Title:
Efficient View-Design Algorithms to Achieve Near-Optimal Performance of Sets of Relational Queries

Project Participants

Senior Personnel

Name: Chirkova, Rada
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Gupta, Shalu
Worked for more than 160 Hours: Yes
Contribution to Project:
Shalu Gupta was the PI's M.Sc. student at NC State University in 2003-05; she collaborated with the PI by working on a PostgreSQL-based (http://www.postgresql.org) implementation for the project. Shalu has also contributed to a paper accepted in 2005 to a conference (10th International Conference on Database Systems for Advanced Applications DASFAA-2005). We are currently working on a journal version of the paper.

Name: Asgharzadeh Talebi, Zohreh
Worked for more than 160 Hours: Yes
Contribution to Project:
Zohreh Asgharzadeh Talebi is the PI's doctoral student at NC State University; she has contributed to a paper accepted in 2005 to a conference (Ninth East-European Conference on Advances in Databases and Information Systems ADBIS 2005), and to a journal version of the paper, submitted in January 2006 to IEEE Transactions on Knowledge and Data Engineering (TKDE). After successfully defending her M.Sc. thesis in May 2006 (thesis title 'Exact and heuristic methods for solving the view-selection problem for aggregate queries'), Zohreh has been working on her Ph.D.-thesis project, focusing on globally optimal design of indexes alongside views for aggregate queries.

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts
1. Professor Yahya Fathi in the Operations Research program at NC State University; he has collaborated with the PI on a journal version of the paper accepted in 2005 to the 9th East-European Conference on Advances in Databases and Information Systems, ADBIS-2005. We submitted the journal version to the IEEE Transactions on Knowledge and Data Engineering (TKDE) in January 2006; the manuscript is available by
request. Dr. Fathi has also been collaborating with the PI on Zohreh Asgharzadeh Talebi's Ph.D.-thesis project 'Exact and Heuristic Methods for Solving the Index Selection Problem for Aggregate Queries'.

2. Professor Matthias Stallmann in the Computer Science department at NC State University; he has been collaborating with the PI on Zohreh Asgharzadeh Talebi's Ph.D.-thesis project 'Exact and Heuristic Methods for Solving the Index Selection Problem for Aggregate Queries'.

3. Michael W. Martin, Ph.D. student in the NCSU Computer Science Department. The PI has been advising Michael W. Martin on his Ph.D. thesis 'Efficient index structure for data warehousing,' with graduation plans for summer 2006. This project has resulted in a publication in the proceedings of the International Conference of the Information Resources Management Association (IRMA), May 2006.

4. Andrew Frick, graduate student in the NCSU Computer Science Department. The PI has been advising Andrew Frick on his M.Sc. project 'Design of views and indexes in data warehouses,' with graduation plans for summer 2006.

5. Professor Michael R. Genesereth at Stanford University; he coauthored with the PI a paper accepted in 2005 to a Logic in Computer Science, LICS-2005 affiliated workshop (Logic and Computational Complexity Workshop, LCC-2005, see the PI's last year's report). We are preparing for publication an extended journal version of that paper; the current version of the manuscript is available by request.

6. Professor Foto Afrati at the National Technical University of Athens, Greece; she has published with the PI three conference papers within the scope of this project and has been collaborating with the PI on two journal submissions for this project. The published conference papers are publications in (a) the 10th International Conference on Database Systems for Advanced Applications (DASFAA-2005), in (b) the Tenth International Conference on Database Theory, ICDT-2005, and in (c) Symposium on Abstraction, Reformulation, and Approximation, SARA-2005. One of the journal versions (extension of our ICDT-2005 publication) was submitted to the Journal of Computer and System Sciences in May 2006, and the other (extension of our SARA-2005 publication) is in the final review stage by the authors, to be submitted to the Acta Informatica journal in June 2006. Both manuscripts are available by request.

7. Professor Manolis Gergatsoulis at the Ionian University of Corfu, Greece. He has coauthored with the PI the publication in the Symposium on Abstraction, Reformulation, and Approximation (SARA-2005), and has been been collaborating with the PI on a journal submission of an extended version of that publication. For the details of the joint project, see under Professor Foto Afrati.

8. Vassia Pavlaki, doctoral student at the National Technical University of Athens, Greece. She has coauthored with the PI the publication in the Symposium on Abstraction, Reformulation, and Approximation (SARA-2005), and has been been collaborating with the PI on a journal submission of an extended version of that publication. For the details of the joint project, see under Professor Foto Afrati.


