FINAL REPORT FOR AWARD # 0238386

Duke University
CAREER: Techniques and Applications of Derived Data Maintenance

Participant Individuals:

Graduate student(s) : Hao He; Adam Silberstein; Junyi Xie; Badrish Chandramouli
Research Experience for Undergraduates(s) : Congyi Wu; Gregory Filpus

Participants' Detail

Partner Organizations:

Other collaborators:

Pankaj K. Agarwal, Shivnath Babu, Carla S. Ellis, and Kamesh Munagala. Faculty members in the Department of Computer Science, Duke University.

Rebecca L. Braynard, Jeff M. Phillips, Ke Yi, and Hai Yu. Students in the Department of Computer Science, Duke University.

Alan E. Gelfand and Gavino Puggioni. Faculty member and graduate student in the Institute of Statistics and Decision Sciences, Duke University.

James S. Clark. Faculty member in the Duke University School of Environment.

Cliburn Chan, Lindsay G. Cowell, and Thomas B. Kelper. Faculty members at the Center for Bioinformatics and Computational Biology, Duke University.

David F. Kong. Duke University Medical Center.

Ioana Stanoi, Haixun Wang, and Philip S. Yu. IBM T. J. Watson Research Center.

Yuguo Chen. Faculty member in the Department of Statistics, University of Illinois at Urbana-Champaign.

Jeffrey Yu Xu. Faculty member at the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong.

Amin Vahdat. Faculty member in the Computer Science Department, University of California at San Diego.

AnHai Doan and Fei Chen. Faculty member and graduate student in the Computer Sciences Department, University of Wisconsin at Madison.

Raghu Ramakrishnan. Yahoo! Research.

Activities and findings:

Research and Education Activities:

Introduction

The focus of this CAREER project has been on the techniques and applications of derived data maintenance. Derived data is the result of applying some transformation, structural or computational, to base data. The use of derived data to facilitate access to base data is a recurring technique in many areas of computer science. Used in hardware and software caches, derived data speeds up access to base data. Used in replicated systems, it improves reliability and performance of applications in a wide-area network. Used as index structures, it provides fast alternative access paths to base data. Used as materialized views in databases or data warehouses, it improves the performance of complex queries over base data. Used as synopses, it provides fast, approximate answers to queries or statistics needed for cost-based optimization. Derived data may vary in complexity: it can be a simple copy of base data, in the cases of caching and replication, or it can be the result of complex transformations, in the cases of indexes and materialized views. Derived data may also vary in accuracy: caches and materialized views are usually exact, while synopses are approximate. Regardless of the varying forms, purposes, complexity, and accuracy of derived data, it must be maintained when base data is updated. Thus, derived data maintenance is a fundamental problem in computer science. It is also an evolving problem: existing techniques are constantly challenged by the explosive growth in data volume and number of data producers and consumers, and by increasing diversity in data formats and storage and communication media.

Traditionally, derived data maintenance has been tackled separately in different contexts, e.g., index updates and materialized view maintenance in databases, cache coherence and replication protocols in distributed systems. Although they share the same underlying theme, these techniques have been developed and applied largely disjointly. Newer and more complex data management tasks, however, call for creative combinations of the traditionally separate ideas. Semantic caching, which has received tremendous interests recently for its applications in caching dynamic Web contents, is a good example of incorporating the idea of materialized views into a cache. With ``outside-the-box'' thinking such as semantic caching, we seek to discover more techniques that combine multiple flavors of derived data to provide better solutions to problems.

Research Progress over the Five Years

In the following, we first summarize the progress made during the first four years of this project; additional details can be found in the annual project reports, available from the project homepage (http://www.cs.duke.edu/dbgroup/ddm). We will then focus on reporting progress made during the final year of the project.

In the first year of this project, we made progress on the following specific research problems: (1) caching for view maintenance; (2) caching for stream data processing; (3) caching for XML indexing; (4) incremental maintenance of XML structural indexes. A detailed description of our contributions can be found below in our 2003-2004 project report.

In the second year of this project, we broadened our study of derived data maintenance to continuous queries and subscriptions, and began investigating derived data in a network setting. We made contributions to the following specific research problems: (1) caching for stream data processing (continued from the first year); (2) incremental maintenance of order-based XML labeling; (3) asymmetric batch incremental view maintenance; (4) scalable continuous query processing; (5) querying networked data. A detailed description of our contributions can be found below in our 2004-2005 project report.

