COMAD 2006 TUTORIAL

Privacy preserving data publication: From Generalization to Anatomy

Yufei Tao
Chinese University of Hong Kong
Sha Tin, New Territories, Hong Kong SAR, China
taoyf@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/~taoyf

ABSTRACT

Companies and organizations often need to publish clients' information to institutions for research purposes. For example, a hospital periodically releases patients' diagnostic records so that medical scientists can study the correlation between diseases and various factors. Privacy preservation is an important topic in data publication. First, the publication should be fuzzy enough to disallow any adversary to figure out the exact medical history of any patient. On the other hand, the released data must be sufficiently precise to enable effective analysis. In this tutorial, we will review the existing techniques for striking an appropriate balance, in order to maximize the accuracy of data investigation, without breaching any patient's privacy.

Speaker's Profile: Yufei Tao is the winner of the Hong Kong Young Scientist Award 2002, conferred by the Hong Kong Institution of Science. He holds a PhD degree in computer science from the Hong Kong University of Science and Technology, and did his post-doc as a visiting scientist in the Computer Science department of the Carnegie Mellon University, from 2002 to 2003. In the next three years, he was an assistant professor at the City University of Hong Kong. Currently he is an assistant professor at the Department of Computer Science and Engineering, the Chinese University of Hong Kong. Prof. Tao is engaged in research of database systems. His research interests include temporal databases, spatial databases, approximate query processing, data privacy and security. He has published extensively in renowned conferences and journals including ACM SIGMOD, VLDB, IEEE ICDE, ACM TODS, IEEE TKDE, VLDB JOURNAL, etc.

(Duration: 1.5 Hours)

COMAD 2006 TUTORIAL

Multilingual Database Systems

Jayant Haritsa
Database Systems Lab, SERC
Indian Institute of Science, Bangalore 560012, India
haritsa@dsl.serc.iisc.ernet.in
http://dsl.serc.iisc.ernet.in/~haritsa

ABSTRACT

Efficient storage and query processing of data spanning multiple natural languages are of crucial importance in today's globalized world. A primary prerequisite to achieve this goal is that the defacto standard data repositories – relational database systems – should efficiently and seamlessly support multilingual data. In this tutorial, we will first present a detailed assessment of how good today's database systems (both commercial and public-domain) are with regard to the storage, management and processing of multilingual data. Our results will show that there are significant performance inefficiencies for languages based on scripts other than Latin (such as Devanagari, Kanji, Cyrillic, etc.). We will also outline techniques for alleviating these problems.

With regard to functionality, a major limitation of SQL is that it does not support querying of data across different natural languages, that is, cross-lingual queries. To address this lacuna, we will propose two new SQL operators that support phoneme-based matching of names, and ontology-based matching of concepts, in the multilingual world.

An algebra for integrating these new operators with relational systems will be defined as well as the associated cost models, selectivity estimators, and access methods. Our experience with a prototype implementation of these operators on PostgreSQL will be highlighted.

In a nutshell, this tutorial will present practical approaches towards realizing the ultimate goal of "natural-language-neutral" database engines.

Speaker's Profile: Jayant Haritsa is a Professor in the Supercomputer Education & Research Centre and in the Department of Computer Science & Automation at the Indian Institute of Science, Bangalore. He received the BTech degree in Electronics and Communications Engineering from the Indian Institute of Technology (Madras), and the MS and PhD degrees in Computer Science from the University of Wisconsin (Madison). His research interests are in database systems and real-time systems. He is a member of IEEE, ACM, and the Computer Society of India, and is an associate editor of the Real-Time Systems journal and the IEEE Data Engineering Bulletin.

(Duration: 3 Hours)

COMAD 2006 TUTORIAL

Secure Data Outsourcing

Radu Sion
Computer Science Department
Stony Brook University
Stony Brook, NY 11794-4400, USA
sion@cs.sunysb.edu
http://www.cs.sunysb.edu/~sion

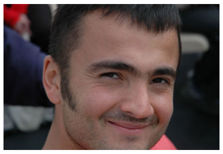
ABSTRACT

The networked and increasingly ubiquitous nature of today's data management services mandates assurances to detect and deter malicious or faulty behavior. This is particularly relevant for outsourced data frameworks in which clients place data management with specialized service providers. Clients are reluctant to place sensitive data under the control of a foreign party without assurances of confidentiality. Additionally, once outsourced, privacy and data access correctness (data integrity and query completeness) become paramount.

