Final Report for Period: 08/2010 - 07/2011
Submitted on: 11/14/2011
Principal Investigator: Ward, Matthew O.
Award ID: 0811510
Organization: Worcester Polytech Inst
Submitted By:
Ward, Matthew - Principal Investigator
Title:
CPA-G&V: Interactive Stream Views: Visual Analysis of Streaming Data

Project Participants

Senior Personnel

Name: Ward, Matthew
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Rundensteiner, Elke
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Yang, Di
Worked for more than 160 Hours: Yes
Contribution to Project:
Di focusses on the data analytics and management issues of the project.

Name: Xie, Zaixian
Worked for more than 160 Hours: Yes
Contribution to Project:
Zaixian focusses on the visualization and interaction issues of the project.

Name: Shastri, Avani
Worked for more than 160 Hours: No
Contribution to Project:
Avani is focusing on the multi-query support for stream pattern extraction for her MS thesis research. She is also assisting with preparing a prototype to be released on our project webpage. She is supported in part on this project over the summer, and she is funded as teaching assistant throughout the regular academic year.

Undergraduate Student

Name: Spitz, Daniel
Worked for more than 160 Hours: No
Contribution to Project:
Dan has helped with development of the core stream engine, in particular its wrappers of sources and protocol for interacting with sources for initial setup.

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

We have been collaborating with colleagues from Mitre Corporation, in particular we are applying our new stream exploration technology on their GMTI stream data sets.

We have also been exchanging ideas with several individuals at Fidelity Corporation about problems they face of detecting trends in their ticker data streams. We have applied some of our techniques on actual stock data streams provided by Fidelity Corporation.

Activities and Findings

Research and Education Activities:

Visualization has been identified as a critical component for
data analysis and decision-making in a wide range of application areas, both for its ability to provide rich overviews and
to permit users to rapidly uncover patterns and outliers using
their perception. While significant research in information visualization has focused on supporting the process of scanning static data sets in search of patterns, there have been few advances in applying visualization to the analysis of continuously streaming data. However, domains such as scientific discovery, homeland security, and outpatient care have to increasingly process digital data streams of ever increasing scale and complexity in real-time.
While major activities in computer networking (sensor networks)
and database systems (stream query processing engines) have sprung up recently, the advancement of information visualization to effectively support visual stream analysis remains in its infancy.

This project address this deficiency by developing a computational infrastructure to visualize, manage, analyze, and interactively explore real-time data streams. Tasks to be undertaken within this project fall into the following categories of key activities:

(1.) Stream Data Management and Visualization:
Technology for the modeling, management, visualization and interaction of unbounded streams of data will be developed. Overload of system and display resources will be tackled by data- and user-driven control for data reduction and multiresolution visualization.

(2.) Stream Query Management and Visualization:
Methods for analysts to express their interests will include selection of features in the data visualization, execution of feature-extraction processes, and specification of query
expressions. Query view displays will allow analysts to monitor active queries, while interactions will support manipulating, linking, and deactivating query objects.

(3.) Stream Query Result Management and Visualization:
Display types and associated interaction methodologies will be devised to capture active stream results, highlighting their attributes and abstractions, while also linking them to their corresponding data.

(4.) Assessment: Following user-centered design principles, domain experts will participate in the design, development,
and testing to insure that the resulting software will be useful and usable for real-time stream analysis. Domains will include areas such as stock market trend analysis and tactical movement tracking.

Findings:

Several insights related to the activities as outlined below have resulted from our project:

1. Data Stream Management and Interaction.
Technology for the management and interaction of unbounded streams of data has been designed. For this, we have designed a framework to deal with the overload of system and display resources by providing DOI (degree of interest) functions to represent users' interest or conversely dis-interest in the particular data within a set of windows. Users can apply two types of pre-defined DOI functions for triggering data reduction. An interactive tool allows users to adjust the DOI function online, in a manner similar to transfer functions in volume visualization, to enable a trial-and-error exploration process. These techniques empower users to either explicitly or implicitly control the data reduction process.

2. Multivariate Stream Visualization.
Technology for the visualization and interaction of unbounded streams of data have been designed -- focusing on applying established multivariate display techniques to stream data. To visually convey not only the multidimensional correlations and trends, but also how patterns change over time, we have designed several layout display strategies. They include: 1) side-by-side displays each from a different time window laid out in a time line, and 2) overlay display of data from different windows within one integrated display using time-based color encoding. These techniques empower users to visually browse possibly multiple views (time window snaphots of the stream data) on the screen at the same time for data analysis goals on infinite data streams.