In the third year of this project, we continued to expand the applications of our derived data maintenance techniques in the following areas: (1) derived data in scalable continuous query processing (continued from the second year); (2) derived data in wide-area network application (including continued efforts on networked data querying from the second year and new thrust on wide-area publish/subscribe); (3) derived data for graph reachability; (4) derived data in wireless sensor networks; (5) asymmetric batch incremental view maintenance (continued from the second year). A detailed description of our contributions can be found below in our 2005-2006 project report.

In the fourth year of this project, we continued to investigate problems related to derived data in a number of areas: (1) query suspend and resume (continued from work started in the second year); (2) derived data for queries over graph-structured data (continued from earlier work on reachability, with emphasis shifted from supporting low-level primitives to user-level queries); (3) derived data in scalable continuous query processing (continued from work in the third year but now with a more practical slant); (4) derived data in computational biology workflows; (5) derived data in wireless sensor networks (continued from the third year).

In the fifth (and final) year of this project, we worked on the following problems related to derived data:

1. Derived data in scalable continuous query processing. Continuous query processing has attracted much interest from the database community recently because of its wide range of traditional and emerging applications, e.g., trigger and production rule processing, data monitoring, stream processing, and publish/subscribe systems. In contrast to traditional query systems, where each query is run once against a snapshot of the database, continuous query systems support standing queries that continuously generate new results (or changes to results) as data updates continue to arrive in a stream. In this sense, continuous query processing has much in common with incremental view maintenance, and can be regarded also as a problem of derived data maintenance. One of the main challenges in continuous query processing is how to handle a large number of continuous queries in a scalable way. For each incoming data update, the system needs to identify the subset of continuous queries whose results are affected by the data update, and compute changes to these results. If there are many continuous queries, a brute-force approach that processes each of them in turn will be inefficient and unable to meet the response-time requirement of most target applications. One important insight gained by research on scalable continuous query processing is the interchangeable roles played by queries and data. In continuous query systems, continuous queries can be treated as data, while each data update can be treated as a query requesting the subset of continuous queries affected by the update. Thus, it is natural to apply indexing and query processing techniques traditionally intended for data to continuous queries.

Since Year 3 of this project, we have been developing efficient processing techniques in collaboration with Pankaj K. Agarwal, a colleague in Duke Computer Science. In Year 3, we focused mostly on theoretical results (ISSAC 2005); in Year 4, we began to work on more practical extensions, implementation, and experimental validation (VLDB 2006). In Year 5, we continued to address the practical aspect of the problem. Our observation is that during the course of continuous query processing, incoming data with different characteristics may warrant change in processing strategy. In a system with a large number of continuous queries, the cost of processing each incoming tuple can be substantial. Given the high cost-saving potential, we argue that is beneficial to support more aggressive data-sensitive processing that makes cost-based decisions to switch among alternative query plans for every input tuple. This approach can be regarded as another example of interchanging the roles of queries and data, where each incoming tuple is optimized as a separate query over the set of the continuous queries. To this end, we have developed and implemented a flexible, cost-based processing framework, which is able to dynamically route incoming data to the most promising query plan based on runtime data and query characteristics. We identify the statistics we need to monitor and build cost models for alternative processing strategies. Combined with efficient group-processing techniques we developed earlier, our approach delivers significantly better performance than traditional approaches for processing a large number of continuous queries. This work is currently under submission to a journal (in preparation for the second round of reviews).

Our work in this direction has also seeded the ProSem project, which has been separately funded since September 2007 (NSF grant IIS-0713498). ProSem aims at building a next-generation, Internet-scale publish/subscribe system, which supports efficient and powerful subscription functionalities, allowing users to control precisely what they want and when they want it. Publish/subscribe is natural extension of continuous queries. ProSem builds on the results of this CAREER award (ISSAC 2005, SIGMOD 2006, VLDB 2006), and seeks to develop end-to-end solutions consisting of techniques from subscription processing and indexing to dissemination network design. Latest results from the ProSem project include solutions for subscriptions with value-based notification conditions (VLDB 2007), select-join subscriptions (VLDB 2008), as well as a system demonstration (SIGMOD 2008). Although primarily funded through NSF grant IIS-0713498, these recent results also draw support from this CAREER award.

