References (1)

Lucent Technologies
Bell Labs Innovations

- [AC99] A. Aboulnaga and S. Chaudhuri. "Self-Tuning Histograms: Building Histograms Without Looking at Data". *ACM SIGMOD 1999*.
- [AGM99] N. Alon, P. B. Gibbons, Y. Matias, M. Szegedy. "Tracking Join and Self-Join Sizes in Limited Storage". *ACM PODS 1999*.
- [AGP00] S. Acharya, P. B. Gibbons, and V. Poosala. "Congressional Samples for Approximate Answering of Group-By Queries". *ACM SIGMOD 2000*.
- [AGP99] S. Acharya, P. B. Gibbons, V. Poosala, and S. Ramaswamy. "Join Synopses for Fast Approximate Query Answering". *ACM SIGMOD 1999*.
- [AMS96] N. Alon, Y. Matias, and M. Szegedy. "The Space Complexity of Approximating the Frequency Moments". *ACM STOC 1996*.
- [BCC00] A.L. Buchsbaum, D.F. Caldwell, K.W. Church, G.S. Fowler, and S. Muthukrishnan. "Engineering the Compression of Massive Tables: An Experimental Approach". *SODA 2000*.
  - Proposes exploiting simple (differential and combinational) data dependencies for effectively compressing data tables.
- [BCG01] N. Bruno, S. Chaudhuri, and L. Gravano. "STHoles: A Multidimensional Workload-Aware Histogram". *ACM SIGMOD 2001*.
- [BDF97] D. Barbara, W. DuMouchel, C. Faloutsos, P. J. Haas, J. M. Hellerstein, Y. Ioannidis, H. V. Jagadish, T. Johnson, R. Ng, V. Poosala, K. A. Ross, and K. C. Sevcik. "The New Jersey Data Reduction Report". IEEE Data Engineering bulletin, 1997.

References (2)

Lucent Technologies
Bell Labs Innovations

- [BFH75] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland. "Discrete Multivariate Analysis". The MIT Press, 1975.
- [BGR01] S. Babu, M. Garofalakis, and R. Rastogi. "SPARTAN: A Model-Based Semantic Compression System for Massive Data Tables". ACM SIGMOD 2001.
  - Proposes a novel, "model-based semantic compression" methodology that exploits mining models (like CaRT trees and clusters) to build compact, guaranteed-error synopses of massive data tables.
- [BKS99] B. Blohsfeld, D. Korus, and B. Seeger. "A Comparison of Selectivity Estimators for Range Queries on Metric Attributes". ACM SIGMOD 1999.
  - Studies the effectiveness of histograms, kernel-density estimators, and their hybrids for estimating the selectivity of range queries over metric attributes with large domains.
- [CCM00] M. Charlikar, S. Chaudhuri, R. Motwani, and V. Narasayya. "Towards Estimation Error Guarantees for Distinct Values". ACM PODS 2000.
- [CDD01] S. Chaudhuri, G. Das, M. Datar, R. Motwani, and V. Narasayya. "Overcoming Limitations of Sampling for Aggregation Queries". IEEE ICDE 2001.
  - Precursor to [CDN01]. Proposes a method for reducing sampling variance by collecting outliers to a separate "outlier index" and using a weighted sampling scheme for the remaining data.
- [CDN01] S. Chaudhuri, G. Das, and V. Narasayya. "A Robust, Optimization-Based Approach for Approximate Answering of Aggregate Queries". ACM SIGMOD 2001.
- [CGR00] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim. "Approximate Query Processing Using Wavelets". VLDB 2000. (Full version to appear in The VLDB Journal)

References (3)

Lucent Technologies
Bell Labs Innovations

- [Chr84] S. Christodoulakis. "Implications of Certain Assumptions on Database Performance Evaluation". *ACM TODS* 9(2), 1984.
- [CMN98] S. Chaudhuri, R. Motwani, and V. Narasayya. "Random Sampling for Histogram Construction: How much is enough?". *ACM SIGMOD* 1998.
- [CMN99] S. Chaudhuri, R. Motwani, and V. Narasayya. "On Random Sampling over Joins". *ACM SIGMOD* 1999.
- [CN97] S. Chaudhuri and V. Narasayya. "An Efficient, Cost-Driven Index Selection Tool for Microsoft SQL Server". *VLDB* 1997.
- [CN98] S. Chaudhuri and V. Narasayya. "AutoAdmin "What-if" Index Analysis Utility". *ACM SIGMOD* 1998.
- [Coc77] W.G. Cochran. "Sampling Techniques". John Wiley & Sons, 1977.
- [Coh97] E. Cohen. "Size-Estimation Framework with Applications to Transitive Closure and Reachability". *JCSS*, 1997.
- [CR94] C.M. Chen and N. Roussopoulos. "Adaptive Selectivity Estimation Using Query Feedback". *ACM SIGMOD* 1994.
  - Presents a parametric, curve-fitting technique for approximating an attribute's distribution based on query feedback.
- [DGR01] A. Deshpande, M. Garofalakis, and R. Rastogi. "Independence is Good: Dependency-Based Histogram Synopses for High-Dimensional Data". *ACM SIGMOD* 2001.

