

Data Mining: Concepts and Techniques (2nd edition)

Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers, 2006

Bibliographic Notes for Chapter 4 Data Cube Computation and Data Generalization

Gray, Chaudhuri, Bosworth, et al. [GCB⁺97] proposed the data cube as a relational aggregation operator generalizing group-by, crosstabs, and subtotals. Harinarayan, Rajaraman, and Ullman [HRU96] proposed a greedy algorithm for the partial materialization of cuboids in the computation of a data cube. Sarawagi and Stonebraker [SS94] developed a chunk-based computation technique for the efficient organization of large multidimensional arrays. Agarwal, Agrawal, Deshpande, et al. [AAD⁺96] proposed several methods for the efficient computation of multidimensional aggregates for ROLAP servers. The chunk-based MultiWay array aggregation method for data cube computation in MOLAP was proposed in Zhao, Deshpande, and Naughton [ZDN97]. Ross and Srivastava [RS97] developed a method for computing sparse data cubes. Iceberg queries were first described in Fang, Shivakumar, Garcia-Molina, et al. [FSGM⁺98]. BUC, a scalable method that computes iceberg cubes from the apex cuboid, downwards, was introduced by Beyer and Ramakrishnan [BR99]. The Star-Cubing method for computing iceberg cubes with a dynamic star-tree structure was introduced by Xin, Han, Li, and Wah [XHLW03]. The shell-fragment-based minimal cubing approach for efficient high-dimensional OLAP introduced in this chapter was proposed by Li, Han and Gonzalez [LHG04]. Han, Pei, Dong, and Wang [HPDW01] introduced an H-Cubing method for computing iceberg cubes with complex measures using an H-tree structure. MMCubing, an efficient iceberg cube computation method that factorizes the lattice space was developed by Shao, Han, and Xin [SHX04].

Aside from computing iceberg cubes, another way to reduce data cube computation is to materialize condensed, dwarf, or quotient cubes, which are variants of closed cubes. Wang, Feng, Lu, and Yu proposed computing a reduced data cube, called a *condensed cube* [WLFY02]. Sismanis, Deligiannakis, Roussopoulos, and Kotidis proposed computing a compressed data cube, called a *dwarf cube*. Lakshmanan, Pei, and Han proposed a *quotient cube* structure to summarize the semantics of a data cube [LPH02], which was further extended to a *qc-tree structure* by Lakshmanan, Pei, and Zhao [LPZ03]. Xin, Han, Shao, and Liu [Xin+06] developed C-Cubing (i.e., *Closed-Cubing*), an *aggregation-based* approach that performs efficient closed-cube computation using a new algebraic measure called *closedness*.

There are also various studies on the computation of compressed data cubes by approximation, such as quasi-cubes by Barbara and Sullivan [BS97], wavelet cubes by Vitter, Wang, and Iyer [VWI98], compressed cubes for query approximation on continuous dimensions by Shanmugasundaram, Fayyad, and Bradley [SFB99], and using log-linear models to compress data cubes by Barbara and Wu [BW00]. Computation of stream data “cubes” for multidimensional regression analysis has been studied by Chen, Dong, Han, et al. [CDH⁺02].

For works regarding the selection of materialized cuboids for efficient OLAP query processing, see Chaudhuri and Dayal [CD97], Harinarayan, Rajaraman, and Ullman [HRU96], and Sristava, Dar, Jagadish, and Levy [SDJL96], Gupta [Gup97], Baralis, Paraboschi, and Teniente [BPT97], and Shukla, Deshpande, and Naughton [SDN98]. Methods for cube size estimation can be found in Deshpande, Naughton, Ramasamy, et al. [DNR⁺97], Ross and Srivastava [RS97], and Beyer and Ramakrishnan [BR99]. Agrawal, Gupta, and Sarawagi [AGS97] proposed operations for modeling multidimensional databases.

The discovery-driven exploration of OLAP data cubes was proposed by Sarawagi, Agrawal, and Megiddo [SAM98]. Further studies on the integration of OLAP with data mining capabilities include the proposal of DIFF and RELAX operators for intelligent exploration of multidimensional OLAP data by Sarawagi and Sathe [SS00, SS01]. The construction of multifeature data cubes is described in Ross, Srivastava, and Chatziantoniou [RSC98]. Methods for answering queries quickly by on-line aggregation are described in Hellerstein, Haas, and Wang [HHW97] and Hellerstein, Avnur, Chou, et al. [HAC⁺99]. A cube-gradient analysis problem, called *cube-graduate*,

was first proposed by Imielinski, Khachiyan, and Abdulghani [IKA02]. An efficient method for multidimensional constrained gradient analysis in data cubes was studied by Dong, Han, Lam, et al. [DHL⁺01].