2. Derived data in wireless sensor networks. Continuing the progress made in previous years, we building a wireless sensor network in Duke Forest to study how various environmental variables influence forest growth, in collaboration with Duke University School of Environment. Wireless sensor networks are capable of generating a vast amount of data; this data, however, must be sparingly extracted to conserve energy, usually the most precious resource in battery-powered nodes.

Our collaborative team is developing novel OS, networking, and data service layers that implement a dynamic, data-driven approach to energy-efficient sensing and communication. While the majority of this effort has been funded under NSF's DDDAS program beginning January 2006 (CNS-0540347), we have been actively studying the use of derived data in sensor data processing through the support of this CAREER award. Continuing with our vision of data-driven processing outlined in CIDR 2007, we considered the particularly thorny challenge of making suppression work in the presence of failures. Suppression is an effective strategy of conserving energy in sensor networks: Instead of transmitting all readings to the base station, a node can suppress a reading if its value can be predicted from previously transmitted readings (to within a user-specified bound). The stream of data received at the base station can be regarded as data derived from the stream of readings taken at sensors using a suppression procedure. The base station needs to interpret this derived data. A critical challenge is message failure, to which sensor networks are particularly vulnerable. Failure creates ambiguity: a non-report may either be a suppression or a failure. Inferring the correct values for missing data and learning the parameters of the underlying process model become quite challenging. We have proposed a novel solution, BaySail, which incorporates the knowledge of the suppression scheme and application-level redundancy in Bayesian inference. Experimental evaluation shows application-level redundancy outperforms retransmissions and basic sampling in both cost and accuracy of inference. The framework shows suppression schemes are generally effective for data collection, despite the presence of failures. This work, published in VLDB 2007, was co-sponsored by CNS-0540347 and this CAREER award.

3. Recovering data from derived data. There has been a recent resurgence of interest in research on noisy and incomplete data. Many applications require information to be recovered from such data. For example, in sensor data processing, errors may occur in collection and transmission; in privacy-preserving data publishing, only summarized or perturbed data is available for analysis. Generally, the problem is how to recover hidden base data from derived data, where the derivation procedure can drop data, add noise, summarize data, etc. Ideally, an approach for recovery should have the following features. First, it should be able to incorporate prior knowledge about the data, even if such knowledge is in the form of complex distributions and constraints for which no close-form solutions exist. Second, it should be able to capture complex correlations and quantify the degree of uncertainty in the recovered data, and further support queries over such data. The database community has developed a number of approaches for information recovery, but none is general enough to offer all above features. To overcome the limitations, we take a significantly more general approach to information recovery based on sampling. We improve efficiency by applying sequential importance sampling, a technique from statistics that works for complex distributions and dramatically outperforms naive sampling when data is constrained. We illustrate the generality and efficiency of this approach in two application scenarios: cleansing RFID data, and recovering information from published data that has been summarized and randomized for privacy. This work has been published in ICDE 2008.

4. Derived data in computational biology workflows. Science is increasingly data-driven. Data comes in huge quantities and many forms, and using data effectively becomes a daunting task. The challenges include discovery, provenance, and dependency, just to name a few. Maintaining adequate metadata is crucial in solving any of these problems. In collaboration with Duke University Laboratory of Computational Immunology, we have been building a system called ERS (Enhanced Repository Service), which captures the metadata on lineage, dependency and versioning. The system will allow users to visualize, explore, and query the graph-structured data, and provides a subscription service that notifies the users whenever an update could potentially affect a derived dataset of interest. This effort has been jointly supported by this CAREER award and NIH. In Year 5, we have completed the implementation of ERS v1, which is currently in use by our collaborators. In addition, we studied the problem of supporting a useful and efficient subscription service for lineage, dependency, and versioning information. A user interested in keeping a particular dataset up to date, for example, can define a subscription that specifies how to notify the user when relevant updates have occurred. Naive subscription processing methods have trouble scaling up to complex dependencies and large numbers of datasets and subscriptions. We propose scalable methods for subscription processing and demonstrate that they significantly outperform the naive ones. This work was the MS thesis of Pradeep Gunda, who

graduated in December 2007.

