Lynne and William Frankel Center for Computer Science
Department of Computer Science, Ben-Gurion University of the Negev
Beer-Sheva, Israel 84105. Tel: (972-8) 6428032 Fax: (972-8) 6428021

Saal Auditorium Rm.202, Alon High-Tech Building 37

Seminar Series Supported by Jeffrey and Holly Ullman

Database Day

June 5th, 2011

10:00 Coffee & Tagging

10:30 Opening Remarks and Greeting

Shlomi Dolev and Ehud Gudes

10:40 A Pattern Based Approach for Secure Database Design

Jenny Abramov, Ben-Gurion University, Israel

Abstract: Security in general and database protection from unauthorized access in particular, are crucial for organizations. Although it has long been accepted that system requirements should be considered from the early stages of the development, non-functional requirements, such as security, tend to be neglected and dealt-with only at the end of the development process. Various methods have been proposed, however, none of them provide a complete framework to guide, enforce and verify the correct implementation of security policies within a system design, and generate source code from it. In this paper we present a novel approach that guides database designers, to design a database schema that complies with the organizational security policies related to access authorization. First, organizational policies are defined in the form of security patterns. Then, during the application development, the patterns guide the implementation of the security requirements and the correct application of the patterns is verified. Finally, the secure database schema is automatically generated .Joint work with Arnon Sturm and Peretz Shoval

Pattern Rewriting Framework for Event Processing Optimization

Ella Rabinovich, Technion, Israel

Abstract: A growing segment of event-based applications require both strict performance goals and support in the processing of complex event patterns. Event processing patterns have multiple complexity dimensions: the semantics of the language constructs (e.g. sequence) and the variety of semantic interpretations for each pattern (controlled by policies). We introduce in this work a novel approach for pattern rewriting that aims at efficiently processing patterns which comprise all levels of complexity. We present a formal model for pattern rewriting and demonstrate its usage in a comprehensive set of rewriting techniques for complex pattern types, taking various semantic interpretations into account. A cost model is presented, balancing processing latency and event throughput according to user's preference. Pattern cost is then estimated using simulation-based techniques. This work advances the state-of-the-art by analyzing complex event processing logic and by using explicit means to optimize elements that were considered "black box." Our empirical study yields encouraging results, with improvement gain of up to tenfold relative to the non optimized solutions that are used in the current state-of-the-art systems.

11:15 Private Data Analysis

Kobbi Nissim, Ben-Gurion University and Microsoft, Israel

Abstract: I will review the (almost)-decade-old research towards rigorous foundations for privacy in data analysis. A notion that has emerged in this line of research is Differential Privacy. I will motivate differential privacy, show basic techniques for constructing differentially private analyses, and relate differential privacy with other notions of theoretical computer science . No prior background will be assumed.

11:55 Coffee Break

12:10 Towards Provenance for SQL (and beyond!)

Daniel Deutch, Ben-Gurion University, Israel

Abstract:The annotation of tuples in the results of database transformations with provenance information was shown to be extremely useful for different purposes, such as trust management, cost, security, incomplete and probabilistic databases, and deletion propagation. Recent work has proposed a generic framework, based on semirings, for capturing these applications uniformly, for query languages of various expressivity. What was critically missing from these works, however, is the support of aggregate queries and queries with difference, used heavily in practice. We identify yardsticks for a ``good framework for '' managing provenance, and use them to formally prove that the previous approach is inapplicable for handling such iesquer. We cconsequently propose a new approach, where we annotate with provenance information not just tuples, but also the individual values within tuples, using provenance to describe the values computation. Finally, we briefly discuss the challenges in extending this approach for managing fine-grained provenance for workflows .

