queries", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1995), pages 71–79.

[Samet (1990)] H. Samet, *The Design and Analysis of Spatial Data Structures*, Addison Wesley (1990).

[Samet (1995)] H. Samet. "Spatial Data Structures", In *[Kim (1995)]*, pages 361–385 (1995).

[Samet (2006)] H. Samet, *Foundations of Multidimensional and Metric Data Structures*, Morgan Kaufmann (2006).

[Sellis et al. (1987)] T. K. Sellis, N. Roussopoulos, and C. Faloutsos, "The R^+-Tree: A Dynamic Index for Multi-Dimensional Objects", In *Proc. of the International Conf. on Very Large Databases* (1987), pages 507–518.

[Vitter (2001)] J. S. Vitter, "External Memory Algorithms and Data Structures: Dealing with Massive Data", *ACM Computing Surveys*, Volume 33, (2001), pages 209–271.

[Wu and Buchmann (1998)] M. Wu and A. Buchmann, "Encoded Bitmap Indexing for Data Warehouses", In *Proc. of the International Conf. on Data Engineering* (1998), pages 220–230.

Credits

The photo of the sailboats in the beginning of the chapter is due to ©Pavel Nesvadba/Shutterstock.