5. Derived data in information extraction. Most current information extraction (IE) approaches have considered only static text corpora, over which we typically have to apply IE only once. Many real-world text corpora however are dynamic. They evolve over time, and to keep extracted information up to date, we often must apply IE repeatedly, to consecutive corpus snapshots. Inspired by materialized views, a prime example of derived data in databases, we explore the approach of incrementally maintaining the IE results when the corpora change, in collaboration with University of Wisconsin and Yahoo! Research. The result is Cyclex, an approach that efficiently executes such repeated IE, by recycling previous IE efforts. Specifically, given a current corpus snapshot U, Cyclex identifies text portions of U that also appear in the previous corpus snapshot V. Since Cyclex has already executed IE over V, it can now recycle the IE results of these parts, by combining these results with the results of executing IE over the remaining parts of U, to produce the complete IE results for U. Realizing Cyclex raises many challenges, including modeling information extractors, exploring the trade-off between runtime and completeness in identifying overlapping text, and making informed, cost-based decisions between redoing IE from scratch and recycling previous IE results. The work was published in ICDE 2008.

Educational Activities over the Five Years

I have continued to incorporate current research topics into both undergraduate and graduate database course at Duke University. Over the course of the project, I offered the undergraduate database course in Fall 2003, Fall 2004, Fall 2005, Fall 2006, and Fall 2007, and the graduate database course in Spring 2004 and Spring 2005. These offerings covered a substantial amount of material drawn from the latest research.

In Spring 2007, I offered a graduate-level topics course on sensor data processing, which covered the most recent research advances in this field. In Fall 2007, I co-taught a course on sensor networks for environmental monitoring at SAMSI (Statistical and Applied Mathematical Sciences Institute), which was cross-listed at Duke, North Carolina State, and UNC Chapel Hill. In Spring 2008, I offered a graduate research seminar on databases with a project component; this course helped a batch of new graduate students get up to speed with research.

In Summer 2007, I worked with undergraduate summer interns (funded through an REU supplement to this CAREER award), both of whom are working on application areas of derived data: Gregory Filpus is working on query processing in wireless sensor networks, while Congyi Wu is working on a system for managing lineage, dependency, and versioning of derived datasets in computational biology workflows.

In addition to undergraduate students directly funded by my CAREER award's REU supplement, I am have actively involved in supervising independent studies and honors thesis research of undergraduate students. Christopher N. Bond, a student from my undergraduate database class, pursued his undergraduate honors thesis research with me based on his course project. The result is a BS degree with High Distinction in Spring 2005, and this work ultimately led to a SIGMOD 2007 paper. Tyler Brock, also a student from my undergraduate database class, pursued his undergraduate honors thesis research with me and graduated with Distinction in Spring 2007. Congyi Wu, who worked as a summer intern supported by REU, also pursued independent studies with me.

This CAREER project has also produced an impressive list of graduate student alumni: four PhD and one MS. See the training and development section of this report for details.

Findings:

We have made significant progress in studying the derived data maintenance problem in multiple application domains, including view maintenance, data warehousing, indexing and querying XML and graph-structured data, continuous query processing in stream and publish/subscribe systems, sensor networks, etc. Published results from this award include:

Traditional settings:

* An efficient method for top-k view maintenance that incorporates the idea of caching (ICDE 2003).
* A new approach to batch incremental view maintenance that exploits asymmetry in maintenance cost components (ICDE 2005 and ESA 2005).
* Efficient support for database query suspend and resume (SIGMOD 2007).

XML and graph-structured data:
* A novel XML structural index (ICDE 2004) utilizing derived data at multiple resolutions.
* Efficient incremental maintenance algorithms for XML structural indexes (SIGMOD 2004), which incorporate the use of auxiliary data.
* Efficient maintenance of order-based labeling for dynamic XML documents, with different degrees of materialization to provide a tradeoff between query and update performance (ICDE 2005).
* A hybrid labeling scheme for graph reachability that identifies different types of substructures within a graph and encodes them using techniques suitable to the characteristics of each (CIKM 2005).
* A labeling scheme supporting constant-time graph reachability queries while remaining space-efficient for sparse graphs (ICDE 2006).
* A bi-level indexing and query processing scheme for top-k keyword search on graphs (SIGMOD 2007).
* PhD dissertation of Hao He, July 2007.

Wide-area network querying:
* A system for distributed network monitoring and resource querying by intelligently placing, locating, and managing bounded approximated caches across the network (DASFAA 2006).