Generalization and concept description methods have been studied in the statistics literature long before the onset of computers. Good summaries of statistical descriptive data mining methods include Cleveland [Cle93] and Devore [Dev95]. Generalization-based induction techniques, such as learning from examples, were proposed and studied in the machine learning literature before data mining became active. A theory and methodology of inductive learning was proposed by Michalski [Mic83]. The learning-from-examples method was proposed by Michalski [Mic83]. Version space was proposed by Mitchell [Mit77, Mit82]. The method of factoring the version space was presented by Subramanian and Feigenbaum [SF86]. Overviews of machine learning techniques can be found in Dietterich and Michalski [DM83], Michalski, Carbonell, and Mitchell [MCM86], and Mitchell [Mit97].

Database-oriented methods for concept description explore scalable and efficient techniques for describing large sets of data. The attribute-oriented induction method described in this chapter was first proposed by Cai, Cercone, and Han [CCH91] and further extended by Han, Cai, and Cercone [HCC93], Han and Fu [HF96], Carter and Hamilton [CH98], and Han, Nishio, Kawano, and Wang [HNKW98].

Bibliography

- [AAD⁺96] S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96)*, pages 506–521, Bombay, India, Sept. 1996.
- [AGS97] R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. In *Proc. 1997 Int. Conf. Data Engineering (ICDE'97)*, pages 232–243, Birmingham, England, April 1997.
- [BPT97] E. Baralis, S. Paraboschi, and E. Teniente. Materialized view selection in a multidimensional database. In *Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97)*, pages 98–12, Athens, Greece, Aug. 1997.
- [BR99] K. Beyer and R. Ramakrishnan. Bottom-up computation of sparse and iceberg cubes. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 359–370, Philadelphia, PA, June 1999.
- [BS97] D. Barbara and M. Sullivan. Quasi-cubes: Exploiting approximation in multidimensional databases. *SIGMOD Record*, 26:12–17, 1997.
- [BW00] D. Barbará and X. Wu. Using loglinear models to compress datacube. In *Proc. 1st Int. Conf. Web-Age Information Management (WAIM'2000)*, pages 311–322, Shanghai, China, 2000.
- [CCH91] Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 213–228. AAAI/MIT Press, 1991.
- [CD97] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *SIGMOD Record*, 26:65–74, 1997.
- [CDH⁺02] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 323–334, Hong Kong, China, Aug. 2002.
- [CH98] C. Carter and H. Hamilton. Efficient attribute-oriented generalization for knowledge discovery from large databases. *IEEE Trans. Knowledge and Data Engineering*, 10:193–208, 1998.
- [Cle93] W. Cleveland. *Visualizing Data*. Hobart Press, 1993.
- [Dev95] J. L. Devore. *Probability and Statistics for Engineering and the Science* (4th ed.). Duxbury Press, 1995.
- [DHL⁺01] G. Dong, J. Han, J. Lam, J. Pei, and K. Wang. Mining multi-dimensional constrained gradients in data cubes. In *Proc. 2001 Int. Conf. on Very Large Data Bases (VLDB'01)*, pages 321–330, Rome, Italy, Sept. 2001.
- [DM83] T. G. Dietterich and R. S. Michalski. A comparative review of selected methods for learning from examples. In R. S. Michalski, J. G. Carbonell, , and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach, Vol. 1*, pages 41–82. Morgan Kaufmann, 1983.

