Svetlozar Evtimov Nestorov

Computer Science Department	**tel:**	(650) 723-3605
Gates Hall 4A, Room 432	**fax:**	(650) 725-2588
Stanford University	**e-mail:**	`evtimov@db.stanford.edu`
Stanford, CA 94305-9040	**URL:**	`http://www-db.stanford.edu/~evtimov`

Research Interests: Databases, knowledge bases, and distributed systems.

I am interested in the knowledge-discovery process for large amounts of structured, semistructured, and multimedia data. In particular, I am interested in extending database technology to address the new challenges of query formulation, optimization, and processing for data mining problems. In this context, I am also interested in information integration, data warehousing, and query optimization.

Education

1994 - PRESENT	*Ph.D.* in Computer Science, *Ph.D. minor* in Engineering-Economic Systems, Stanford University (expected AUG 1999) **Advisor:** Prof. Jeffrey D. Ullman **Dissertation:** *Efficient, On-line, Ad-hoc, Data Mining*
JUN 1997	*M.S.* in Computer Science, Stanford University
1992 - 1994	*B.S.* (double) in Computer Science and Mathematics, Stanford University
1989 - 1992	Mathematics major, Sofia University, Bulgaria

Honors

1993 - 1994	*Frances Morgan Swain Scholarship*, Stanford University
DEC 1992	All US 12th, *Putnam Mathematical Competition*, Stanford University
JUN 1992	1st prize, *National Mathematical Olympiad*, Sofia, Bulgaria
JUN 1991	1st prize, *National Mathematical Olympiad*, Sofia, Bulgaria
MAY 1989	1st prize, *Balkan Mathematical Olympiad*, Split, Croatia

Synopsis of Research Contributions

Efficient, On-line, Ad-hoc, Data Mining: With the advent of the World Wide Web, the amount of data stored and accessible electronically on the Internet and private Intranets has grown tremendously. The process of knowledge discovery (data mining) by sophisticated and intricate analysis of this data has become very important for the business and research communities alike. The current state of the art in data mining is specialized techniques and algorithms, designed for limited types of data, that solve specific problems. My research has primarily focused on the design and development of a general framework and algorithms for formulating, optimizing, and processing complex data mining queries in a uniform manner.

Synopsis of Research Contributions (cont.)

Query Flocks: I have developed a framework where a large class of data mining problems can be expressed declaratively as query flocks. Each query flock has a query part and a filter part. The query part consists of a set of Datalog queries with parameters. The filter part consists of several SQL-style filter conditions over the results of the query part. The semantics of the query flock is defined as the set of parameter values that satisfy the filter conditions. In this formal setting it becomes easy to express ad-hoc data-mining queries and optimize them in a uniform way. For example, I have shown that a general version of the a-priori trick can be applied for a certain class of filter conditions [2,8].

Optimization Techniques for Data-Mining Queries: The early success of data mining has been based on association-rule mining of market-basket data. This success is attributed not only to the appropriateness of the objective but also to the use of new query-optimization ideas, such as the a-priori trick, that make association-rule mining run fast. Using the query flock framework, I have developed several rule-based algorithms (of which the a-priori trick is a simple instance) for generating query plans for complex data-mining queries. I have also developed several cost-based optimization algorithms [9]. My current research is focused on extending the cost-based techniques, developing methods for obtaining the required estimates and statistics, and combining them with traditional query optimization techniques.

Integration of Data Mining and Relational Database Systems: There are several architectural approaches to integrating data mining with a traditional DBMS. Most current data-mining systems employ a loosely-coupled approach, using the DBMS only to store and retrieve the data. In my work, I have focused on the tightly-coupled approach because it takes full advantage of the assortment of services offered by a conventional database system. Under this approach, the query processing is performed within the DBMS. I also showed that the performance with this approach can be improved dramatically with the help of an external optimizer that makes up for the limitations of commonly implemented optimization techniques [1,8].