10. Kyoung-hwa Kim, the PI's M.Sc. student at NC State University in 2003-05; she collaborated with the PI by working on a PostgreSQL-based (http://www.postgresql.org) implementation for the project. Kyoung-hwa has also contributed to a paper accepted to a conference (International Advanced Database Conference IADC-2005).

11. Professor Chen Li at the University of California, Irvine; he coauthored with the PI a paper accepted to The VLDB Journal in 2005.

12. Jia Li, doctoral student at the University of California, Irvine; she contributed to a paper accepted to The VLDB Journal in 2005.

13. Jingni Li, M.Sc. student at NC State University; she contributed to a paper accepted in 2005 to a conference (Ninth East-European Conference on Advances in Databases and Information Systems ADBIS-2005).

14. Charles Loftis was a graduate student at NC State University in 2003-05; he contributed to a paper accepted in 2005 to a conference (10th International Conference on Database Systems for Advanced Applications DASFAA-2005).

15. Filip Perich, Cougar Software, Inc; Filip has a Ph.D. from the University of Maryland Baltimore County (UMBC). While he was a student at UMBC, he collaborated with the PI on a book chapter "Data Management for Mobile Ad-Hoc Networks' in "Enabling Technologies for Wireless e-Business Applications' (edited by W. Kou and Y. Yesha, Springer, July 2005).
Activities and Findings

Research and Education Activities:

Activities:

[Note: NCSU stands for "North Carolina State University."]

A1. The PI worked in 2003-05 with her NCSU M.Sc. students Shalu Gupta (female) and Kyoung-hwa Kim (female) on a PostgreSQL-based (http://www.postgresql.org) implementation for the project. The latest version of the code (as of 04/01/2005) and an accompanying technical report are available for the public at http://research.csc.ncsu.edu/selftune/.

The goals of the project were to build a framework for generation and rewriting of derived data in an open-source database-management system PostgreSQL (http://www.postgresql.org), and to use the framework for experimental validation of research projects in my group, including the NSF project 0307072.

As part of the coding project, the PI and Kyoung-hwa Kim studied the precision/efficiency tradeoff in view-size estimation in self-organizing databases. An outcome of the work is their joint paper (titled "View-size estimation in self-organizing databases") accepted to the International Advanced Database Conference (IADC), San Diego, CA, June 2005.

A2. The PI has been working with Dr. Foto Afrati (NTUA Greece) on three research projects within the scope of this NSF project.

In the first joint project (2003-05) they studied how views with or without aggregation can be (i) used in answering queries with aggregation, or (ii) designed to efficiently evaluate queries with aggregation. This project resulted in a publication in the Tenth International Conference on Database Theory (ICDT) in January 2005. In May 2006, Dr. Afrati and the PI submitted a journal extension of this publication to the Journal of Computer and System Sciences (JCSS); the manuscript is available by request.

The second joint project (Dr. Afrati and the PI, 2003-04) concerned efficient design and use of views with aggregation in star-schema data warehouses. An outcome of this project was a publication in April 2005, jointly with NCSU graduate students Shalu Gupta and Charles Loftis, in the Tenth International Conference on Database Systems for Advanced Applications (DASFAA-2005). Shalu Gupta, and the PI are currently preparing a journal version of the paper.

The goal of the third joint project (Dr. Afrati and the PI) is to study tractable cases of view design (in the absence of dependencies), which include (a) set semantics for queries without self-joins, (b) bag semantics, and (c) bag-set semantics. This is a joint project with Dr. Manolis Gergatsoulis (Ionian University of Corfu, Greece) and with Dr. Afrati's doctoral student Vassia Pavlaki (NTUA Greece). This project resulted in a publication, in July 2005, in the Symposium on Abstraction, Reformulation and Approximation (SARA-2005). In 2005, the PI and her collaborators extended the findings to efficient algorithms for view design under bag and bag-set semantics for queries without aggregation. We are currently working on a journal version of the SARA-2005 publication for submission to the Acta Informatica journal; this version will include our new results. The current version of the manuscript is available by request.

