

# ICDE 2006 Ph.D. Workshop Report

Wai Gen Yee  
Dept. of Computer Science  
Illinois Institute of Technology  
10 W. 31<sup>st</sup> Street  
Chicago, IL 60616  
yee@iit.edu

Shamkant B. Navathe  
College of Computing  
Georgia Institute of Technology  
801 Atlantic Drive  
Atlanta, GA 30332  
sham@cc.gatech.edu

## ABSTRACT

This report summarizes the Ph.D. Workshop held in conjunction with the 2006 IEEE International Conference on Data Engineering. This report includes a summary of the technical presentations as well as the panelist discussion.

## 1. INTRODUCTION

The goal of the 2006 IEEE International Conference on Data Engineering's Ph.D. Workshop is to increase the exposure of Ph.D. students to the conference and academic environments in a supportive and constructive way. This year's Workshop consisted of two parts: technical presentations of attendees' work as well as a panelist discussion. More details on the workshop, including papers and slides can be found on the Web at [1].

This structure of the report corresponds to the structure of the Workshop. Technical content and then panelist discussion are summarized in sequence.

## 2. SUMMARY OF TECHNICAL PROGRAM

The topics selected for the Workshop's technical program roughly mirrors the current research general research directions in data management. Contributors were also from several parts of the world, reflecting, perhaps, regional area focuses.

We divided the paper presentation into four sessions: Data Stream Management Systems, Query Processing, Data Mining and Semantics, and Distributed Systems. Because of the nature of the Workshop, some work may be ongoing, and we will describe outstanding issues where appropriate.

The first talk of the Data Stream Management System session was entitled *Control-Based Load Shedding in Data Stream Management Systems* by Yi-Chen Tu and Sunil Prabakar. This work shows how query accuracy can be maintained while reducing data loss by applying control theory and related feedback mechanisms.

The second talk of the Data Stream Management System session was entitled *Scalable and Adaptable Distributed Stream Processing* by Yonghuan Zhou. This work describes how a hierarchical architecture of streamed data

workload coordinators that distribute load among stream data clients can be scalable, simple to implement, and cost-effective.

The third talk of the Data Stream Management System session was entitled *Processing High-Volume Stream Queries on a Supercomputer* by Eric Zeitler and Tore Risch. This talk described an experience of implementing a stream query processor on a supercomputer for the LOFAR system (a software radio telescope) and went into detail about how design decisions impacted cost and performance.

The first talk of the Query Processing session was entitled *Supporting Predicate-Window Queries in a Data Stream Management System* by Thanaa M. Ghanem. This work claims that the use of a "predicate window" over streaming data, which considers a streamed tuple being semantically equivalent to a traditional update of to a materialized view, is a more powerful generalization of existing "sliding window" operations on stream data. This semantic equivalence allows queries over the predicate window to be expressed using minimum extensions to SQL.

The second talk of the Query Processing session was entitled *Optimization of Complex Queries in Relational Databases* by Bin Cao. This talk described how SQL queries containing nested subqueries that are either aggregate or non-aggregate can be optimized by transformations that limit redundant computations.

The third talk of the Query Processing session was entitled *Twig Query Processing under Concurrent Updates* by Christian Mathis and Theo Harder. This work describes a technique for allowing high concurrency in an environment where XML data are being updated, subject to transactional properties, by using fine-grain locks.

The first talk of the Data Mining and Semantics session was entitled *A Generalized Framework for Mining Spatial and Spatio-temporal Patterns in Scientific Data* by Hui Yang and Srinivasan Parthasarathy. This talk described the application of data mining techniques to extract information scalably from geometric objects, possibly capturing temporal information, for scientific computing applications.

The second talk of the Data Mining and Semantics session was entitled *Searching and Ranking Documents Based on Semantic Relationships* by Boarnerges Aleman-Meza. This talk described the use of named entities on documents and their semantic relationships to improve query result ranking, much in the same way link analysis is a structural means of improving result ranking.

The third talk of the Data Mining and Semantics session was entitled *Using Data Extraction Ontology to Foster Automating Semantic Annotation* by Yihong Ding and David W. Embley. This talk described an automated Web page annotator that works by using ontologies as an input to the data extraction process to help guide the annotation process and to make system adaptive to a large number of domains. Further, object extraction is integrated into this process to improve performance.

The fourth talk of the Data Mining and Semantics session was entitled *Model Video Semantics with Constraints Considering Temporal Structure and Typed Events* by Yu Wang, Lizhu Zhou, and Jianyong Wang. This talk described a way of modeling and constraining entities and events temporally in a way that is suitable for video annotation and querying.

The first talk of the Distributed Systems session was entitled *Location-based Spatial Queries with Data Sharing in Mobile Environments* by Wei-Shinn Ku and Roger Zimmermann. This talk describes a system for finding nearest neighbor queries in a mobile peer-to-peer environment. Each query uses partial results from other peers as well as the triangle inequality to estimate nearest neighbor results.

