ANNUAL 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. Agarawal, 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 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.

Activities and findings:

Research and Education Activities:

The focus of this CAREER project is on 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.

In Year 4 of this project, we have investigated the following specific research problems:

1. Query suspend and resume. Back in Year 2 of this project, an undergraduate student, Christopher N. Bond, worked with me on the interesting problem of suspending and resuming a database query. Query suspend and resume are useful in many applications. For example, suppose a long-running analytical query is executing on a database server and has been allocated a large amount of physical memory. A high-priority task comes in and we need to run it immediately with all available resources. We have several choices. We could swap out the old query to disk, but writing out a large execution state may take too much time. Another option is to terminate the old query and restart it after the new task completes, but we would waste all the work already performed by the old query. Yet another alternative is to periodically checkpoint the query during execution, but traditional synchronous checkpointing carries high overhead. In Year 4, we resume the investigation of this problem based on the insights and preliminary ideas we developed in Year 2. We advocate a database-centric approach to implementing query suspension and resumption, with negligible execution overhead, bounded suspension cost, and efficient resumption. The basic idea is to let each physical query operator perform lightweight checkpointing according to its own semantics, and coordinate asynchronous checkpoints among operators through a novel contracting mechanism. At the time of suspension, we find an optimized suspend plan for the query, which may involve a combination of dumping current state to disk and going back to previous checkpoints. Indeed, the choice between dumping state and reconstructing state from the previous checkpoint is a classic tradeoff involving derived data: to recompute or to materialize. The optimization seeks to minimize the suspend/resume overhead while observing the constraint on suspension time. Our approach requires

only small changes to the iterator interface, which we have implemented in the PREDATOR database system. Experiments with our implementation demonstrate significant advantages of our approach over traditional alternatives. The results have been published in SIGMOD 2007.

2. Derived data for queries over graph-structured data. Graph-structured data has found a growing number of important applications recently. In bioinformatics, protein interactions, metabolic pathways, and gene regulatory networks are modeled as directed graphs. In Semantic Web, two key technologies, RDF and OWL, are designed to capture graph data. In earlier years of this project, we have focused on efficiently supporting fundamental query primitives such as testing reachability between two nodes in a graph. In Year 4, we have begun to tackle complex user-level queries. In particular, a top-k keyword search query on a graph finds the top k answers according to some ranking criteria, where each answer is a substructure of the graph containing all query keywords. Current techniques for supporting such queries on general graphs suffer from several drawbacks, e.g., poor worst-case performance, not taking full advantage of indexes, and high memory requirements. To address these problems, we propose BLINKS, a bi-level indexing and query processing scheme for top-k keyword search on graphs. BLINKS follows a search strategy with provable performance bounds, while additionally exploiting a bi-level index for pruning and accelerating the search. To reduce the index space, BLINKS partitions a data graph into blocks: The bi-level index stores summary information at the block level to initiate and guide search among blocks, and more detailed information for each block to accelerate search within blocks. Our experiments show that BLINKS offers orders-of-magnitude performance improvement over existing approaches. The results have been published in SIGMOD 2007.

3. 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. Most existing work on indexing continuous relational queries has focused on selections. As far as we know, there has been little work on how to process more complex continuous queries (e.g., joins) scalably. 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 particular, we have developed novel, 'input-sensitive' schemes for indexing continuous joins with range conditions. These schemes exploits the clusteredness of the range conditions being indexed: More clustered queries lead to more efficient processing. We have also

obtained other results including lower bounds on the inherent complexity of the problem, and data structures with space-time tradeoffs. In Year 3, we have focused on developing more theoretical results, published in ISAAC 2005. In Year 4, we worked on more practical extensions, implementation, and experimental validation. We developed the concept of hotspots, which are heavily clustered groups of query ranges; the hotspots are processed differently from the rest of the query ranges. This two-pronged approach allows us to exploit clusteredness for processing where most beneficial, and it is much more robust than our previous approach in cases where scattered ranges coexist with clustered ones. These results have been published in VLDB 2006. In Year 5, we plan to make processing more dynamic: For each incoming data update, we will decide how to process it based on the estimated costs of different processing strategies.

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. So far, we have been focusing on system implementation. A prototype system is under beta testing by our collaborators. We hope to work on aspects of the system that are more novel from a computer science perspective in Year 5.