A3. The PI has been working with Dr. Yahya Fathi and doctoral student Zohreh Asgharzadeh Talebi (female), both in NCSU Operations Research, on developing linear-programming and integer-programming models for designing views and indexes for queries with aggregation. The importance of this work is that its results would provide globally optimal solutions to the view- and index-selection problems. One outcome of this project so far has been a publication in 2005 (jointly with Jingni Li, female, M.Sc. student in NCSU Operations Research) in the 9th East-European Conference on Advances in Databases and Information Systems (ADBIS-2005). Dr. Fathi, Zohreh, and the PI submitted in January 2006 a journal extension of this publication to IEEE Transactions on Knowledge and Data Engineering (TKDE); the manuscript is available by request.

After successfully defending her M.Sc. thesis in May 2006 (thesis title 'Exact and heuristic methods for solving the view-selection problem for aggregate queries'), doctoral student Zohreh Asgharzadeh Talebi is currently working with the PI and with Drs. Fathi and Stallmann (in the NCSU Computer Science department) on her Ph.D.-thesis project, focusing on globally optimal design of indexes alongside views for aggregate queries. We are currently investigating proofs of NP-hardness of the view- and index-selection problems.

A4. The PI has worked with Dr. Michael R. Genesereth (the PI's Ph.D. advisor, Stanford University) on view design under set semantics and integrity constraints. The findings were accepted in June 2005 to the Logic and Computational Complexity Workshop (LCC-2005, in conjunction with LICs-2005). In 2005, Dr. Genesereth and the PI extended these results to the contexts of (a) bag and bag-set semantics and (b)
aggregate queries. We are currently working on a journal version of the LCC-2005 publication for submission to the Journal of Computer and System Sciences (JCSS); this version will include all the new results. The current version of the manuscript is available by request.

A5. In 2006, the PI published a journal paper (in The VLDB Journal) with Dr. Chen Li and doctoral student Jia Li, both at the University of California, Irvine. This work is an extension of their previous results published in the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS-2003).


A7. The PI has been advising doctoral student Michael W. Martin (NCSU Computer Science) on his Ph.D. thesis project 'Efficient index structure for data warehousing,' with graduation plans for summer 2006. In 2006, this project resulted in a publication in the proceedings of the International Conference of the Information Resources Management Association (IRMA), May 2006.

A8. The PI has been advising graduate student Andrew Frick (NCSU Computer Science) on his M.Sc. thesis project 'Design of views and indexes in data warehouses,' with graduation plans for summer 2006. We are preparing a publication on the results of the research, possibly for submission to the ACM Ninth International Workshop on Data Warehousing and OLAP (DOLAP-2006).

B. In the summer semester 2005, I worked with an NCSU undergraduate student Nathaniel Derbinsky on the project (jointly with NCSU M.Sc. student Andrew Frick) on designing indexes for query-evaluation efficiency, see under A8 above.

C. In the spring semester 2005, I taught two courses, NCSU CSC (computer science) 440 (intro database systems for undergraduate students) and 742 (advanced topics on database systems for graduate students). Parts of the courses were devoted (especially in CSC 742) to discussing the research directions of this NSF proposal. To increase novelty and excitement in the curriculum and to give students a research experience that could motivate them to go to graduate school and to academia after they graduate, I was acting as an expert-in-residence on research in the project.

This project was continued in the fall 2005 in the course NCSU CSC 540 (intro database systems for graduate students). I am planning to continue with this education project in spring 2007, when I will teach courses CSC 440 and CSC 540 again.

Findings:
Findings on the projects A1 through A8, under Activities above (projects B and C are of educational value and thus do not involve research findings):

F1. We are using the PostgreSQL-based (http://www.postgresql.org) implementation to validate the findings in the project, including those for the individual projects A2, A3, A8, B, and C. In addition, we have used the implementation to discover and study guidelines on using the precision/efficiency tradeoff in view-size estimation in self-organizing databases. An outcome of the work is a paper 'View-size estimation in self-organizing databases' accepted to the International Advanced Database Conference (IADC), San Diego, CA, June 2005.

The latest version of the code (as of 04/01/2005) and an accompanying technical report are available for the public at http://research.csc.ncsu.edu/selftune/.

F2. In the context of obtaining equivalent rewritings of aggregate queries using views, we have considered the problem of minimizing the cost of computing a query workload, and looked at query rewriting using existing views and at view selection. In the query-rewriting problem, we have discovered sufficient and necessary conditions for a rewriting to exist. For view selection, we have proved complexity results. We have also developed algorithms for obtaining rewritings and selecting views. See publication 'Selecting and Using Views to Compute Aggregate Queries' in the Tenth International Conference on Database Theory (ICDT) in January 2005, and our submission (in May 2006) of an extended version of this publication to the Journal of Computer and System Sciences (JCSS); the journal-submission manuscript is available by request.