Continuous query processing in stream and publish/subscribe systems:
* Framework and techniques for managing the state of a stream join to maximize result completeness, which is related to the classic caching problem (SIGMOD 2005).
* New, input-sensitive approaches to scalable processing of continuous join queries (ISAAC 2005).
* A new approach towards wide-area publish/subscribe that examines the spectrum of possibilities of interfacing subscription processing and notification delivery for more efficient support of stateful subscriptions (SIGMOD 2006).
* Practical extensions and improvements to the ISAAC 2005 paper, including hotspot-based processing and experimental evaluation (VLDB 2006).
* Internet-scale publish/subscribe for subscriptions with value-based notification conditions (VLDB 2007), and for select-join subscriptions (VLDB 2008).
* Demonstration of ProSem, a scalable publish/subscribe system supporting complex, stateful subscriptions over a wide-area network (SIGMOD 2008).
* PhD dissertation of Junyi Xie, September 2007.
* PhD dissertation of Badrish Chandramouli, July 2008.

Applications of derived data in sensor networks:
* A model-driven approach to snapshot top-k queries that uses samples of past sensor readings and linear programming for optimization (ICDE 2006).
* Energy-efficient algorithms for continuously monitoring extreme values using a hierarchy of local constraints, or thresholds (SIGMOD 2006).
* Energy-efficient monitoring using spatio-temporal suppression and a chain of locally monitored constraints for reconstructing the global view (poster paper in ICDE 2006; full paper in SIGMOD 2006).
* Vision and challenges of data-driven processing, whose goal is to support continuous sensor data collection without continuous reporting; it uses models for optimization and interpretation, but never substitutes model for actual data (CIDR 2007).
* Efficient support for computing multiple aggregates in a sensor network, where the relationship between sources and destinations of aggregates is many-to-many (ICDE 2007).
* BaySail, an approach to handling failures in suppression-enabled sensor networks, which incorporates the knowledge of the suppression scheme and application-level redundancy in Bayesian inference (VLDB 2007).
* PhD dissertation of Adam Silberstein, February 2007.

Other settings:
* Recovering data from derived data that has been subject to drops, noise, and summarization (ICDE 2008).
* Incremental information extraction by detect changes in corpora and reusing previous extraction results (ICDE 2008).
* Managing lineage, dependency, and versioning information for computational biology workflows (MS thesis of Pradeep Gunda, Fall

2007).

The software artifacts produced by this award include the ProSem system for Internet-scale publish/subscribe (demonstrated at SIGMOD 2008) and ERS (Enhanced Repository Service) v1 (in use by collaborators in computational immunology).

For detailed descriptions of the above findings, please refer to the section of this report on research and education activities.

Training and Development:

I have advised the following students in the context of this project:

Ph.D. students:

Adam Silberstein (defended in February 2007);
Hao He (defended in July 2007);
Junyi Xie (defended in September 2007);
Badrish Chandramouli (defended in July 2008).

M.S. students:

Zhihui Wang (thesis completed in Summer 2003);
Wenbin Pan (thesis completed in Fall 2004);
Pradeep K. Gunda (thesis completed in Fall 2007).

Undergraduate students:

Christopher N. Bond (BS with High Distinction, 2005);
Gregory Filpus;
Congyi Wu;
Tyler J. Brock (BA with Distinction, 2007).

Outreach Activities:

I have been active in running the Carolina Database Research Group (http://www.cs.duke.edu/cdb/) with a group of database researchers in North Carolina, including members from Duke, North Carolina State University, University of North Carolina at Chapel Hill, Charlotte, and Greensboro. We hold monthly meetings and are currently running a seminar series, which have been a great resource for facilitating student and faculty interaction across institutions and attracting student interests in database research. I was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

In addition to serving on program committees, I also served as program committee co-chair of DMSN (International Workshop on Data Management for Sensor Networks) 2007, and general co-chair of DMSN 2008.

I have given a large number of research seminars and presentations on the results from this award (the following list does not include conference presentations):

* 'Data-Driven Processing in Sensor Networks,' seminars at University of Pennsylvania, University of Waterloo, and New England Database Society, April 2007 - October 2007.

* 'Scalable Continuous Query Processing and Result Dissemination,' seminars at IBM T. J. Watson Research Center, University of Maryland at College Park, University of Pittsburgh/Carnegie Mellon University Joint Database Seminar, Brown University, University of Illinois at Urbana-Champaign, and University of California at Berkeley, February 2006 - December 2006.