3. Change-centric stream compression and history matching. In order to cope with long periods of user interest, we have developed techniques for efficient hierarchical compression of data windows based on the rate of change in the data. Thus windows that exhibit small change can be merged, and the user can monitor the stream based on interactively selected change degrees. Furthermore, rather than discarding the data when buffers become full, we compress the data in a history storage system, allowing users to retrieve past changes for patterns of interest (note that only the pattern parameters are stored, rather than the original data).

4. Stream pattern detection via data mining algorithms.
Instead of attempting to apply existing data mining algorithms designed for static data to each of the windows in the possibly infinite data stream from scratch, we have developed a customized solution for incremental detection of neighbor-based patterns specific to sliding window scenarios, in particular, for density-based clusters and distance-based outliers. Incremental pattern computation in highly dynamic streaming environments is challenging, because purging a large amount of to-be-expired data from previously formed patterns may cause complex pattern changes, including migration, splitting, merging and termination of these patterns. We exploit the ``predictability'' property of sliding windows to discount the effect of expiring objects on the remaining pattern structures. Our solution achieves minimal CPU utilization, while still keeping the memory utilization linear in the number of objects in the window. Our proposed principles are general. As proof of concept, we have also successfully applied them to speed up other stream data analysis methods, including top-k most interesting objects and reverse-nearest neighbor requests.

5. Managing Multiple Pattern Mining Requests over Streaming Data.
We have observed that many pattern detection methods are parameterized, i.e., for a clustering algorithm one may be able to indicate the number of neighbors required in order for a data point to be considered core, the density of a neighborhood function, and so on. Unfortunately, analysts may not be aware of what are the best parameter settings for achieving their analysis task, especially if the stream of data is frequently changing. To tackle this, we design algorithms for the integrated processing of a shared set of parameterized data mining requests on the same data stream. These algorithms first analyze a query workload to determine interrelationships between the different mining requests (mining algorithms instrumented with particular parameter settings). Thereafter, they achieve optimal shared processing among these requests. Our experimental performance study demonstrates that this proposed shared execution method is highly effective - avoiding redundant computations and achieving effective sharing of both computations as well as memory. This strategy has been effectively applied to several classes of neighbor-based mining algorithms, including density-based clustering, outliers, and top-k.

6. Evolution Tracking of Cluster Pattern Changes.
We have designed a framework for detecting patterns across different stream windows. For this, we have designed a model for characterizing the classes of evolution that have occurred between two sets of clusters. Evolution includes both primitive changes, such as split, merge, shrink, and grow, and composite changes. Second, we have designed the corresponding display and interaction technology to allow such patterns to be displayed not only on the concrete data stream level (see the description of data stream visualization bullet above), but also at the higher level of abstraction. In the later, each cluster is mapped to one single point in the display. The size of the point is determined by the size of the membership of the corresponding cluster. The line between two points indicates via color encoding the type of evolution that occurred.

7. Evaluation Using Experimental Performance Studies. Extensive experimental studies have been conducted for both the neighborhood-based trend discovery algorithms (clusters and outliers), as well as the multi-mining request strategy for density-based clustering and top-k type object extraction. The experimental results have confirmed that our proposed methods are significantly more efficient than

state-of-the-art methods, including DBScan and incremental DBScan, in terms of both CPU utilization and memory consumption. These experiments have been conducted on synthetic data streams so to be able to stress-test the performance of the algorithms under different forms of duress. They have also been evaluated on real data streams, namely, the GMTI data set from Mitre Corporation and on public stock stream data sets.

8. Evaluation using Case Studies. We have evaluated our technologies via two case studies so far: (1) the methodology of coping with streams by windowing and sampling via DOI functions and (2) the methodology to provide abstract level change tracking support. The first case study aims to assess the relative effectiveness of the different visualization techniques to visually convey not only the multidimensional correlations and trends, but also how patterns change over time. In particular, those include two layout strategies and three ways to combine various visualization techniques. Case studies are discussed to show the effectiveness and relative applicability of the various visualization techniques. This work has been completed and is currently in submission. The second case study evaluated the effectiveness of our proposed tool suite in detecting trends in datasets over time compared to not providing this multi-level evolution tracking framework.

9. Evaluation using user studies. For many components of our system we have evaluated their effectiveness with small to medium-sized user studies, testing the speed and accuracy of the various visualization components as well as assessing the ease at which users can learn to use the system. Results are used to enhance our displays and interaction tools, or to tune the underlying algorithms, and are reported in our publications.

Training and Development:

Two Ph.D. students, one MS and one undergraduate student have been partially supported under this grant.