References (4)

- [FK97] C. Faloutsos and I. Kamel. "Relaxing the Uniformity and Independence Assumptions Using the Concept of Fractal Dimension". JCSS 55(2), 1997.
- [FM85] P. Flajolet and G.N. Martin. "Probabilistic counting algorithms for data base applications". JCSS 31(2), 1985.
- [FMS96] C. Faloutsos, Y. Matias, and A. Silbershcatz. "Modeling Skewed Distributions Using Multifractals and the `80-20' Law". VLDB 1996.
  - Proposes the use of "multifractals" (i.e., 80/20 laws) to more accurately approximate the frequency distribution within histogram buckets.
- [GGM96] S. Ganguly, P.B. Gibbons, Y. Matias, and A. Silberschatz. "Bifocal Sampling for Skew-Resistant Join Size Estimation". ACM SIGMOD 1996.
- [Gib01] P. B. Gibbons. "Distinct Sampling for Highly-Accurate Answers to Distinct Values Queries and Event Reports". VLDB 2001.
- [GK01] M. Greenwald and S. Khanna. "Space-Efficient Online Computation of Quantile Summaries". ACM SIGMOD 2001.
- [GKM01a] A.C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M.J. Strauss. "Optimal and Approximate Computation of Summary Statistics for Range Aggregates". ACM PODS 2001.
  - Presents algorithms for building "range-optimal" histogram and wavelet synopses; that is, synopses that try to minimize the total error over all possible range queries in the data domain.

References (5)

Lucent Technologies
Bell Labs Innovations

- [GKM01b] A.C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M.J. Strauss. "Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries". VLDB 2001.
- [GKT00] D. Gunopulos, G. Kollios, V.J. Tsotras, and C. Domeniconi. "Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes". ACM SIGMOD 2000.
- [GKS01a] J. Gehrke, F. Korn, and D. Srivastava. "On Computing Correlated Aggregates over Continual Data Streams". ACM SIGMOD 2001.
- [GKS01b] S. Guha, N. Koudas, and K. Shim. "Data Streams and Histograms". ACM STOC 2001.
- [GLR00] V. Ganti, M.L. Lee, and R. Ramakrishnan. "ICICLES: Self-Tuning Samples for Approximate Query Answering". VLDB 2000.
- [GM98] P. B. Gibbons and Y. Matias. "New Sampling-Based Summary Statistics for Improving Approximate Query Answers". ACM SIGMOD 1998.
  - Proposes the "concise sample" and "counting sample" techniques for improving the accuracy of sampling-based estimation for a given amount of space for the sample synopsis.
- [GMP97a] P. B. Gibbons, Y. Matias, and V. Poosala. "The Aqua Project White Paper". Bell Labs tech report, 1997.
- [GMP97b] P. B. Gibbons, Y. Matias, and V. Poosala. "Fast Incremental Maintenance of Approximate Histograms". VLDB 1997.

References (6)