12:50 Towards Proactive Event-driven Computing

Opher Etzion, IBM Haifa Research, Israel

Abstract: The processing of events allow enterprises and individuals to move from responsive mode, where the computerized system responds to the user's request to reactive mode, where the computerized system reacts to an event or a detected situation based on event. This talks discusses the next phase is this evolution - the proactive mode, where the computerized system proactively reacts to a predicted event or state, in order to mitigate or eliminate reaching undesired states. The talk will describe work done in IBM Research in this area, which involves event processing, real-time optimization and causality model. The talk is self-contained and accompanied by various examples from different areas, and explains the principles of the concepts and facilities of such a system, and the research and engineering challenges that are required to make such technology in pervasive use.

13:30 Lunch Break

14:15 Keynote lecture – Map Reduce And Its Children

Jeff Ullman, Stanford University, USA

Abstract: Map-reduce is a programming system that supports easy parallelism on computing clusters and at the same time copes with hardware faults during a computation. It has been extended to allow arbitrary acyclic data flows in a number of ways. Extension to recursive data flows is more difficult for two reasons. First, fault management depends on tasks never having delivered output if they fail (the "blocking property"). Recursive tasks can never be blocking. Second, many recursive algorithms require many rounds, the first few of which produce most of the answer. In a cluster-computing environment, communication overhead makes it expensive to communicate many small files (the "endgame" problem). We shall discuss how extensions to map reduce deal with faults when the tasks are non-blocking. Then, we look at the redesign of certain recursive algorithms such as transitive closure to cope with the endgame problem by minimizing the number of recursive rounds without increasing the total communication. In other cases, we can show it is impossible to do so, although many open questions of this kind remain .

15:15 Parallel XML Query Algorithm Lila Shnaiderman, Technion

Abstract: XML is based on a tree-structured data model. Naturally, the most popular XML querying language (XPath) uses patterns of selection predicates on Multiple elements related by a tree structure. These are abstracted by twig patterns. Finding all occurrences of such a twig pattern in an XML database is a core operation for XML query processing.

We present the Parallel Path Stack algorithm (PPS) and the Parallel Twig Stack algorithm (PTS). PPS and PTS are novel and efficient algorithms for matching XML query twig patterns in a parallel multi-threaded computing platform. PPS and PTS are based on the PathStack and TwigStack algorithms [Bruno et al(2002)Bruno, Koudas, and Srivastava.[These algorithms employ a sophisticated search technique for limiting stream processing to specific subtrees. We propose a novel scheme for working with Terabyte sized documents. This storage scheme is based on the (Extended XML stream storage) model for managing XML streams. This model allows us to encode some structural details for specific subtrees of the main XML document tree, within the stream representation of the document. We conducted extensive experimentation with PPS and PTS. We compared PPS and PTS to the standard (sequential) PathStack and Twig Stack algorithms in terms of run time (to completion). We checked their performance for varying numbers of threads. Experimental results indicate that using PPS/PTS 1 significantly reduces the running time of queries in comparison with the PathStack/TwigStack algorithm(up to 44 times faster for DBLP queries and up to 22 times faster for XMark queries).

On Provenance Minimization Yael Amsterdamer,TAU, Israel

Abstract: Provenance information has been proved to be very effective in capturing the computational process performed by queries, and has been used extensively as the input to many advanced data management tools (e.g. view maintenance, trust assessment, or query answering in probabilistic databases). We study here the core of provenance information, namely the part of provenance that appears independently of the query plan that is in use. This provenance core is informative as it describes the part of the computational process that is inherent to the query. It is also useful as a compact input to the above mentioned data management tools. We study algorithms that, given a query, compute an equivalent query that realizes the core provenance for all tuples in its result. We study these algorithms for queries of varying expressive power. Finally, we observe that, in general, one would not want to require database systems to execute a specific query plan that realizes the core provenance, but instead to be able to find, possibly off-line, the core provenance of a given tuple in the output (computed by an arbitrary query plan), without re-evaluating the query. We provide algorithms for such direct computation of the core provenance.