Zaixian Xie has focused on the display and interaction technologies associated with stream analytics. In particular, Zaixian has conducted and successfully completed his dissertation research in this area of stream analytics. Zaixian has developed a methodology for coping with continuous multivariate data streams by applying sampling and windowing strategies. This has resulted in a conference poster, a CHI workshop paper, and two conference papers; an extended paper on this work is currently in submission to a journal. He has also done presentations in the department Image Science Research Group. Zaixian has recently successfully defended his dissertation and is currently employed in Microsoft Corp.

Di Yang, also a Ph.D. student on the project, has focused on the stream analytics and data management issues. Di has developed automated techniques for the extraction of interesting trends in the data streams in particular, density-based clusters, outliers and continuous top-k interesting objects. His effort included new results in the area of incremental algorithms for continuous pattern extraction. Furthermore, he has devised several strategies for multi-query optimization of continuous mining requests. The latter was successfully applied to 4 different mining algorithms to illustrate their generality. He has conducted experimental performance studies using real data sets, including data from Mitre Corporation and stock market data. He has also conducted user studies on the overall pattern extraction framework he has developed. His work has resulted in 3 journal papers, 7 conference papers, and two SIGMOD software demonstrations. Di Yang has completed a draft of his dissertation, and anticipates defending his dissertation this academic year 2012. Di Yang is currently employed at Oracle Corporation, Nashua.

The MS student, Avani Shastri, has designed and implemented shared query execution strategies for multiple top-k exploration queries in the context of our stream framework. The results have been integrated with the stream code base. This forms the core of her MS thesis research, which was completed during the summer of 2011, and has resulted in a conference paper. She is employed at Oracle Corporation.

The BS student, Daniel Spitz, has worked jointly with Di Yang on the development of some of the core modules of the stream engine, in particular, the stream source server.

Outreach Activities:

In the K12 REK project supported by NSF in 2008/2009 by one of the PIs, we have worked with K12 students on small research projects with the goal to increase their awareness and interest in science and technology. In this outreach context, we have made an effort to expose K12 students to visual stream exploration, as studied and developed as part of this research NSF grant.

Journal Publications

Di Yang, E. A. Rundensteiner, and M. O. Ward, "A Shared Execution Strategy for Multiple Pattern Mining Requests over Streaming Data", Very Large Databases Conference, p. 1, vol. 1, (2009). Accepted,

Yang, Di, Rundensteiner, E.A., and Ward, M., "Neighbor-Based Pattern Detection for Windows Over Streaming Data", Extending Database Technology (EDBT), p. 1, vol. 1, (2009). Published,

Zaixian Xie, "Towards Exploratory Visualization of Multivariate Streaming Data", IEEE Vis/InfoVis/VAST Doctoral Colloquium, p. 1, vol. 1, (2007). Published,

Z. Xie, M. O. Ward, and E. A. Rundensteiner, "Exploring multivariate data streams using windowing and sampling strategies", Interacting with Temporal Data, CHI'2009 Workshop, p. 1, vol. 1, (2009). Published,

Di Yang, Zhenyu Guo, Zaixian Xie, Elke A. Rundensteiner, Matthew O. Ward, "Interactive visual exploration of neighbor-based patterns in data streams", ACM SIGMOD Conference Proceedings - software demonstration, p. 1, vol. , (2010). Published,

Zaixian Xei, Zhenyu Guo, Matthew O. Ward, Elke A. Rundesteiner, "Operator-Centric Design Patterns for Information Visualization Software", Proc. of SPIE-IS&T Electronic Imaging, SPIE, Visualization and Data Analysis, p. 1, vol. 7530, (2010). Published,

Xie, Z., Ward, M.O., and Rundensteiner, E.A., "Visual exploration of stream pattern changes using a data-driven framework", Proc. International Symposium on Visual Computing, p. 522, vol. , (2010). Published,

D. Yang, A. Shastri, E. Rundensteiner, M. Ward, "An optimal strategy for monitoring top-k queries in streaming windows", Proc. EDBT, p. 57, vol. , (2011). Published,

Abhishek Mukherji, Elke A. Rundensteiner, Matthew O. Ward, "Achieving High Freshness and Optimal Throughput in CPU-limited Execution of Multi-Join Continuous Queries", Proc. British National Conference on Databases, p. 1, vol. , (2011). Published,

Di Yang, Elke A. Rundensteiner, and Matthew O. Ward, "Multiple Query Optimization for Neighbor-Based Pattern Mining Requests over Data Streams", ACM Transactions on Database Systems, p. , vol. , (2011). Accepted,