Lucent Technologies
Bell Labs Innovations

- [GTK01] L. Getoor, B. Taskar, and D. Koller. "Selectivity Estimation using Probabilistic Relational Models". ACM SIGMOD 2001.
  - Proposes novel, Bayesian-network-based techniques for approximating joint data distributions in relational database systems.
- [HAR00] J. M. Hellerstein, R. Avnur, and V. Raman. "Informix under CONTROL: Online Query Processing". Data Mining and Knowledge Discovery Journal, 2000.
- [HH99] P. J. Haas and J. M. Hellerstein. "Ripple Joins for Online Aggregation". ACM SIGMOD 1999.
- [HHW97] J. M. Hellerstein, P. J. Haas, and H. J. Wang. "Online Aggregation". ACM SIGMOD 1997.
- [HNS95] P.J. Haas, J.F. Naughton, S. Seshadri, and L. Stokes. "Sampling-Based Estimation of the Number of Distinct Values of an Attribute". VLDB 1995.
  - Proposes and evaluates several sampling-based estimators for the number of distinct values in an attribute column.
- [HNS96] P.J. Haas, J.F. Naughton, S. Seshadri, and A. Swami. "Selectivity and Cost Estimation for Joins Based on Random Sampling". JCSS 52(3), 1996.
- [HOT88] W.C. Hou, Ozsoyoglu, and B.K. Taneja. "Statistical Estimators for Relational Algebra Expressions". ACM PODS 1988.
- [HOT89] W.C. Hou, Ozsoyoglu, and B.K. Taneja. "Processing Aggregate Relational Queries with Hard Time Constraints". ACM SIGMOD 1989.

References (7)

Lucent Technologies
Bell Labs Innovations

- [IC91] Y. Ioannidis and S. Christodoulakis. "On the Propagation of Errors in the Size of Join Results". *ACM SIGMOD 1991.*
- [IC93] Y. Ioannidis and S. Christodoulakis. "Optimal Histograms for Limiting Worst-Case Error Propagation in the Size of join Results". *ACM TODS 18(4), 1993.*
- [Ioa93] Y.E. Ioannidis. "Universality of Serial Histograms". VLDB 1993.
  - The above three papers propose and study serial histograms (i.e., histograms that bucket "neighboring" frequency values, and exploit results from majorization theory to establish their optimality wrt minimizing (extreme cases of) the error in multi-join queries.
- [IP95] Y. Ioannidis and V. Poosala. "Balancing Histogram Optimality and Practicality for Query Result Size Estimation". *ACM SIGMOD 1995.*
- [IP99] Y.E. Ioannidis and V. Poosala. "Histogram-Based Approximation of Set-Valued Query Answers". VLDB 1999.
- [JKM98] H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K. Sevcik, and T. Suel. "Optimal Histograms with Quality Guarantees". VLDB 1998.
- [JMN99] H. V. Jagadish, J. Madar, and R.T. Ng. "Semantic Compression and Pattern Extraction with Fascicles". VLDB 1999.
  - Discusses the use of "fascicles" (i.e., approximate data clusters) for the semantic compression of relational data.
- [KJF97] F. Korn, H.V. Jagadish, and C. Faloutsos. "Efficiently Supporting Ad-Hoc Queries in Large Datasets of Time Sequences". *ACM SIGMOD 1997.*

References (8)

Lucent Technologies
Bell Labs Innovations

  - Proposes the use of SVD techniques for obtaining fast approximate answers from large time-series databases.
- [Koo80] R. P. Kooi. "The Optimization of Queries in Relational Databases". PhD thesis, Case Western Reserve University, 1980.
- [KW99] A.C. Konig and G. Weikum. "Combining Histograms and Parametric Curve Fitting for Feedback-Driven Query Result-Size Estimation". VLDB 1999.
  - Proposes the use of linear splines to better approximate the data and frequency distribution within histogram buckets.
- [Lau96] S.L. Lauritzen. "Graphical Models". Oxford Science, 1996.
- [LKC99] J.H. Lee, D.H. Kim, and C.W. Chung. "Multi-dimensional Selectivity Estimation Using Compressed Histogram Information". ACM SIGMOD 1999.
  - Proposes the use of the Discrete Cosine Transform (DCT) for compressing the information in multi-dimensional histogram buckets.
- [LM01] I. Lazaridis and S. Mehrotra. "Progressive Approximate Aggregate Queries with a Multi-Resolution Tree Structure". ACM SIGMOD 2001.
  - Proposes techniques for enhancing hierarchical multi-dimensional index structures to enable approximate answering of aggregate queries with progressively improving accuracy.
- [LNS90] R.J. Lipton, J.F. Naughton, and D.A. Schneider. "Practical Selectivity Estimation through Adaptive Sampling". ACM SIGMOD 1990.
  - Presents an adaptive, sequential sampling scheme for estimating the selectivity of relational equi-join operators.