Structure Discovery for Semistructured Data: The current surge of interest in semistructured data reflects its wide proliferation, especially on the Web. The lack of a fixed schema makes semistructured data easy to generate or extract but hard to browse and query. To alleviate this problem, I have developed several algorithms that uncover the implicit structure in semistructured data. The main contribution is proposing approximate schema and investigating the tradeoff between the size of the schema and its precision. I have considered varying degrees of schema complexity: from approximate typing based on attribute importance to greatest-fixpoint semantics of monadic datalog programs [3,4].

Semistructured Databases: I have developed algorithms for concise representation of semistructured data based on the idea of the lowest common denominator of structure among many semistructured objects. In effect, this concise representation provides a summary of the represented data that closely resembles Document Type Definition (DTD) in XML. I have also designed and implemented a prototype in the Lore project that evolved into a query-by-example interface [5,7].

Synopsis of Research Contributions (cont.)

Information Integration: I have actively participated in the TSIMMIS project at Stanford. I was involved in redesigning and implementing the engine of the on-line information integration TSIMMIS prototype. I added a schema management module as a first step towards automatic structure discovery of autonomous information sources [6].

Research and Work Experience

1995 - PRESENT — *Database Group, Stanford University*, Research Assistant. Conducting research in the areas of data mining (MIDAS project), semistructured data (TSIMMIS and Lore projects), and on-line integration of heterogeneous information sources (TSIMMIS project).

- Developed general optimization techniques for data-mining queries [9].
- Designed and implemented a system for efficient data mining integrated with relational database systems [1,8].
- Co-designed framework generalizing association-rule data mining and implemented prototype [2].
- Developed algorithms and implemented prototypes for obtaining approximate schema for semistructured data [3,4].
- Participated in redesigning and implementing the engine of on-line data integration TSIMMIS prototype [6].
- Developed algorithms for concise representation of semistructured data. Designed and implemented prototype in the Lore project [5,7].

SUMMER 1998 — *Research and Development Lab, Hitachi America,Ltd., Santa Clara, California.* Designed and implemented the Query Flock Compiler, a system for efficient data mining integrated with relational database system. Developed algorithms for translation of complex data mining queries into a sequence of optimizable SQL queries. Performed experiments on health-care data and achieved speedup of several orders of magnitude.

SUMMER 1994 — *BARRA,Inc., Berkeley, California*, Member of the KRONOS team. Participated in the development of a software application for portfolio optimization and risk characterization. Designed and implemented parts of the front-end and report-generating components.

1993 - 1994 — *Robotics Lab, Stanford University*, Cooperated in the development of the Machine Learning Library in C++. Responsible for the module generating random data sets according to specifications.

Teaching Experience

SPRING 1998 *Teaching Assistant* for CS99H: Stanford Introductory Seminar: Programming and Problem Solving Seminar, at Stanford University. Assisted in designing course format and content, mentored undergraduate students.

WINTER 1998 *Teaching Assistant* for CS245A: Database System Principles, at Stanford University. Prepared and graded assignments, assisted in preparing exams, occasionally lectured.

FALL 1996 *Teaching Assistant* for CS345: Advanced Topics in Database Systems, at Stanford University. Supervised student projects, prepared and graded homeworks, assisted in preparing exams, occasionally lectured.

1992 - 1994 *Tutor* for Center for Teaching and Learning, School of Engineering, and Athletics Department, at Stanford University. Tutored undergraduate and graduate students, individually and in groups, in numerous courses in Mathematics, Computer Science, and Physics.

Professional Activities

- Reviewer for *ACM SIGMOD Conference*, *ACM Principles of Database Systems Conference*, *IEEE ICDE Conference*, *Knowledge Discovery and Data Mining Conference*, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, *IEEE Transactions on Knowledge and Data Engineering*, *SIGMOD Record.*
- Member of ACM SIGMOD.

Other Activities

1995 - 1998 *Bulgarian Association at Stanford*, member of the Board.

1996 - 1997 *ACM Programming Contest at Stanford*, organizer of the local competition and assistant coach of the Stanford team.

OCT 1995 *Bulgarian Association at Stanford*, founder.

