Exploiting Storage Redundancy and Parallelism for Efficient Retrieval of Multimedia Data

Divyakant Agrawal & Amr El Abbadi

Department of Computer Science
University of California at Santa Barbara

Contact Information

Department of Computer Science
University of California at Santa Barbara
Santa Barbara, CA 93106
Phone: (805) 893-4321
Fax : (805) 893-8553
Email: agrawal@cs.ucsb.edu
amr@cs.ucsb.edu

WWW Page

List of Supported Students and Staff (optional)

Hakan Ferhatosmanoglu, Research Assistant
JoAnne Holliday , Research Assistant
Kevin O'Gorman, Research Assistant
Mirek Riedewald, Research Assistant
Ioana Stanoi, Research Assistant
Chengyu Sun, Research Assistant
Yi-Leh Wu, Research Assistant

Project Award Information

- Award Number: IIS-9970700
- Duration: 10/01/1999 -- 09/30/2002
- Title: Exploiting Storage Redundancy and Parallelism for Efficient Retrieval of Multimedia Data

Keywords

Multidimensional data, parallel disks, declustering, databases, digital libraries.

Project Summary

With the rapid advances in computing and communication technologies, there is increasing demand and reliance to store a large variety of data types in digital form. Once data is stored digitally, it becomes available for access and retrieval using a variety of parameters. This proposal addresses the problem of

retrieval of multimedia data, which tends to be very large and hence requires a vast amount of storage. To capture the information content of data objects succinctly, it is represented as points/vectors in a multidimensional data space. Development of efficient placement and retrieval mechanisms for multidimensional data by exploiting redundant and parallel storage is proposed. The role of data partitioning/declustering and its interaction with indexing to support efficient retrieval of multidimensional data is explored. Efficient partitioning/declustering techniques for retrieving multimedia data and NASD (networked attached secure disk) and active disk architectures developed during this proposal will be integrated with the NSF sponsored Alexandria Digital Library project at University of California at Santa Barbara.

Publications and Products

- Epidemic Quorums for Managing Replicated Data. J. Holliday, D. Agrawal, and A. El Abbadi. Proceedings of the 19th IEEE International Performance, Computing and Communications Conference (IPCCC 2000). February 2000, Phoenix, Arizona, pp. 93--100.

- Exploiting Planned Disconnections in Mobile Environemts. J. Holliday, D. Agrawal, and A. El Abbadi. Proceedings of the Tenth International Workshop on Research Issues in Data Engineering (RIDE 2000), February 2000, San Diego, California, pp. 25--30.

- Reverse Nearest neighbor Queries for Dynamic Databases. I. Stanoi, D. Agrawal, and A. El Abbadi. Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Dallas, Texas, May 2000, pp. 44--53.

- pCube: Update-Efficient Online Aggregation with Progressive Feedback and Error Bounds. M. Riedewald, D. Agrawal, and A. El Abbadi. Proceedings of the 12th International conference on Scientific and Statistical Database Management (SSDBM 2000), Berlin, Germany, July 2000, pp. 95--108

- Data Replication using Epidemic Communication. J. Holliday, D. Agrawal, and A. El Abbadi. Proceedings of the 6th International Euro-Par Conference (Euro-Par 2000), Munich, germany, August/September 2000 (Lecture Notes in Computer Science 1900, ed. Bode, Ludwig, Karl and Wismuller), pp. 427--434.

- Space-efficient Data Cubes for Dynamic Environments. M. Riedewald, D. Agrawal, A. El Abbadi, and R. Pajarola. Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000), London, UK, September 2000, (Lecture Notes in Computer Science, LNCS 1676, ed. Kambayashi, Mohania and Min Tjoa), pp. 24--34.

- View Derivation Graph with Edge Fitting for Adaptive Data Warehousing. I. Stanoi, D. Agrawal, and A. El Abbadi. Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000), London, UK, September 2000, (Lecture Notes in Computer Science, LNCS 1676, ed. Kambayashi, Mohania and Min Tjoa), pp. 67--76.

