Annual Report for Period: 10/2004 - 10/2005
Submitted on: 06/21/2005
Principal Investigator: Yang, Li
Organization: Western Michigan Univ
Title:
Projection and Interactive Exploration of Large Relational Data

Project Participants

Senior Personnel

Name: Yang, Li
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Sanver, Mustafa
Worked for more than 160 Hours: Yes
Contribution to Project:
Mustafa Sanver is a Ph.D. student who has been working on research and development of the data visualization component.

Name: Yang, Young-ik
Worked for more than 160 Hours: Yes
Contribution to Project:
Young-ik Yang has worked on the database component.

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Findings: (See PDF version submitted by PI at the end of the report)

Training and Development:
Two graduate students have worked on this project. They are:
1. Mustafa Sanver
2. Young-Ik Yang

Outreach Activities:
No outreach activities to report in the first year.

Journal Publications

Li Yang, "Distance-preserving projection of high dimensional data for nonlinear dimensionality reduction", IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1243, vol. 26, (2004). Published


Books or Other One-time Publications

Web/Internet Site

URL(s):
http://www.cs.wmich.edu/~yang

Description:
The URL provided is the PI's home page. A web site dedicated to this project will be setup once we finish the development of the first release of the proposed software.

Other Specific Products

Contributions within Discipline:
Our primary contribution in this first year of the project is the idea of using high dimensional indexing tree as a data structure to store and organize data aggregations on the secondary storage for the purpose of visual exploration of high dimensional data. Our attention has focused on the k-d-b tree and its variations. In addition to information kept in internal pages of a k-d-b tree for range research, we use each internal node of the tree to keep aggregation information of data points in the region represented by
the node. Each node is partitioned such that all nodes at a certain depth level of the tree contain approximately the same number of data points. Therefore, data required by the visualization component can be provided by visiting internal nodes (each represents a hyper-rectangle in the high dimensional space) of the k-d-b tree at a certain depth level. Sibling internal pages of the tree are linked together for the purpose of visual panning. Zooming is supported by accessing data aggregation values at another depth level of the tree.

In the data projection component, four methods have been proposed to build connected neighborhood graphs. Using these methods, connectivity of the constructed neighborhood graphs is guaranteed. These methods have made data embedding techniques applicable to under-sampled data and non-uniformly distributed data, for example, data distributed among clusters.

We have discovered that many-to-many generalized association rules can be visually explored by a novel use of parallel coordinates. An item taxonomy tree is displayed at each coordinate. Each rule is visualized as a sequence of Bezier curves connecting items on the coordinates. The item taxonomy tree can be partly displayed by user interaction and such interaction effectively introduces a display border in the itemset lattice. Strategies for summarizing and pruning association rules have been developed. This approach allows us to visually explore many frequent itemsets and many-to-many association rules.

Contributions to Other Disciplines:
The k-d-b data aggregation tree may potentially be used in the optimization of traditional database queries such as multi-attribute join, data aggregation and OLAP queries. These exciting issues are currently under investigation.

Contributions to Human Resource Development:
The PI and two graduate students have gained great research experience in working on this project.

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Special Requirements

Special reporting requirements: None
Change in Objectives or Scope: None
Unobligated funds: less than 20 percent of current funds
Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Organizational Partners
Any Book
Any Product
Contributions: To Any Resources for Research and Education
Contributions: To Any Beyond Science and Engineering
Major Findings

Data Visualization Component:

- The two-step bootstrapping strategy of footprint splatting is applicable if the data are not aggregated into evenly partitioned data cubes. This enables us to visually explore high dimensional data by directly visualizing aggregated data stored in internal pages of a high dimensional indexing tree. Zooming is easily supported through accessing (multi-resolution) data aggregations at different levels in the internal nodes of the tree.

- For visual exploration of association rules, we have proposed to visualize generalized association rules by a novel use of parallel coordinates where each coordinate is used to display an item taxonomy tree. The display of child nodes of each node can be toggled on/off by user interaction. Such an interaction effectively introduces a display border in the itemset lattice. Only frequent itemsets or rules that are on the border are visualized. This technique allows us to visually explore many frequent itemsets and many-to-many association rules.

Database Component:

- We have found that most database research issues (multi-resolution data aggregation, interaction-driven query optimization, high dimensional data indexing) listed in the original proposal can be properly addressed if we use a k-d-b data aggregation tree as an index structure to organize data aggregations on the secondary storage. The data aggregation tree is used for two purposes: (1) to store data aggregations at multiple resolutions; (2) to serve as an indexing mechanism for high dimensional search and range query. Together with the bootstrapping strategy of footprint splatting, visual exploration of high dimensional data can be supported by interactively browsing internal pages of the data aggregation tree. Panning and zooming are directly supported by the index.