In the context of efficient design and use of views with aggregation in star-schema data warehouses, we have developed an extensible system architecture for Query-Performance Enhancement by Tuning (QPET). QPET combines design and use of derived data in an end-to-end approach to automated query-performance tuning, and selects appropriate data-design algorithms depending on the characteristics of the prevalent queries. Our focus in automated query-performance tuning is on a tradeoff between the amount of system resources spent on
designing derived data and the degree of the resulting improvement in query performance. We have developed algorithms and reported experimental results in designing and using materialized views for practically important classes of aggregate queries, including range-aggregate queries on star-schema data warehouses. See publication 'Designing and Using Views to Improve Performance of Aggregate Queries' in the Tenth International Conference on Database Systems for Advanced Applications (DASFAA) in April 2005.

In the project on studying view design for efficiently answering queries without aggregation, we have obtained complexity results on selecting views to answer queries in relational databases under set, bag, and bag-set semantics. The results can be used under each of the three assumptions, to find sound and complete algorithms for designing views and rewriting queries efficiently. See publication 'Designing Views to Efficiently Answer Real SQL Queries' in the Symposium on Abstraction, Reformulation and Approximation (SARA) in July 2005, and our planned submission (June 2006) of an extended version of this publication to the Acta Informatica journal; this version will include our new results. The current version of the manuscript is available by request.

F3. In the project on obtaining optimal solutions for view design, we have developed an integer-programming model to obtain optimal solutions to the problem of view selection for aggregate queries on data warehouses. We have also performed post-optimality analysis to determine/observe the impact of changing certain input characteristics on the optimal solution. The importance of this work is that its results would provide globally optimal solutions to the view-selection problem. See publication 'A Formal Model for the Problem of View Selection for Aggregate Queries' in the Ninth East-European Conference on Advances in Databases and Information Systems (ADBIS), in September 2005. Also see our submission (in January 2006) of an extended version of this publication to IEEE Transactions on Knowledge and Data Engineering (TKDE); the manuscript is available by request. Finally, based on the findings in this project, doctoral student Zohreh Asgharzadeh Talebi defended in May 2006 her M.Sc. thesis in May 2006; the thesis title is 'Exact and heuristic methods for solving the view-selection problem for aggregate queries'.

Zohreh Asgharzadeh Talebi's Ph.D.-thesis project (globally optimal design of indexes alongside views for aggregate queries) is work in progress.

F4. In the project on view design in presence of integrity constraints, we have obtained complexity results and algorithms for database reformulation for conjunctive queries and for several types of constraints, including functional and inclusion dependencies. To obtain better complexity results, we have introduced an unchase technique, which reduces the problem of query equivalence under constraints to equivalence in the absence of constraints without increasing query size. See publication 'Database Reformulation with Integrity Constraints' in the Logic and Computational Complexity Workshop (LCC, in conjunction with the Logic in Computer Science Conference - LICS), in June 2005.

In 2005, Dr. Genesereth and the PI extended these results to the contexts of (a) bag and bag-set semantics and (b) aggregate queries, with aggregation MAX, MIN, SUM, COUNT. We gave a complete characterization of which types of dependencies can be used in chase (or unchase) of conjunctive queries under each of bag and bag-set semantics for query evaluation. We also studied implications of these results for aggregate queries. We are currently working on a journal version of the LCC-2005 publication for submission to the Journal of Computer and System Sciences (JCSS); this version will include all the new results. The current version of the manuscript is available by request.

F5. The PI's 2006 journal publication (jointly with Dr. Chen Li and doctoral student Jia Li, both at the University of California, Irvine) in The VLDB Journal studies the following problem. Given a database and a set of queries, we want to find a set of views that can compute the answers to the queries, such that the amount of space, in bytes, required to store the viewset is minimum on the given database. (We also handle problem instances where the input has a set of database instances, as described by an oracle that returns the sizes of view relations for given view definitions.) This problem is important for applications such as distributed databases, data warehousing, and data integration. Our results in this publication concern the decidability and complexity of the problem for workloads of conjunctive queries. We show that results differ significantly depending on whether the workload queries have self-joins. Further, for queries without self-joins we describe a very compact search space of views, which contains all views in at least one optimal viewset. We present techniques for finding a minimum-size viewset for a single query without self-joins by using the shape of the query and its constraints, and validate the approach by extensive experiments.

