References (1)

- Proposes exploiting simple (differential and combinational) data dependencies for effectively compressing data tables.
References (2)


  - 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.

  - Studies the effectiveness of histograms, kernel-density estimators, and their hybrids for estimating the selectivity of range queries over metric attributes with large domains.


  - 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.


References (3)

  - Presents a parametric, curve-fitting technique for approximating an attribute’s distribution based on query feedback.

Garofalakis & Gibbons, VLDB 2001 #3
References (4)


  - Proposes the use of “multifractals” (i.e., 80/20 laws) to more accurately approximate the frequency distribution within histogram buckets.


  - 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)

  - 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.
References (6)

  - Proposes novel, Bayesian-network-based techniques for approximating joint data distributions in relational database systems.


  - Proposes and evaluates several sampling-based estimators for the number of distinct values in an attribute column.


References (7)

  - 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.
  - Discusses the use of “fascicles” (i.e., approximate data clusters) for the semantic compression of relational data.
References (8)

- Proposes the use of SVD techniques for obtaining fast approximate answers from large time-series databases.


  - Proposes the use of linear splines to better approximate the data and frequency distribution within histogram buckets.


  - Proposes the use of the Discrete Cosine Transform (DCT) for compressing the information in multi-dimensional histogram buckets.

  - Proposes techniques for enhancing hierarchical multi-dimensional index structures to enable approximate answering of aggregate queries with progressively improving accuracy.

  - Presents an adaptive, sequential sampling scheme for estimating the selectivity of relational equi-join operators.
References (9)

  - Presents adaptive-sampling-based techniques and estimators for approximating the result size of a relational projection operation.
References (10)

  - 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.
References (11)

  - 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.


• 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

• Related Tutorials
    • http://www.research.att.com/~drknow/pubs.html
    • http://control.cs.berkeley.edu/sigmod01/
    • http://atlas.eml.org/ICDE/index_html

• Research Project Homepages
  - The AQUA and NEMESIS projects (Bell Labs)
    • http://www.bell-labs.com/project/{aqua, nemesi}/
  - 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/