- On the Importance of Tuning in Incremental View Maintenance: An Experience Case Study. K. O'Gorman, D. Agrawal and A. El Abbadi. Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000), London, UK, September 2000, (Lecture Notes in Computer Science, LNCS 1676, ed. Kambayashi, Mohania and Min Tjoa), pp. 77--82.

- Planned disconnections for mobile databases J. Holliday, D. Agrawal, and A. El Abbadi. Proceedings of the Eleventh International Workshop on Database and Expert Systems Applications, Greenwich, UK, September 2000, pp. 165--169.

- Using multicast communication to reduce deadlock in replicated databases. J. Holliday, D. Agrawal, and A. El Abbadi. Proceedings of the 19th Symposium on reliable distributed systems (SRDS 2000), October 2000, Nurnberg, Germany, pp. 196--205.

- Vector approximation based indexing for non-uniform high dimensional data sets. H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, and A. El Abbadi. Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM 2000), November 2000, McLean, VA, pp. 202--209.

- Using wavelet decomposition to support progressive and approximate range-sum queries over data cubes. Yi-Leh Wu, D. Agrawal, and A. El Abbadi. Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM 2000), November 2000, McLean, VA, pp. 414--421.

- A comparison of DFT and DWT based similarity search in time-series databases. Yi-Leh Wu, D. Agrawal, and A. El Abbadi. Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM 2000), November 2000, McLean, VA, pp. 488--495.

- View maintenance and analytical processing at data warehouse (Invited talk) D. Agrawal, and A. El Abbadi. International Workshop on Databases in Networked Information Systems. Aizu, Japan, December 2000, pp. 107--118 (LNCS 1966).

- Flexible data cubes for online aggregation. M. Riedewald, D. Agrawal, and A. El Abbadi. Proceedings of the 8th International Conference on Database Theory (ICDT 2001), London, UK, January 2001, pp. 159--173 (LNCS 1973).

- Storage and retrieval of moving objects. Hae Don Chon, D. Agrawal, and A. El Abbadi. Proceedings of the 2nd International Conference on Mobile Data Management (MDM 2001), Hong Kong, China, January 2001, 173--185 (LNCS 1987).

- Approximate Nearest Neighbor Searching in Multimedia Databases. H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, and A. El Abbadi. Proceedings of the International Conference on Data Engineering (ICDE'2001), April 2001, Germany.

Project Impact

- The award has supported 5-6 graduate students since its inception.
- Two PhD students graduated in 2000: JoAnne Holliday is currently an Assistant Professor at Santa Clara University and Ioana Stanoi is currently a researcher at IBM T.J. Watson Labs.
- We have incorporated various aspects of our research into two graduate courses, one in databases and the other in distributed systems.
- We have collaborated with R. Pajarola at UC Irvine to develop space efficient data cube implementaions.
- Some of the proposed datacube techniques are being incorporated into the Alexandria Digital Library project to support efficient browsing of large data sets.

Goals, Objectives, and Targeted Activities

- Validate the effectiveness of hyperspaces-based partitioning through detailed analysis and simulation studies.
- Design and develop disk allocation algorithms for data spaces partitioned based on alternative geometries (i.e., hypercubes and hyperpyramids).
- Extend existing allocation methods to support range in a different form (e.g., conic) other than rectilinear. Feasibility of an integrated allocation method to support different types of range queries.
- Developed a variety of datacube implementations that support both efficient querying and updates. The proposed techniques provide an data warehouse administrator a large spectrum of tradeoff that support differing and novel OLAP applications.
- Developed novel approaches for query size estimation for query optimzation and spatial data browsing.
- Implemented and validated the effectivness of epidemic-based replication in a variety of envirnoments, including mobile settings and transaction-based applications.
- Currently investigating data partitioning approach to maintain information about mobile objects.