Languages

Bulgarian (native), English (fluent), French (good), Russian (good).

Extracurricular Interests

soccer, tennis, skiing, cinema, travel

Personal Information

Born on September 22, 1971 in Plovdiv, Bulgaria. Citizen of Bulgaria, holding an F-1 student visa.

Publications (available from `http://www-db.stanford.edu/~evtimov/pubs.html`)

[1] S. Nestorov and D.Tsur, "Integrating Data Mining with Relational DBMS: A Tightly-Coupled Approach"; *Proceedings of 4th Workshop on Next Generation Information Technologies and Systems (NGITS'99)*, Zikhron-Yaakov, Israel, July 1999

[2] D. Tsur, J. Ullman, C. Clifton, S. Abiteboul, R. Motwani, S. Nestorov, and A. Rosenthal, "Query Flocks: A Generalization of Association-Rule Mining"; *Proceedings of 27th ACM SIGMOD Conference*, pages 1-12, Seattle, Washington, June 1998

[3] S. Nestorov, S. Abiteboul, and R. Motwani, "Extracting Schema from Semistructured Data"; *Proceedings of 27th ACM SIGMOD Conference*, pages 295-306, Seattle, Washington, June 1998

[4] S. Nestorov, S. Abiteboul, and R. Motwani, "Inferring Structure from Semistructured Data"; *SIGMOD Record (Special Issue on Management of Semistructured Data)*, pages 39-43, December 1997; also appears in *Proceedings of SIGMOD Workshop for Semistructured Data*, held in conjunction with ACM SIGMOD'97, Tucson, Arizona, May 1997.

[5] S. Nestorov, J. Ullman, J. Wiener, and S. Chawathe "Representative Objects: Concise Representations of Semistructured, Hierarchical Data "; *Proceedings of 13th ICDE Conference*, pages 79-90, Birmingham, England, April 1997.

[6] J. Hammer, M. Breunig, H. Garcia-Molina, S. Nestorov, V. Vassalos, and R. Yerneni, "Template-Based Wrappers in the TSIMMIS System"; Demonstration Section of the *Proceedings of 26th ACM SIGMOD Conference*, pages 532-535, Tucson, Arizona, May 1997.

[7] D. Quass, J. Widom, R. Goldman, K. Haas, Q. Luo, J. McHugh, S. Nestorov, A. Rajaraman, H. Rivero, S. Abiteboul, J. Ullman, and J. Wiener "LORE: A Lightweight Object REpository for Semistructured Data"; Demonstration Section of the *Proceedings of 25th ACM SIGMOD Conference*, page 549 , Montreal, Canada, June 1996.

Technical Reports

[8] S. Nestorov and D.Tsur, "Query Flock Compiler: Efficient Implementation of Query Flocks"; *Research and Development Lab, Hitachi America, Ltd.*, October 1998.

In Preparation

[9] S. Nestorov, "Optimization Techniques for Data-Mining Queries", February 1999.

References

Prof. Jeffrey Ullman
Stanford University
Computer Science Department
Gates Hall 4A, Room 411
Stanford, CA 94305-9040
(650) 725-4802
`ullman@db.stanford.edu`

Prof. Rajeev Motwani
Stanford University
Computer Science Department
Gates Hall 4B, Room 474
Stanford, CA 94305-9040
(650) 723-6045
`rajeev@cs.stanford.edu`

Dr. Serge Abiteboul
INRIA-Rocquencourt
78153 Le Chesnay Cedex
France
(33-1) 39 63 55 37
`Serge.Abiteboul@inria.fr`

Dr. Shalom (Dick) Tsur
Surromed, Inc.
1060 East Meadow Circle
Palo Alto, CA 94303
(650) 849-6331
`tsur@surromed.com`

Prof. Joachim Hammer
Dept. of Computer and Information Science and Engineering
CSE Building, Room E452
Box 116125
University of Florida
Gainesville, FL 32611-6125
(352) 392-2687
`jhammer@cise.ufl.edu`