* 'Layers and Boxes: Efficient and Maintainable Indexes for XML,' seminar at IBM T. J. Watson Research Center, July 2004.

For further outreach, I have also given higher-level talks aimed at more general audiences:

* 'Thoughts on Data Sharing: A Database Researcher's Perspective,' presentation at the Primate Life History Working Group Meeting, NESCent (National Evolutionary Synthesis Center), August 2007.

* 'Continuous Query Processing over Networked Data,' presentation at

IBM Research Triangle Park University Day, October 2006.

* Panel discussion at SIGMOD '06 Life after Graduation Symposium, June 2006.

* 'Querying Networked Data,' presentation at IBM Research Triangle Park University Day, October 2005.

* 'An Overview of Database Research at Duke,' presentation at inDuke Meeting, Duke University, May 2005.

Journal Publications:

Book(s) of other one-time publications(s):

Ke Yi, Hai Yu, Jun Yang, Gangqiang Xia, and Yuguo Chen, "Efficient Maintenance of Materialized Top-k Views" , bibl. Bangalore, India, (2003). *Proceedings* Published
of Collection: , "Proceedings of the 19th International Conference on Data Engineering (ICDE '03)"

Junyi Xie, Jun Yang, and Yuguo Chen, "On Joining and Caching Stochastic Streams" , bibl. Baltimore, Maryland, June 2005, (2005). *Proceedings* Published
of Collection: , "Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data"

Adam Silberstein and Jun Yang, "NeXSort: Sorting XML in External Memory" , bibl. Boston, Massachusetts, March 2004, (2004). *Proceedings* Published
of Collection: , "Proceedings of the 20th International Conference on Data Engineering (ICDE '04)"

Hao He and Jun Yang, "Multiresolution Indexing of XML for Frequent Queries" , bibl. Boston, Massachusetts, March 2004, (2004). *Proceedings* Published
of Collection: , "Proceedings of the 20th International Conference on Data Engineering (ICDE '04)"

Ke Yi, Hao He, Ioana Stanoi, and Jun Yang, "Incremental Maintenance of XML Structural Indexes" , bibl. Paris, France, June 2004, (2004). *Proceedings* Published
of Collection: , "Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD '04)"

Zhihui Wang, "Multiple-View Maintenance with Semantic Caching" , bibl. Durham, North Carolina, August 2003, (2003). *Thesis* Published
of Collection: , "M.S. Thesis, Duke University"

Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu, "On Scalable Processing of Continuous Joins" , bibl. Durham, North Carolina, December 2004, (2004). *Technical Report* Submitted
of Collection: , "Technical Report, Department of Computer Science, Duke University"

Adam Silberstein, Hao He, Ke Yi, and Jun Yang, "BOXes: Efficient Maintenance of Order-Based Labeling for Dynamic XML Data" , bibl. Tokyo, Japan, April 2005, (2005). *Proceedings* Published
of Collection: , "Proceedings of the 21st International Conference on Data Engineering"

Hao He, Junyi Xie, Jun Yang, and Hai Yu, "Asymmetric Batch Incremental View Maintenance" , bibl. Tokyo, Japan, April 2005, (2005). *Proceedings* Published
of Collection: , "Proceedings of the 21st International Conference on Data Engineering"

Kamesh Munagala, Jun Yang, and Hai Yu, "Online View Maintenance Under a Response-Time Constraint" , bibl. Mallorca, Spain, October 2005, (2005). *Proceedings* Published
of Collection: , "Proceedings of the 13th Annual European Symposium on Algorithms (ESA '05)"

Hao He, Haixun Wang, Jun Yang, and Philip S. Yu, "Compact Reachability Labeling for Graph-Structured Data" , bibl. Bremen, Germany, November 2005, (2005). *Proceedings* Published
of Collection: , "Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)"

Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu, "Monitoring Continuous Band-Join Queries over Dynamic Data" , bibl. Sanya, Hainan, China, December 2005, (2005). *Proceedings* Published
of Collection: , "Proceedings of the 16th Annual International Symposium on Algorithms and Computation (ISAAC '05)"

Adam Silberstein, Rebecca Braynard, and Jun Yang, "Energy-Efficient Continuous Isoline Queries in Sensor Networks (Poster Paper)" , bibl. Atlanta, Georgia, USA, April 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)"