- [DNR⁺97] P. Deshpande, J. Naughton, K. Ramasamy, A. Shukla, K. Tufte, and Y. Zhao. Cubing algorithms, storage estimation, and storage and processing alternatives for OLAP. *Bull. Technical Committee on Data Engineering*, 20:3–11, 1997.
- [FSGM⁺98] M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. D. Ullman. Computing iceberg queries efficiently. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 299–310, New York, NY, Aug. 1998.
- [GCB⁺97] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery*, 1:29–54, 1997.
- [Gup97] H. Gupta. Selection of views to materialize in a data warehouse. In *Proc. 7th Int. Conf. Database Theory (ICDT'97)*, pages 98–112, Delphi, Greece, Jan. 1997.
- [HAC⁺99] J. M. Hellerstein, R. Avnur, A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth, and P. J. Haas. Interactive data analysis: The control project. *IEEE Computer*, 32:51–59, July 1999.
- [HCC93] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 5:29–40, 1993.
- [HF96] J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 399–421. AAAI/MIT Press, 1996.
- [HHW97] J. Hellerstein, P. Haas, and H. Wang. Online aggregation. In *Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'97)*, pages 171–182, Tucson, AZ, May 1997.
- [HNKW98] J. Han, S. Nishio, H. Kawano, and W. Wang. Generalization-based data mining in object-oriented databases using an object-cube model. *Data and Knowledge Engineering*, 25:55–97, 1998.
- [HPDW01] J. Han, J. Pei, G. Dong, and K. Wang. Efficient computation of iceberg cubes with complex measures. In *Proc. 2001 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'01)*, pages 1–12, Santa Barbara, CA, May 2001.
- [HRU96] V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96)*, pages 205–216, Montreal, Canada, June 1996.
- [IKA02] T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades: Generalizing association rules. *Data Mining and Knowledge Discovery*, 6:219–258, 2002.
- [LHG04] X. Li, J. Han, and H. Gonzalez. High-dimensional OLAP: A minimal cubing approach. In *Proc. 2004 Int. Conf. Very Large Data Bases (VLDB'04)*, pages 528–539, Toronto, Canada, Aug. 2004.
- [LPH02] L. V. S. Lakshmanan, J. Pei, and J. Han. Quotient cube: How to summarize the semantics of a data cube. In *Proc. 2002 Int. Conf. on Very Large Data Bases (VLDB'02)*, pages 778–789, Hong Kong, China, Aug. 2002.
- [LPZ03] L. V. S. Lakshmanan, J. Pei, and Y. Zhao. QC-Trees: An efficient summary structure for semantic OLAP. In *Proc. 2003 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'03)*, pages 64–75, San Diego, CA, June 2003.
- [MCM86] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 2*. Morgan Kaufmann, 1986.
- [Mic83] R. S. Michalski. A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach, Vol. 1*, pages 83–134. Morgan Kaufmann, 1983.

- [Mit77] T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proc. 5th Int. Joint Conf. Artificial Intelligence (IJCAI'77)*, pages 305–310, Cambridge, MA, 1977.
- [Mit82] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.
- [Mit97] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [RS97] K. Ross and D. Srivastava. Fast computation of sparse datacubes. In *Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97)*, pages 116–125, Athens, Greece, Aug. 1997.
- [RSC98] K. A. Ross, D. Srivastava, and D. Chatziantoniou. Complex aggregation at multiple granularities. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, pages 263–277, Valencia, Spain, Mar. 1998.
- [SAM98] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, pages 168–182, Valencia, Spain, Mar. 1998.
- [SDJL96] D. Sristava, S. Dar, H. V. Jagadish, and A. V. Levy. Answering queries with aggregation using views. In *Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96)*, pages 318–329, Bombay, India, Sept. 1996.
- [SDN98] A. Shukla, P. M. Deshpande, and J. F. Naughton. Materialized view selection for multidimensional datasets. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 488–499, New York, NY, Aug. 1998.
- [SF86] D. Subramanian and J. Feigenbaum. Factorization in experiment generation. In *Proc. 1986 Nat. Conf. Artificial Intelligence (AAAI'86)*, pages 518–522, Philadelphia, PA, Aug. 1986.
- [SFB99] J. Shanmugasundaram, U. M. Fayyad, and P. S. Bradley. Compressed data cubes for OLAP aggregate query approximation on continuous dimensions. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pages 223–232, San Diego, CA, Aug. 1999.
- [SHX04] Z. Shao, J. Han, and D. Xin. MM-Cubing: Computing iceberg cubes by factorizing the lattice space. In *Proc. 2004 Int. Conf. on Scientific and Statistical Database Management (SSDBM'04)*, pages 213–222, Santorini Island, Greece, June 2004.
- [SS94] S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. In *Proc. 1994 Int. Conf. Data Engineering (ICDE'94)*, pages 328–336, Houston, TX, Feb. 1994.
- [SS00] S. Sarawagi and G. Sathe. Intelligent, interactive investigation of OLAP data cubes. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, page 589, Dallas, TX, May 2000.
- [SS01] G. Sathe and S. Sarawagi. Intelligent rollups in multidimensional OLAP data. In *Proc. 2001 Int. Conf. Very Large Data Bases (VLDB'01)*, pages 531–540, Rome, Italy, Sept. 2001.
- [VWI98] J. S. Vitter, M. Wang, and B. R. Iyer. Data cube approximation and histograms via wavelets. In *Proc. 1998 Int. Conf. Information and Knowledge Management (CIKM'98)*, pages 96–104, Washington, DC, Nov. 1998.
- [WLFY02] W. Wang, H. Lu, J. Feng, and J. X. Yu. Condensed cube: An effective approach to reducing data cube size. In *Proc. 2002 Int. Conf. Data Engineering (ICDE'02)*, pages 155–165, San Francisco, CA, April 2002.
- [XHLW03] D. Xin, J. Han, X. Li, and B. W. Wah. Star-cubing: Computing iceberg cubes by top-down and bottom-up integration. In *Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03)*, pages 476–487, Berlin, Germany, Sept. 2003.
- [ZDN97] Y. Zhao, P. M. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In *Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'97)*, pages 159–170, Tucson, AZ, May 1997.