5. Derived data in wireless sensor networks. In collaboration with Duke University School of Environment, we are building a wireless sensor network in Duke Forest to study how various environmental variables influence forest growth. 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, we have been actively studying the use of derived data in sensor data processing through the support of this CAREER grant, and have obtained a series of results. (1) In our work published in ICDE 2007, we tackled the problem of supporting many-to-many aggregation in a sensor network. An application of many-to-many aggregation is in-network control of sensors. For expensive sensing tasks such as sap flux measurements and camera repositioning, we use low-cost information obtained at multiple other nodes in the network to control such tasks, e.g., decreasing sampling rates when readings are predictable or unimportant, while increasing sampling rates when there are interesting activities. In general, there is a many-to-many relationship between sources (nodes providing control inputs) and destinations (nodes requiring control outputs). We present a method for implementing many-to-many aggregation in a sensor network that minimizes the communication cost by optimally balancing a combination of multicast and in-network aggregation. Our optimization technique is efficient in finding the initial solution and handling dynamic updates. (2) The work we published in CIDR 2007 was jointly supported by DDDAS, CAREER, and in particular, a REU supplement to CAREER. This work outlines our vision for building a battery-powered wireless sensor network for environmental monitoring. The straightforward solution of instructing all nodes to report their measurements as they are taken to a base station will quickly consume the network's energy. On the other hand, the solution of building models for node behavior and substituting these in place of the actual measurements is in conflict with the end goal of learning environmental models. To address this dilemma, we

propose data-driven processing, the goal of which is to provide continuous data without continuous reporting, but with checks against the actual data. Our primary strategy for this is suppression, which uses in-network monitoring to limit the amount of communication to the base station. Suppression employs models for optimization of data collection, but not at the risk of correctness. We discuss techniques for designing data-driven collection, such as building suppression schemes and incorporating models into them. We then present and address some of the major challenges to making this approach practical, such as handling failure and avoiding the need to co-design the network application and communication layers. This work includes contribution from our undergraduate research intern, Gregory Filpus, who was supported by the REU supplement.

The progress we have made in Year 4, as summarized above, are in line with the modified research plan outlined in the project report from Years 2 and 3. We have broaden our study of derived data maintenance to continuous queries in a networked setting, including wireless sensor networks. In parallel, we continue to work on derived data in the forms of views and indexes for traditional databases and graph-structured data. In Year 5, we plan to (1) continue our work on scalable subscription processing and notification in the context of a wide-area publish/subscribe system with a rich subscription language; (2) continue our work on querying graph-structured data; (3) continue to work with our bioinformatics collaborators to apply our techniques (for both subscription processing and graph indexing) to manage lineage, dependency, and versioning of derived datasets that arise in computational biology workflows; (4) continue to work with with our collaborators in Duke University School of Environment to apply derived data techniques to wireless sensor networks. We are currently seeking additional funding to complement our work in the area of wide-area publish/subscribe systems. A proposal has been submitted to NSF IIS in 2006.

In terms of educational activities, I have continued to incorporate current research topics into both undergraduate and graduate database course at Duke University. The undergraduate database course I offered in Fall 2006 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.

Findings:

We have made significant progress in studying the derived data maintenance problem in multiple application domains, including view maintenance, data warehousing, stream data processing, indexing and querying XML and graph-structured data, continuous query processing in wide-area networks and sensor networks. Published results from this grant so far 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).

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).

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).

We are actively working with our collaborators in bioinformatics and ecology and applying the above results to real-world problems. Based on these findings we believe that the direction we are currently pursuing is a promising one. For detailed descriptions of these findings please refer to the section of this report on research and education activities.

Training and Development:

The PI has advised the following students in the context of this project:

Ph.D. students: Adam Silberstein (defended in February 2007), Hao He, Junyi Xie, Badrish Chandramouli.

M.S. student: Zhihui Wang (thesis completed in 2003), Wenbin Pan (thesis completed in 2004).

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

2007).

Outreach Activities:

The PI has 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. The PI was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

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)"

Other Specific Products:

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, stream data processing, indexing and querying XML and graph-structured data, continuous query processing in wide-area networks and sensor networks. A number of the contributions have been published in premier conferences (7 full papers in SIGMOD 2004-2007, 7 full papers in ICDE 2003-2007, one paper each in CIKM 2005, ESA 2005, ISAAC 2005, DASFAA 2006, and VLDB 2006). 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, the PI has 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:

The PI has been actively applying derived data techniques to areas beyond computer science. Specifically, the PI has been working with a group of computational immunologists led by Dr. Thomas B. Kepler at Duke University on developing a system called ERS for tracking lineage, dependency, and versioning of derived datasets in computational biology workflows. Also, the PI has 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 Resources for Science and Technology:

The PI has 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. The PI was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

Special Requirements for Annual Project Report:

Unobligated funds: less than 20 percent of current funds

Categories for which nothing is reported:

Participants: Partner organizations
Products: Journal Publications
Products: Other Specific Product
Contributions to Education and Human Resources
Contributions Beyond Science and Engineering
Special Reporting Requirements
Animal, Human Subjects, Biohazards

Submit Return
View Attached PDF File

We welcome comments on this system