Haixun Wang, Hao He, Jun Yang, Philip S. Yu, and Jeffrey Xu Yu, "Dual Labeling: Answering Graph Reachability Queries in Constant Time" , bibl. Atlanta, Georgia, USA, April 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)"

Adam Silberstein, Rebecca Braynard, Carla Ellis, Kamesh Munagala, and Jun Yang, "A Sampling-Based Approach to Optimizing Top-k Queries in Sensor Networks" , bibl. Atlanta, Georgia, USA, April 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)"

Badrish Chandramouli, Jun Yang, and Amin Vahdat, "Distributed Network Querying with Bounded Approximate Caching" , bibl. Singapore, April 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA '06)"

Adam Silberstein, Kamesh Munagala, and Jun Yang, "Energy-Efficient Monitoring of Extreme Values in Sensor Networks" , bibl. Chicago, Illinois, USA, June 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)"

Adam Silberstein, Rebecca Braynard, and Jun Yang, "Constraint-Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks" , bibl. Chicago, Illinois, USA, June 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)"

Badrish Chandramouli, Junyi Xie, and Jun Yang, "On the Database/Network Interface in Large-Scale Publish/Subscribe Systems" , bibl. Chicago, Illinois, USA, June 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)"

Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu, "Scalable Continuous Query Processing by Tracking Hotspots" , bibl. Seoul, Korea, September 2006, (2006). *Proceedings* Published
of Collection: , "Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06)"

Adam Silberstein, Rebecca Braynard, Gregory Filpus, Gavino Puggioni, Alan Gelfand, Kamesh Munagala, and Jun Yang, "Data-Driven Processing in Sensor Networks" , bibl. Asilomar, California, USA, January 2007, (2007). *Proceedings* Published
of Collection: , "Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR '07)"

Adam Silberstein and Jun Yang, "Multiple Aggregation for In-Network Control of Sensors" , bibl. Istanbul, Turkey, April 2007, (2007). *Proceedings* Published
of Collection: , "Proceedings of the 23rd International Conference on Data Engineering (ICDE '07)"

Badrish Chandramouli, Christopher N. Bond, Shivnath Babu, and Jun Yang, "On Suspending and Resuming Dataflows" , bibl. Istanbul, Turkey, April 2007. Poster paper. Results in this paper are subsumed by those in the SIGMOD '07 paper titled "Query Suspend and Resume.", (2007). *Proceedings* Published
of Collection: , "Proceedings of the 23rd International Conference on Data Engineering (ICDE '07)"

Hao He, Haixun Wang, Jun Yang, and Philip S. Yu, "BLINKS: Ranked Keyword Searches on Graphs" , bibl. Beijing, China, June 2007, (2007). *Proceedings* Published
of Collection: , "Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD '07)"

Badrish Chandramouli, Christopher N. Bond, Shivnath Babu, and Jun Yang, "Query Suspend and Resume" , bibl. Beijing, China, June 2007, (2007). *Proceedings* Published
of Collection: , "Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD '07)"

Adam Silberstein, Gavino Puggioni, Alan Gelfand, Kamesh Munagala, and Jun Yang, "Suppression and Failures in Sensor Networks: A Bayesian Approach" , bibl. Vienna, Austria, September 2007, (2007). *Proceedings* Published
of Collection: , "Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB '07)"

Badrish Chandramouli, Jeff M. Phillips, and Jun Yang, "Value-Based Notification Conditions in Large-Scale Publish/Subscribe Systems" , bibl. Vienna, Austria, September 2007, (2007). *Proceedings* Published
of Collection: Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB '07), ""

Junyi Xie, Jun Yang, Yuguo Chen, Haixun Wang, and Philip S. Yu, "A Sampling-Based Approach to Information Recovery" , bibl. Cancun, Mexico, April 2008, (2008). *Proceedings* Published
of Collection: , "Proceedings of the 24th International Conference on Data Engineering (ICDE '08)"

Fei Chen, AnHai Doan, Jun Yang, and Raghu Ramakrishnan, "Efficient Information Extraction over Evolving Text Data" , bibl. Cancun, Mexico, April 2008, (2008). *Proceedings* Published
of Collection: , "Proceedings of the 24th International Conference on Data Engineering (ICDE '08)"

