Multiple Query Optimization for Complex Pattern Mining Requests over Data Streams (Elke Angelika Rundensteiner)

In this presentation, I'll describe our recent research, supported by a National Science Foundation grant on providing high-performance support for processing data mining requests on data streams. This work has been done jointly with collaborators Di Yang and Matt Ward.
We observe that in diverse applications ranging from stock trading to traffic monitoring, data streams are continuously monitored by multiple analysts for extracting patterns of interest in real-time. Such complex pattern mining requests cover a broad range of popular mining query types, including detection of clusters, outliers, nearest neighbors, and top-k requests. These analysts often submit similar pattern mining requests yet customized with different parameter settings. In this work, we exploit classical principles for core database technology, namely, multi-query optimization, now in the context of data mining. That is, we present a methodology for optimizing and shared processing of a large number of such pattern mining requests with arbitrary parameter settings. Given the high algorithmic complexity of the mining process, serving multiple such requests in a single system is extremely resource intensive. The naive method of detecting and maintaining patterns for different queries independently is often infeasible in practice, as its demands on system resources increase dramatically with the cardinality of the query workload.

As foundation, we analyze the interrelations between the patterns identified by queries with different parameters settings, including both pattern-specific and window-specific parameters. First, we introduce an incremental representation of patterns identified by queries with different pattern-specific parameters. This leads to a characterization of the conditions under which the patterns identified by multiple such queries can be incrementally maintained in a single compact structure. Second, by leveraging the potential overlaps among sliding windows, we

propose a metaquery strategy which utilizes a single query to answer multiple queries with different window-specific paramters. By combining these two techniques, our framework realizes integrated maintenance for patterns identified by queries with arbitrary parameter settings. It achieves significant savings of computational and memory resources due to shared execution. Our comprehensive experimental study, using real data streams from domains of stock trades and moving object monitoring, demonstrates that our solution is significantly faster than the independent execution strategy, while using only a small portion of memory space compared to the independent execution. We also show that our solution scales in handling large numbers of queries on the order of hundreds or even thousands under high input data rates. Lastly, we will briefly introduce a visual paradigm we have developed to simplify the task of analysts in tracking such extracted patterns in real time.


Elke Rundensteiner of WPI.

Elke A. Rundensteiner is a Full Professor in the Computer Science Department of Worcester Polytechnic Institute (WPI), and the director of the database systems research laboratory (DSRG[1]) at WPI. Elke received her B.S. degree (Vordiplom) from the Johann Wolfgang Goethe University, Frankfurt, West Germany, in 1984, a Master's degree from the Florida State University, Tallahassee, in 1987, and a Ph.D. degree from the University of California, Irvine, in 1992; all in Computer Science.

Prof. Rundensteiner is an internationally recognized expert in databases and information systems, having spend 20 years of her career focussing on the development of scalable data management technology in support of advanced applications including business, engineering, and sciences. Her current research interests include scalable data stream processing, query optimization, complex event analytics, information integration and visual exploration, and data

[1] http://davis.wpi.edu:8180/DSRG/

warehousing for distributed systems. She has over 300 publications in these and related areas. Her publications on view technology, database integration, and data evolution are widely cited, and her research software prototypes released to public domain have been used by academic and non-profit groups around the world. Her research has been funded by government agencies including NSF, NIH, DOE and by industry and government labs including IBM, Verizon Labs, GTE, HP, NEC, Mitre Corporation, and others.

She has been recipient of numerous honors, including NSF Young Investigator, Sigma Xi Outstanding Senior Faculty Researcher, and WPI Board of Trustees' Outstanding Research and Creative Scholarship award, and the 2010 Chairman's Exemplary Faculty Prize. She is on program committees of prestigious conferences in the database field, has been editor of several journals, including Associate Editor of the IEEE Transactions on Data and Knowledge Engineering Journal, and of the VLDB Journal, and PC chair of several conferences, most recently EDBT'2012.