The second talk of the Distributed Systems session was entitled *MoSCoE: A Framework for Modeling Web Service Composition and Execution* by Jyotishman Pathak, Samik Basu, Robyn Lutz, and Vasant Honavar. This talk described a system for incremental composition of Web services that allows for a high degree of flexibility in case of partial failure of the composition.

The third talk of the Distributed Systems session was entitled *Ant Algorithms for Search in Unstructured Peer-to-Peer Networks* by Elke Michlmayr. This talk described a way of routing peer-to-peer queries in a random network by incrementally learning best paths by using a technique similar to pheromone trail blazing done by ants.

### 3. SUMMARY OF PANELIST DISCUSSION

The goal of the panelist session was two-fold: ask for technical advice on their work or ask about perspectives on career choices. As many of the students attendees were about to graduate, they focused almost exclusively on the latter goal. Furthermore, as most of the student attendees were interested in academic

jobs, most questions focused on attaining and succeeding in an academic career.

The panelists were selected in an attempt to mix up seniority levels. They were:

- Laura Bright (post-doctoral researcher, Portland State University)
- Ramez Elmasri (professor, University of Texas, Arlington)
- Christopher Jermaine (assistant professor, University of Florida)
- Vijay Kumar (professor, University of Missouri, Kansas City)
- Wai Gen Yee (assistant professor, Illinois Institute of Technology)
- Sham Navathe (professor, moderator, Georgia Institute of Technology)

Students were first asked about how they would be judged when looking for an academic position. The criteria include (in no particular order) pedigree, publication record, letters of recommendation, demographic profile. Pedigree (i.e., having a degree from a top-ranked institution) indicates that a student has trained with the leaders in an area. Hiring someone with a good pedigree increases the chance for collaborations and associations between the respective institutions. On a more subtle note, the embarrassment on the part of the hiring institution of hiring someone with a good pedigree who ultimately fails is less than if that someone came from a less prestigious institution.

There is no doubt that a graduating student must publish. The question is whether it is the quality or the quantity of publications that matters. The consensus is that quality of publication is more important than quantity. Hiring committees want to know what the peak intellectual output of a person is, and assume that, given enough resources (e.g., equipment, students), publication quantity will happen.

Letter of recommendations are important to help hiring committees understand how much a particular student contributed to his/her stated research. For example, was the student the initiator of his/her research ideas? Was he the workhorse? Does s/he work well in teams? Does he understand the big picture? Hiring committees would likely avoid candidates who are not going to be able to produce results independently but will also avoid those who isolate themselves.

The demographic profile of a candidate is important as well. Although an institution will first search for scholarship and departmental need, all else being equal, an American university is likely to give preference to Americans and to underrepresented ethnic and gender groups. The justification for this is that it may introduce more perspectives into the environment and improve the educational potential of the department. Note that foreign applications are welcome at most institutions; however, many smaller institutions do not have the recruiting budget to invite such candidates.

A few of the students asked about the possibility of working in industry for a few years before returning to academia. Most of the panelists scoffed at this idea: it is simply too difficult, in most settings, to maintain a publication stream while working a full-time industry job.

#### **4. CONCLUSIONS AND POSSIBLE DIRECTIONS OF FUTURE WORK**

From our feedback from students, the ICDE06 Ph.D. Workshop was a success, having achieved its stated goals. Students particularly appreciated the ability to ask questions and get advice from Ph.D.s of various ranks from various backgrounds.

Much of the success was due to the contributions of the authors and how they were able to interact despite their disparate areas of specialization. We were also fortunate to have contributions from more senior members of the research community join various sessions and act as panelists. Their input made a significant impact.

Future Ph.D. Workshops may be improved by shortening each author's presentation time and having a poster session. This would allow more specialized student-to-student interaction as well as greater student participation,

as we learned in organizing the Midwest Database Research Colloquium [2].

Furthermore, we also noticed that many of the questions posed during the panel session were the same questions that have been echoed in our personal advisement of students as well as when we were students. We believe that such questions are general to all Ph.D. students and believe that a repository of such questions and answers should be compiled as a student reference. Such a repository could, say, store the collected wisdom from other recent attempts at group advisement (e.g., [4]).

In effect, we could form a persistent community advisement system. Technically, a Web blog may be the appropriate means of implementing such a system (perhaps [3]). The success of such a system would depend on its broad support from the entire community. We would be happy to hear any feedback or pledges of support for such an idea.

#### **5. ACKNOWLEDGMENTS**

We would like to thank the paper contributors, the attendees, the program committee and the ICDE organizers for their support of the Workshop. We would also like to acknowledge Brian Cooper for his idea on community advisement of database students.

#### **REFERENCES**

- [1] ICDE 2006 Ph.D. Workshop Web Site. <http://ir.iit.edu/~waigen/icde06phd/>.
- [2] Midwest Database Research Symposium Web Site. <http://dais.cs.uiuc.edu/mwdbrs/>.
- [3] DB Advisor Blog. <http://dbadvisor.blogspot.com/>.
- [4] SIGMOD 2006 Life After Graduation Symposium Web Site. <http://tangra.si.umich.edu/clair/sigmod-pods06/graduation.htm>.