Di Yang, Zhenyu Guo, Elke A. Rundensteiner, and Matthew O. Ward, "CLUES: A Unified Framework Supporting Interactive Exploration of Density-Based Clusters in Streams", Proc. 20th ACM Conference on Information and Knowledge Management, p. 1, vol. , (2011). Published,

Avani Shastri, Di Yang, Elke A. Rundensteiner, and Matthew O. Ward, "MTopS: Scalable Processing of Continuous Top-K Multi-Query Workloads", Proc. 20th ACM Conference on Information and Knowledge Management, p. 1, vol. , (2011). Published,

Di Yang,Elke A. Rundensteiner, and Matthew O. Ward, "Summarization and Matching of Density-Based Clusters in Steaming Environments", PVLDB, p. 121, vol. 5, (2012). Accepted,

Di Yang,Elke A. Rundensteiner, and Matthew O. Ward, "Mining Neighbor-Based Patterns from Streaming Data", Data and Knowledge Engineering Journal, p. , vol. , (2011). Submitted,

Books or Other One-time Publications

Web/Internet Site

URL(s):

http://davis.wpi.edu/~xmdv

Description:

Other Specific Products

Product Type:

Software demonstration at major conference.

Product Description:

Di Yang, Zhenyu Guo, Zaixian Xie, Elke A. Rundensteiner, Matthew O. Ward, " Interactive visual exploration of neighbor-based patterns in data streams", Software Demonstration, SIGMOD Conference, 2010.

Our work has resulted in a demonstration at ACM SIGMOD in June 2010. We are currently refining the software based on the feedback received by attendees of our demonstration, and aim to have a release of this system available for public download.

Sharing Information:

We plan to release this software to the general public, once code and usage documentation and installation guides have been developed.

Contributions

Contributions within Discipline:

Within the visualization field, we have developed a new approach to visual stream analysis that allows us to redeploy established multivariate display techniques within this new context of streaming data exploration. The key idea is to either lay side-by-side several time-related displays next to each other, from one stream window to the next, or to overlay them all within one single integrated window (each encoded with time-driven color). This enables the analyst to not only see many time windows worth of data concurrently, but also allows her to more easily track changes over time.

We also have created a tool for monitoring the evolution of clusters in a data stream. The cluster view shows the size and relative distances between clusters at each time step, and hypothesizes about likely splits, merges, insertions, and deletions of clusters. The corresponding data views are shown below this view, color-coded to match the cluster view.

Another contribution is a mechanism for change-based history compression, which allows users to view longer segments of the data stream by merging windows where little change has occurred. Users can also query past data for changes similar to what is occurring in the present and recent windows. At present, the work focuses on changes to regression parameters as time progresses; future work will include exploring other measures of change.

We have also contributed to the data modeling and management field by creating novel algorithms for the shared execution of multiple data mining requests on a data stream. While such technology has been applied to core SQL type operations, such as sharing joins across multiple SQL query plans, this is the first time that we are aware of that shared execution of data mining has been pursued. Our results based on the MITRE data set are encouraging, indicating the effectiveness of our methodology to be able to execute 100 or even 1,000 mining requests concurrently on a data stream within both limited memory footprint and CPU utilization. This represents a major achievement in the data management field. We have demonstrated the application of our core principles on other data mining request types, which helps to establish their scope of applicability.

Contributions to Other Disciplines:

Providing technology for discovering trends within data streams has the potential to lead to contributions in multiple other domains. In particular, by providing domain experts with the tools to conduct their scientific exploration on data streams in real-time in a more effective manner, new scientific discoveries may be facilitated by our technology.

Contributions to Human Resource Development:

As indicated earlier, two Ph.D. students, one MS student and one undergraduate student have been trained in state-of-the-art technology as part of this project effort.

Contributions to Resources for Research and Education:

We plan to distribute the software generated by our research to the public domain -- after having developed a first prototype of both the back-end stream engine as well as the front-end display and interaction support modules. This prototype technology has been demonstrated at ACM SIGMOD in summer 2010. We are finalizing this technology, extending upon its features, making it robust and documented - so to be able to release our system to the public domain.

Researchers at several universities and research labs use our exploratory visualization software, XmdvTool, for their work, and educators at numerous schools use the software in their courses. We expect to be able to achieve the same popularity with our new stream technology, once the software development has been completed.

Contributions Beyond Science and Engineering:

Exploratory stream analysis touches nearly every aspect of our society, from medicine to manufacturing to homeland security. Interactive visualization of data streams has been recognized as a critical technology in all such fields. Over the years, the techniques we developed for this project may get integrated into commercial visualization tools and thus be used in a wide range of disciplines.

Conference Proceedings

Categories for which nothing is reported:

Organizational Partners

Any Book

Any Conference