References (9)

Lucent Technologies
Bell Labs Innovations

- [LNS93] R.J. Lipton, J.F. Naughton, D.A. Schneider, and S. Seshadri. "Efficient sampling strategies for relational database operators", Theoretical Comp. Science, 1993.
- [MD88] M. Muralikrishna and D.J. DeWitt. "Equi-Depth Histograms for Estimating Selectivity Factors for Multi-Dimensional Queries". ACM SIGMOD 1988.
- [MPS99] S. Muthukrishnan, V. Poosala, and T. Suel. "On Rectangular Partitionings in Two Dimensions: Algorithms, Complexity, and Applications". ICDT 1999.
- [MVW98] Y. Matias, J.S. Vitter, and M. Wang. "Wavelet-based Histograms for Selectivity Estimation". ACM SIGMOD 1998.
- [MVW00] Y. Matias, J.S. Vitter, and M. Wang. "Dynamic Maintenance of Wavelet-based Histograms". VLDB 2000.
- [NS90] J.F. Naughton and S. Seshadri. "On Estimating the Size of Projections". ICDT 1990.
  - Presents adaptive-sampling-based techniques and estimators for approximating the result size of a relational projection operation.
- [Olk93] F. Olken. "Random Sampling from Databases". PhD thesis, U.C. Berkeley, 1993.
- [OR92] F. Olken and D. Rotem. "Maintenance of Materialized Views of Sampling Queries". IEEE ICDE 1992.
- [PI97] V. Poosala and Y. Ioannidis. "Selectivity Estimation Without the Attribute Value Independence Assumption". VLDB 1997.

References (10)

Lucent Technologies
Bell Labs Innovations

- [PIH96] V. Poosala, Y. Ioannidis, P. Haas, and E. Shekita. "Improved Histograms for Selectivity Estimation of Range Predicates". *ACM SIGMOD 1996.*
- [PSC84] G. Piatetsky-Shapiro and C. Connell. "Accurate Estimation of the Number of Tuples Satisfying a Condition". *ACM SIGMOD 1984.*
- [Poo97] V. Poosala. "Histogram-Based Estimation Techniques in Database Systems". PhD Thesis, Univ. of Wisconsin, 1997.
- [RTG98] Y. Rubner, C. Tomasi, and L. Guibas. "A Metric for Distributions with Applications to Image Databases". IEEE Intl. Conf. On Computer Vision 1998.
- [SAC79] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. T. Price. "Access Path Selection in a Relational Database Management System". *ACM SIGMOD 1979.*
- [SDS96] E.J. Stollnitz, T.D. DeRose, and D.H. Salesin. "Wavelets for Computer Graphics". Morgan-Kauffman Publishers Inc., 1996.
- [SFB99] J. Shanmugasundaram, U. Fayyad, and P.S. Bradley. "Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions". KDD 1999.
  - Discusses the use of mixture models composed of multi-variate Gaussians for building compact models of OLAP data cubes and approximating range-sum query answers.
- [V85] J. S. Vitter. "Random Sampling with a Reservoir". *ACM TOMS, 1985.*

References (11)

Lucent Technologies
Bell Labs Innovations

- [VL93] S. V. Vrbsky and J. W. S. Liu. "Approximate—A Query Processor that Produces Monotonically Improving Approximate Answers". IEEE TKDE, 1993.
  - Uses class hierarchies on the data to iteratively fetch blocks relevant to the answer, producing tuples certain to be in the answer while narrowing the possible classes containing the answer.
- [VW99] J.S. Vitter and M. Wang. "Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets". ACM SIGMOD 1999.

- *This is only a partial list of references on Approximate Query Processing. Further important references can be found, e.g., in the proceedings of SIGMOD, PODS, VLDB, ICDE, and other conferences or journals, and in the reference lists given in the above papers.*

Additional Resources

Lucent Technologies
Bell Labs Innovations

- Related Tutorials
  - [FJ97] C. Faloutsos and H.V. Jagadish. "Data Reduction". KDD 1998.
    - http://www.research.att.com/~drknow/pubs.html
  - [HH01] P.J. Haas and J.M. Hellerstein. "Online Query Processing". SIGMOD 2001.
    - http://control.cs.berkeley.edu/sigmod01/
  - [KH01] D. Keim and M. Heczko. "Wavelets and their Applications in Databases". IEEE ICDE 2001.
    - http://atlas.eml.org/ICDE/index_html
- Research Project Homepages
  - The AQUA and NEMESIS projects (Bell Labs)
    - http://www.bell-labs.com/project/{aqua, nemesis}/
  - The CONTROL project (UC Berkeley)
    - http://control.cs.berkeley.edu/
  - The Approximate Query Processing project (Microsoft Research)
    - http://www.research.microsoft.com/research/dmx/ApproximateQP/
  - The Dr. Know project (AT&T Research)
    - http://www.research.att.com/~drknow/