- The k-d-b data aggregation tree extends the original k-d-b tree in the following major ways: (1) values of data aggregation (count, sum, max, min, etc.) are stored in internal pages of the tree; (2) sibling internal pages are linked with each other for the purpose of visual panning. For data insertion and deletion, we have modified the k-d-b tree’s page splitting strategy: when a page is full, it splits by following the first partition of the hyper-rectangle represented by the page. This may create unbalanced tree and the minimum occupancy of each page cannot be guaranteed. We are investigating techniques to avoid this problem by using data histogram and by re-balancing the tree after bulk-loading of data records.

- The k-d-b data aggregation tree has potential applications to accelerate data aggregation queries and OLAP queries. These issues are currently under investigation.

Data Projection Component:

- Four methods ($k$-MST, Min-$k$-ST, $k$-EC and $k$-VC) have been proposed to build connected neighborhood graphs. The first three methods build $k$-edge-connected neighborhood graphs and $k$-VC builds $k$-connected neighborhood graphs. These methods can be used for both geodesic-distance-based data embedding and locality-preserving data embedding. They make the data embedding algorithms applicable to non-uniformly distributed data, including data distributed among clusters.

- Geodesic distance estimation using neighborhood graph performs better in estimating long geodesic distances than in estimating short ones. Geodesic-distance-based data embedding methods (e.g. Isomap) work well in keeping global structure of data.

- Locality-preserving embedding methods (e.g. LLE) support bidirectional mapping between high dimensional data and their projections. They also support incremental mapping of new data records. A challenge is how to find a method that keeps advantages of both geodesic-distance-preserving methods and locality preserving methods. Such a method should be scalable to large data sets.
Major Research and Education Activities

The proposed project consists of three components: data visualization, database support, and data projection. Activities in each of the three components are reported in the following:

Data Visualization Component:

Footprint splatting of \(n\)-D aggregated data has been implemented by using a two-step splatting strategy: the first step calculates 2D footprint of a \(n\)-D cube by partitioning the \(n\)-D cube into smaller cells and by integrating the footprints of the cells, each of which is approximated by a Gaussian kernel; the second step visualizes the \(n\)-D aggregated data by integrating the footprints of all data cubes, where the color and the opacity of each footprint are set according to the aggregation values of the corresponding data cube.

This two-step bootstrapping strategy can be applied to footprint calculation of arbitrarily shaped \(n\)-D regions. The strategy is still applicable if the high dimensional data are not aggregated into evenly partitioned data cubes. This enables us to visually explore high dimensional data by interactively browsing data aggregations stored in the internal nodes (which represent hyper-rectangles) of a high dimensional indexing tree.

We did experiments of using 3D footprints for footprint splatting.

We have developed techniques for pruning and visualizing frequent itemsets and many-to-many association rules. A paper on related algorithms is published on the IEEE Transactions on Knowledge and Data Engineering.

Database Component:

Visual exploration of large high dimensional data requires that the data are aggregated at multiple resolutions. Various data structures for multidimensional indexing and data aggregation have been studied. We have focused particularly on the k-d-b tree and its variations.

We have extended the k-d-b tree so that it can be used to store aggregation information of high dimensional data. The original k-d-b tree is modified in the following ways: (1) values of data aggregation (count, sum, max, min, etc.) are stored at internal pages; (2) sibling internal pages are linked to each other for the purpose of visual panning.

Data insertion and bulk loading to the k-d-b tree have also been studied. When an internal page is full and has to split, it is divided by following the first partition in the page. This avoids cascading splitting of descendent pages. However, this may create unbalanced tree and the minimum occupancy of each page cannot be guaranteed. We are investigating techniques to re-balance the tree after bulk loading of all data records. In a dynamic database, the indexing tree needs to be re-balanced periodically.

Performance of Unix memory-mapped I/O (\texttt{mmap}) has been tested.

Data Projection Component:

Existing data embedding techniques have been evaluated. Their advantages and disadvantages have been identified. Four methods have been proposed to build connected neighborhood graphs for the purpose of defining neighbors of each data point. These methods have been implemented in Matlab and experiments have been conducted on different kinds of datasets.

Papers on these methods for constructing connected neighborhood graphs have been published on IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition Letters, and on the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD’2005).

Education Activities:

Visual data exploration will be taught as one of the major parts of CS603 — Knowledge Discovery and Data Mining in Fall 2005. Course material has been prepared.