F6. In the PI's 2005 book-chapter publication (with Filip Perich and Dr. Anupam Joshi at the University of Maryland Baltimore County) 'Data Management for Mobile Ad-Hoc Networks' (in 'Enabling Technologies for Wireless e-Business Applications', Springer), we presented an overview of challenges arising in the area of mobile data management and surveyed existing solutions, with the emphasis on data management in mobile ad-hoc networks. We concentrated on specific data-management challenges in three areas, and presented the MoGATU model - a novel peer-to-peer data-management model for mobile ad-hoc networks. Finally, we outlined future work required to bring data management and processing in mobile ad-hoc networks close to their goal to allow individual devices to compute the information they need in a timely manner.

F7. The PI's 2006 publication (jointly with doctoral student Michael W. Martin) in the proceedings of the International Conference of the Information Resources Management Association (IRMA) introduced MultiDimensional Dynamic Clustering (MDDH), a new multidimensional
data-clustering method that supports efficient processing of OLAP queries while requiring fewer materialized views, aggregates, and secondary indexes. MDDC also allows dynamic inserts, updates, and deletes on the stored data without reorganization of the underlying data structures.

F8: Work in progress

Training and Development:
Research and teaching skills and experience the project helped provide:

I have been working with four female students at NCSU in the context of the project: M.Sc. students Shalu Gupta and Kyoung-hwa Kim (see Activity A1), and M.Sc. student Jingni Li and Ph.D. student Zohreh Asgharzadeh Talebi (see Activity A3). The project has contributed to education of female NCSU students and of other NCSU students (including doctoral student Michael W. Martin, M.Sc. student Andrew Frick, graduate student Charles Loftis, and undergraduate student Nathaniel Derbinsky, see Activities A7, A8, and B), by giving them unique research and implementation skills in order to give them a competitive edge in the workforce.

One outcome of this project so far are M.Sc. theses for NCSU CSC student Kyoung-hwa Kim (graduated with a M.Sc. degree in January 2005), NCSU CSC student Shalu Gupta (graduated with a M.Sc. degree in July 2005), and NCSU Operations Research student Zohreh Asgharzadeh Talebi. In addition, this grant provides financial support for Zohreh Asgharzadeh Talebi in her Ph.D.-thesis project.

In addition, the project has contributed to increasing novelty and excitement in the curriculum and to give students (in my courses CSC 440/540/742) a research experience that could motivate them to go to graduate school and to academia after they graduate, see Activity C.

Outreach Activities:
I have made all project reports and an implementation version and technical report available to the general public on the Internet, at http://research.csc.ncsu.edu/selftune/

Journal Publications


Books or Other One-time Publications

Web/Internet Site

URL(s):
http://research.csc.ncsu.edu/selftune/

Description:
This site has up-to-date information on the progress of the project, including an overview of the project (http://research.csc.ncsu.edu/selftune/manifestoFall2003.pdf), NSF annual reports (http://research.csc.ncsu.edu/selftune/reportAugust2003.html), technical reports and software to download.

Other Specific Products

Product Type:
Software (or netware)

Product Description:
This software a is a database-management system that includes a framework for generation and rewriting of derived data. The software has been built in an open-source database-management system PostgreSQL (http://www.postgresql.org). The purpose of the software is experimental validation of research projects in my group, including the NSF project 0307072.

Sharing Information:
On the Internet via http://research.csc.ncsu.edu/selftune/

Contributions

Contributions within Discipline:
Findings on the problem of designing views to reduce the evaluation costs of database queries contribute to the body of knowledge on designing and using derived data to answer queries in data-intensive systems. Please see the details under 'Findings' in this report.

Contributions to Other Disciplines:

Contributions to Human Resource Development:
Yes; please see under 'Findings' in this project. Briefly, work on the project is contributing to education of diverse skilled workforce, within the context of the mission of NCSU.

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Special Requirements

Special reporting requirements:
Additional information: no changes in the research plan.

**Change in Objectives or Scope:** None

**Unobligated funds:** $12,120.00

**Animal, Human Subjects, Biohazards:** None

**Categories for which nothing is reported:**

Organizational Partners
Any Book
Contributions: To Any Other Disciplines
Contributions: To Any Resources for Research and Education
Contributions: To Any Beyond Science and Engineering