Today's solutions are fundamentally insecure and vulnerable to illicit behavior, because they do not handle these dimensions. In this tutorial we will discuss existing solutions and future designs for robust, efficient, and scalable data outsourcing mechanisms providing strong security assurances of (1) correctness, (2) confidentiality, and (3) data access privacy.

There exists a strong relationship between such assurances; for example, the lack of access pattern privacy usually allows for statistical attacks compromising data confidentiality. Confidentiality can be achieved by data encryption. However, to be practical, outsourced data services should allow expressive client queries (e.g., relational joins with arbitrary predicates) without compromising confidentiality. This is a hard problem because decryption keys cannot be directly provided to potentially untrusted servers. Moreover, if the remote server cannot be fully trusted, protocol correctness become essential. Therefore, solutions that do not address all three dimensions are incomplete and insecure.

It is important to design query mechanisms targeting outsourced relational data that (i) ensure queries have been executed with integrity and completeness over their respective target data sets, (ii) allow queries to be executed with confidentiality over encrypted data, (iii) guarantee the privacy of client queries and data access patterns. We will discuss protocols that adapt to the existence of trusted hardware – so critical functionality can be delegated securely from clients to servers. We will exemplify with practical protocols handling binary predicate JOINs with full privacy in outsourced scenarios.



Speaker's Profile: Radu Sion is an Assistant Professor in Computer Sciences at Stony Brook University. He is a member of the Network Security and Applied Cryptography (NSAC) Lab. His research interests are in Information Assurance, Applied Cryptography and Network Security. Instances are: wireless and sensor networks security, digital rights management, secure data outsourcing, queries over encrypted data, reputation systems, integrity proofs in sensor networks, secure storage in peer to peer and ad-hoc environments, data privacy and bounds on illicit inference over multiple data sources, security and policy management in computation/data grids.

(Duration: 3 Hours)

COMAD 2006 TUTORIAL

High Performance Data Mining: Consolidation and Renewed Bearing

Srinivasan Parthasarathy
Dept. of Computer Science and Engineering and Dept. of Biomedical Informatics
The Ohio State University
Columbus, OH-43210, USA
srini@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~srini

ABSTRACT

Over the years the definition of high performance computing has taken on various forms as a function of the types of technical and creative uses and the underlying semantics of applications driving them. Traditional definitions often refer to the problem of using high end parallel computers to meet the need of applications. However in the modern context, high performance computing ranges from fast sequential algorithms that target memory and I/O performance on modern processors all the way to work on the computational and data grid.

In this tutorial, I will describe the impact of high performance computing, under this broad definition, on the field of data mining. I will review and consolidate some of the key algorithmic developments over the last decade as they pertain to high performance data mining. Specifically we will examine parallel and sequential performance of such algorithms on modern high performance systems. In the latter part of this tutorial, I will present an outlook towards the future paying particular attention to recent technological advances (e.g. multi-core architectures, InfiniBand networks etc.) that I expect will have a bearing on this field of research. I will conclude by describing, why in light of these advances I expect that existing data mining algorithms will need to be re-architected and improved to realize performance commensurate with these technological advances. I also will describe why I believe they will lead to a renewed set of challenges for algorithm development and associated systems support spanning areas such as compilers, runtime, database, middleware and hardware systems.

Speaker's Profile: Srinivasan Parthasarathy is an Associate Professor in the Computer Science and Engineering Department at the Ohio State University (OSU). He heads the data mining research laboratory and has a joint appointment in the department of biomedical informatics at OSU. He is a recipient of an NSF CAREER award, a DOE Early Career Award, and an Ameritech Faculty fellowship. His papers have received several awards from leading conferences in the field including an IEEE Data Mining 2002 best paper, a SIAM Data Mining 2003 best paper, the Very Large Databases Conference (VLDB) 2005 best research paper and a "Best of SIAM Data Mining 2005" selection. He is a member of the ACM and the IEEE and serves on the editorial board of IEEE Intelligent Systems and is currently serving as one of the program chairs of SIAM Data Mining in 2007.