15:50 Coffee Break

16:05 Non-homogeneous anonymizations

Tamir Tassa, The Open University, Israel

Abstract: Up until recently most of the research on Privacy Preserving Data Publishing considered partition-based anonymization models. The approach in such models is to partition the database records into groups and then homogeneously generalize the quasi-identifiers in all records within a group. We describe in this talk alternative anonymization models which are not based on partitioning and homogeneous generalization. Such models extend the set of acceptable anonymizations of a given table, whence they allow achieving similar privacy goals with much less information loss. We shall briefly review the basic models of homogeneous anonymization (e.g. k-anonymity and l-diversity) and then define non-homogeneous anonymization, discuss its privacy, describe algorithms and demonstrate the advantage of such anonymizations in reducing the information loss. We shall then discuss the usefulness of those models for data mining purposes. In particular, we will show that the reduced information loss that characterizes such anonymizations translates also to enhanced accuracy when using the anonymized tables to learn classification models.
Based on joint works with Aris Gionis, Arnon Mazza, Mark Last and Sasha Zhmudyak

16:45 Efficient Entity resolution with MFIblocks

Batya Kenig, Technion, Israel

Abstract: Entity resolution is the process of discovering groups of tuples that correspond to the same real world entity. In order to avoid the prohibitively expensive comparison of all pairs of tuples, blocking algorithms separate the tuples into blocks which are highly likely to contain matching pairs. Tuning is a major challenge in the blocking process. In particular, contemporary blocking algorithms need to provide a blocking key, based on which tuples are assigned to blocks. Due to the tuning complexity, blocking keys are manually designed by domain experts. In this work, we introduce a blocking approach that avoids selecting a blocking key altogether, relieving the user from this difficult task. The approach is based on maximal frequent item sets selection. This approach also allows early evaluation of block quality based on the overall commonality of its members. A unique feature of the proposed algorithm is the use of prior knowledge of the estimated sizes of duplicate sets in enhancing the blocking accuracy. We report on a thorough empirical analysis, using common benchmarks of both real-world and synthetic datasets which exhibit the efficiency of our approach.
Joint work with Avigdor Gal

Diversified Recommendations for Semantic-less Collaborative Filtering Rubi Boim, TAU, Israel

Abstract: This paper considers a popular class of recommender systems that are based on Collaborative Filtering (CF) and proposes a novel technique for diversifying the recommendations that they give to users. Items are clustered based on a unique notion of priority-medoids that provides a natural balance between the need to present highly ranked items vs. highly diverse ones. Our solution estimates items diversity by comparing the rankings that different users gave to the items, thereby enabling diversification even in common scenarios where no semantic information on the items is available. It also provides a natural zoom-in mechanism to focus on items (clusters) of interest and recommending diversified similar items. We present DiRec, a plug-in that implements the above concepts and allows CF Recommender systems to diversify their recommendations. We illustrate the operation of DiRec in the context of a movie recommendation system and present a thorough experimental study that demonstrates the effectiveness of our recommendation diversification technique and its superiority over previous solutions.

Evaluation of Schema Matching Tasks: Between Prophecy and Hindsight Tomer Sagi, Technion, Israel

Abstract: Schema matching is recognized to be one of the basic operations required by the process of data and schema integration. Over the years different algorithms have been proposed and refined for matching schemata and the related task of ontology alignment. These algorithms have been roughly divided to similarity measures (a.k.a First line matchers) and matchers (a.k.a second line matchers) . When presented with a specific schema matching or ontology alignment task, there is a need to a candidate set of algorithms an appropriate subset for the task. However, in order to choose wisely, a method of evaluating schema matchers w.r.t the problem at hand is required. To date, evaluation of schema matching and ontology alignment has been done using various posterior measures, most notably precision and recall. All of these measures compare the correspondence list generated by the matching algorithm with an exact match or reference alignment. This hindsight is of little use when evaluating algorithms for a new task. Asking an oracle's opinion is always an option, however few researchers seem to choose it. Our ultimate goal is to provide a generalized evaluation framework that allows evaluation of both similarity measures and matchers without the benefit of neither hind sight nor prophets. In the talk we will present several of the directions we are currently exploring.

17:40 End of Database Day