Badrish Chandramouli, Jun Yang, Pankaj K. Agarwal, Albert Yu, and Ying Zheng, "ProSem: Scalable Wide-Area Publish/Subscribe" , bibl. Vancouver, Canada, June 2008; system demonstration description, (2008). *Proceedings* Published
of Collection: , "Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)"

Badrish Chandramouli and Jun Yang, "End-to-End Support for Joins in Large-Scale Publish/Subscribe Systems" , bibl. Auckland, New Zealand, August 2008, (2008). *Proceedings* Published
of Collection: , "Proceedings of the 34th International Conference on Very Large Data Bases (VLDB '08)"

Other Specific Products:

Software (or netware)

ProSem is a scalable wide-area publish/subscribe system that supports complex, stateful subscriptions as well as simple ones. One unique feature of ProSem is its cost-based joint optimization of both subscription processing and notification dissemination. ProSem uses novel reformulation techniques to expose new alternatives for processing and disseminating data using standard stateless content-driven network components.

ProSem was demonstrated at SIGMOD 2008. Its development is still ongoing. When ready, it will be released open-source.

Software (or netware)

ERS (Enhanced Repository System) v1 captures the metadata on lineage, dependency and versioning for computational biology workflows. It allows users to visualize, explore, and query the graph-structured data, and provides a subscription service that notifies the users whenever an update could potentially affect a derived dataset of interest.

This system is currently being used by researchers in computational immunology at Duke University. Its development is still ongoing. Once ready, it will be released as open-source.

Internet Dissemination:

http://www.cs.duke.edu/dbgroup/ddm/

Contributions:

Contributions within Discipline:

We have made contributions to multiple application domains of derived data maintenance, including view maintenance, data warehousing, indexing and querying XML and graph-structured data, continuous query processing in stream and publish/subscribe systems, sensor networks, etc. Many contributions have been published in premier conferences (7 full papers in SIGMOD 2004-2007, one demo paper in SIGMOD 2008, 4 full papers in VLDB 2006-2008, 9 full papers in ICDE 2003-2008, one paper each in CIKM 2005, ESA 2005, ISAAC 2005, and DASFAA 2006, and CIDR 2007). For detailed descriptions of these contributions please refer to the section of this report on research and education activities.

In addition to serving on numerous program committees, I served as program committee co-chair of DMSN (International Workshop on Data Management for Sensor Networks) 2007, and general co-chair of DMSN 2008. I have also been active in running the Carolina Database Research Group, and was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

Contributions to Other Disciplines:

I have been actively applying derived data techniques to areas beyond computer science. Specifically, I have been working with a group of computational immunologists led by Dr. Thomas B. Kepler at Duke University on developing a system for tracking lineage, dependency, and versioning of derived datasets in computational biology workflows. The product is a system called ERS (Enhanced Repository Service) that is now being used by these immunologists.

I have also been collaborating with a group of ecologists led by Dr. James S. Clark at the Duke University School of Environment on developing a wireless sensor network in Duke Forest to study how various environmental variables influence forest growth.

Contributions to Education and Human Resources:

I have advised the following students in the context of this project:

Ph.D. students:

Adam Silberstein (defended in February 2007, now at Yahoo! Research);
Hao He (defended in July 2007, now at Google);
Junyi Xie (defended in September 2007, now at Oracle);
Badrish Chandramouli (defended in July 2008, now at Microsoft Research).

M.S. students:

Zhihui Wang (thesis completed in Summer 2003);
Wenbin Pan (thesis completed in Fall 2004);
Pradeep K. Gunda (thesis completed in Fall 2007).

Undergraduate students:

Christopher N. Bond (BS with High Distinction, 2005);
Gregory Filpus;
Congyi Wu;
Tyler J. Brock (BA with Distinction, 2007).

Contributions to Resources for Science and Technology:

I have been actively applying derived data techniques to areas beyond computer science. Specifically, I have been working with a group of computational immunologists led by Dr. Thomas B. Kepler at Duke University on developing a system for tracking lineage, dependency, and versioning of derived datasets in computational biology workflows. The product is a system called ERS (Enhanced Repository Service) that is now being used by these immunologists.

I have also been collaborating with a group of ecologists led by Dr. James S. Clark at the Duke University School of Environment on developing a wireless sensor network in Duke Forest to study how various environmental variables influence forest growth.

Categories for which nothing is reported:
Participants: Partner organizations
Products: Journal Publications
Contributions Beyond Science and Engineering

Submit Return

We welcome comments on this system