(Duration: 3 Hours)

COMAD 2006 TUTORIAL

Scalable Information Extraction and Integration

Sunita Sarawagi
KR School of Information Technology
Indian Institute of Technology, Bombay, India
sunita@it.iitb.ac.in
http://www.it.iitb.ac.in/~sunita

ABSTRACT

Many applications over text require efficient methods for extracting and integrating structured data from large unstructured sources. This tutorial reviews the state of the art approaches for information extraction and duplicate elimination. First, we present an overview of the methods used, with particular emphasis on machine learning based approaches including sequential models like Conditional Random Fields and their generalizations. Second, we present scalable techniques for deploying these models on large unstructured text collections. We review key approaches for scaling up information extraction, including using general-purpose search engines as well as indexing techniques specialized for information extraction applications. We also overview scalable techniques for integrating the extracted information using approximate join algorithms and fuzzy index lookups. We highlight research opportunities and challenges that remain.

Speaker's Profile: Sunita Sarawagi researches in the fields of databases, data mining, machine learning and statistics. She is associate professor at IIT Bombay. Prior to that she was a Research Staff Member at IBM Almaden Research Center. She got her PhD in databases from the University of California at Berkeley and a bachelors degree from IIT Kharagpur. She has several publications in databases and data mining including a best paper award at the 1998 ACM SIGMOD conference and several patents. She is on the editorial board of the ACM TODS and ACM KDD journals and was editor-in-chief of the ACM SIGKDD newsletter. She has served as program committee member for ACM SIGMOD, VLDB, ACM SIGKDD and IEEE ICDE, ICML conferences.

(Duration: 3 Hours)

COMAD 2006 TUTORIAL

An Introduction to Data Grid Management Systems (DGMS)

Arun Jagatheesan
San Diego Supercomputer Center,
University of California at San Diego
La Jolla, CA 92093-0505, USA
arun@sdsc.edu
http://www.sdsc.edu/~arun

ABSTRACT

If we analyze the history of computer science, most of the contributions including the DBMS were the result of a requirement that was pushing the limits of an existing technology. One such requirement today is to manage very large unstructured data that is distributed in multiple countries using traditional file systems. Some of the FORTUNE 500 companies face this problem today, due to out-sourcing and distributed global teams that collaborate with each other. Data Grid Management Systems (DGMS) manage collaborative global sharing of very large amounts of unstructured data amongst multiple teams. The core concepts of a DGMS are very similar to traditional RDBMS. A DGMS could be considered as a logical namespace (or a logical distributed file system) of heterogeneous data storage resources from multiple sub-organizations. DGMS are powered by relational databases and provide both system and user-defined schema to organize and query data. In this tutorial, we introduce DGMS concepts and explain with real use cases why such a system is needed in very large academic data centers and major companies. Novices and experts in distributed data management will have a chance to learn about this emerging technology, research problems and business opportunities.

Speaker's Profile: Arun Jagatheesan ("Arun") is a Dataflow/Data grid Specialist at the San Diego Supercomputer Center (SDSC) in University of California, San Diego. His research interests include Data Grid Management Systems (DGMS), peer-to-peer data management, and workflow management systems. Arun works on research, development and standardization of data grid technologies by collaborating with multiple academic and commercial organizations, as part of the SDSC Storage Resource Broker (SRB) Project. He is the founder and technical lead of the SRB Matrix Project on Gridflow Management Systems.

He is currently involved in many data grid projects at SDSC including the new LUSciD collaboration, a joint effort by the University of California and Lawrence Livermore National Laboratory (LLNL) exploring the software requirements for managing very large amount of data. Arun also plays an active role in the LSST project that will manage hundreds of petabytes of data.

Arun was previously an OPS faculty member at the University of Florida. He has published many papers and provided multiple invited talks or tutorials on data grids at multiple technical conferences.

(Duration: 2 Hours)