Area Background

Advances in processor and network technologies have catalyzed the growth of data intensive applications such as image repositories and digital libraries. The lack of commensurate improvements in storage systems has resulted in I/O becoming a major bottleneck in modern systems. It is necessary to find techniques that can improve the access performance of storage devices despite their high latency. The use of parallel I/O from multiple devices to overlap latency is a well known technique for improving performance. A key factor in exploiting parallel I/O is knowledge of the access pattern - the sets of data items that are likely to be accessed concurrently should be declustered across the disks. We have developed novel methods to overcome the bottleneck of I/O latency in data-intensive applications. Many scientific and engineering applications process large multidimensional datasets. An important access pattern for these applications is the retrieval of data corresponding to ranges of values in multiple dimensions. Performance of these accesses is limited by the poor performance of magnetic disks, largely due to high disk latencies. Tiling the multidimensional data space and exploiting parallel I/O from multiple disks is an effective technique for improving performance. The gains from parallelism are chiefly determined by the distribution of the tiles across the parallel disks such that as many accesses as possible are maximally parallelized. Several schemes for declustering multidimensional data to improve the performance of range queries have been proposed in the literature. We have extended the class of cyclic schemes which have been developed earlier for two-dimensional data to multiple dimensions. They have established fundamental properties of these allocation schemes. Based upon these properties, we have identified techniques for reducing the search space for determining good declustering schemes within the class of cyclic schemes. Through experimental evaluation, his research has established that the cyclic schemes are superior to other declustering schemes both in terms of the degree of parallelism achieved and robustness to variations in parameters. In particular, we have shown that the exhaustive cyclic scheme gives better performance than the state-of-the-art scheme for declustering multidimensional data for range queries. Range and nearest-neighbor (similarity) queries are the most important class of queries for multimedia databases. Declustering schemes tailored for improving the performance of range only or similarity only queries have been proposed in the literature. The problem of declustering for combined range and similarity queries has not been addressed in the literature. We have investigated the problem of improving the performance of combined range and similarity queries. This research has established that the Cyclic allocation schemes give the best overall performance improvement for combined queries and are most robust with respect top variations in system parameters. Most of the earlier schemes partitioned multi-dimensional data in terms of a rectangular tiles or a grid. In a recent paper, we have have shown that significant performance gains can be obtained if alternative partitioning strategies are considered. This partitioning is based concentric hyperspaces. Disk allocation methods have been developed for concentric partitioning methods. Research results indicate that partitioning based on concentric hyperspaces has a significant advantage over balanced partitioning approach for parallel I/O.

Area References

- David Hung-Chang Du, J. S. Sobolewski: Disk Allocation for Cartesian Product Files on Multiple-Disk Systems. TODS 7(1): 82-101(1982)
- Stefan Berchtold, Christian Böhm, Bernhard Braunmüller, Daniel A. Keim, Hans-Peter Kriegel: Fast Parallel Similarity Search in Multimedia Databases. SIGMOD Conference 1997: 1-12
- Sunil Prabhakar, Khaled A. S. Abdel-Ghaffar, Divyakant Agrawal, Amr El Abbadi: Cyclic Allocation of Two-Dimensional Data. ICDE 1998: 94-101
- Stefan Berchtold, Christian Böhm, Hans-Peter Kriegel: The Pyramid-Tree: Breaking the Curse of Dimensionality. SIGMOD Conference 1998: 142-153
- Hakan Ferhatosmanoglu, Divyakant Agrawal, Amr El Abbadi.Clustering Declustered Data for Efficient Retrieval. ACM CIKM 1999
- Hakan Ferhatosmanoglu, Divyakant Agrawal, Amr El Abbadi: Concentric Hyperspaces and Disk Allocation for Fast Parallel Range Searching. ICDE 1999: 608-615

```
  *All award information can be found on the on the NSF on-line
Awards Abstracts system http://www.fastlane.nsf.gov/a6/A6Start.htm.
```

Back to the IDM '01 Homepage