Data Management Research at the Middle East Technical University

Nihan K. Cicekli, Ahmet Cosar, Asuman Dogac, Faruk Polat, Pinar Senkul, I. Hakki Toroslu and Adnan Yazici

Department of Computer Engineering Database Group
Middle East Technical University (METU)
06531 Ankara Turkey

1. INTRODUCTION

The Middle East Technical University (METU) (http://-www.metu.edu.tr) is the leading technical university in Turkey. The department of Computer Engineering (http://-www.ceng.metu.edu.tr) has twenty seven faculty members with PhDs, 550 undergraduate students and 165 graduate students. The major research funding sources include the Scientific and Technical Research Council of Turkey (TÜBÍTAK), the European Commission, and the internal research funds of METU. Data management research conducted in the department is summarized in this article.

2. WEB SERVICES AND SEMANTIC WEB TECHNOLOGIES

The research in the semantic Web services area has concentrated upon the application of this technology in two important sectors: healthcare through the Artemis project (http://www.srdc.metu.edu.tr/webpage/projects/-artemis/) and tourism through the Satine project (http://-www.srdc.metu.edu.tr/webpage/projects/satine/).

2.1 The Artemis Project

The Artemis project provides the interoperability of medical information systems through semantically enriched Web services. An essential element in defining the semantic of Web services is domain knowledge. Medical informatics is one of the few domains to have considerable domain knowledge exposed through standards as mentioned above. These standards offer significant value in terms of expressing the semantic of Web services in the healthcare domain. In the Artemis project, prominent healthcare standards are used to semantically annotate Web services as follows:

- HL7 has categorized the events in the healthcare domain by considering service functionality that reflects the business logic in this domain. We use this classification as a basis for defining the service action semantics through a "Service Functionality Ontology". In this way, semantic discovery of Web services is facilitated.
- Given the complexity of clinical domain, the Web service messages exchanged have innumerous segments of different types and optionality. To make any use of these messages at the receiving end, their semantics must be clearly defined. We annotate the Web services through the reference information models of Electronic Healthcare Record (EHR) standards.

The details of this work is presented in the following publication:

- Dogac, A., Laleci, G., Kirbas, S., Kabak, Y., Sinir, S., Yildiz, A., Gurcan, Y., "Artemis: Deploying Semantically Enriched Web Services in the Healthcare Domain", Information Systems Journal, Elsevier, to appear.

Using archetypes is a promising approach for providing semantic interoperability among healthcare systems. To realize archetype based interoperability, healthcare systems need to discover the existing archetypes based on their semantics, annotate their archetypes with ontologies, compose templates from archetypes and retrieve corresponding data from the underlying medical information systems. In the Artemis project, we use ebXML Registry semantic constructs for annotating, storing, discovering and retrieving archetypes.

The details of this work is presented in the following publication:

- Dogac, A., Laleci, G. B., Kabak, Y., Unal, S., Beale, T., Heard, S., Elkin, P., Najmi, F., Mattocks, C., Webber, D., "Exploiting ebXML Registry Semantic Constructs for Handling Archetype Metadata in Healthcare Informatics", Intl. Journal of Metadata, Semantics and Ontologies, to appear.

In the Artemis project, AMEF (Artemis Message Exchange Framework) is developed to provide the exchange of meaningful clinical information among healthcare institutes through semantic mediation. The framework involves first providing the mapping of source ontology into target message ontology with the help of a mapping tool that produces a mapping definition. This mapping definition is then used to automatically transform the source ontology message instances into target message instances. Through a prototype implementation, we demonstrate how to mediate between HL7 Version 2 and HL7 Version 3 messages. However, the framework proposed is generic enough to mediate between any incompatible healthcare standards that are currently in use. The details of this work is presented in the following publication:

- Bicer, V., Laleci, G., Dogac, A., Kabak, Y., "Artemis Message Exchange Framework: Semantic Interoperability of Exchanged Messages in the Healthcare Domain", ACM Sigmod Record, Vol. 34, No. 3, September 2005.

2.2 The SATINE Project

The SATINE project addresses the interoperability in the travel domain. The tourism industry today is the second largest economic sector, after manufacturing in the world. Tourism embarked on eBusiness earlier than other sectors. Currently, travel information services are dominantly provided by Global Distribution Systems (GDSs). All the airlines, many hotel chains and car rental companies list their inventory with major GDSs. A GDS gives its subscribers pricing and availability information for multiple travel products such as flights. Travel agents, corporate travel departments, and even Internet travel services, subscribe to one or more GDSs. However, small and medium-sized enterprises, for example "bed and breakfast" type accommodation or companies hiring bicycles, restaurants and a host of others cannot participate to GDS-based eBusiness activities because selling their products through GDSs is too expensive for them.

Furthermore, GDSs are legacy systems and suffer from a number of problems: they mostly rely on private networks, are mainly for human use, have difficult to use cryptic interfaces, have limited speed and search capabilities, and are difficult to interoperate with other systems and data sources. The implication is that the tour operators, travel agencies, etc. cannot benefit fully from the advantages of electronic business-to-business trading.

In order to facilitate eBusiness, the travel industry has formed a consortium called the Open Travel Alliance (OTA). OTA produces XML schemas of message specifications to be exchanged between the trading partners, including availability checking, booking, rental, reservation, query services, and insurance. However, not every travel company's applications can be expected to produce and consume OTA compliant messages.

In the SATINE project, we describe how to deploy semantically enriched travel Web services and how to exploit semantics through Web service registries to addres the problems mentioned. We also address the need to use the semantics in discovering both Web services and Web service registries through peer-to-peer technologies. The mechanisms are described in detail in the following publication:

- Dogac, A., Kabak, Y., Laleci, G., Sinir, S., Yildiz, A., Kirbas, S., Gurcan, Y., "Semantically Enriched Web Services for the Travel Industry", ACM Sigmod Record, Vol. 33, No. 3, September 2004.

Web services, similar to their real life counterparts, have several properties and, thus, truly useful semantic information can only be defined through standard ontology languages. In the SATINE project, mechanisms to enrich ebXML registries through OWL-S ontologies for describing the Web service semantics are developed. Particularly, how the various constructs of OWL can be mapped to ebXML classification hierarchies and how the services are discovered through standardized queries by using the ebXML query facility are described.

Detailed information on these mechanisms is available in the following publication:

- Dogac, A., Kabak, Y., Laleci, G. B., Mattocks, C., Najmi, F., Pollock, J., "Enhancing ebXML Registries to Make them OWL Aware", Distributed and Parallel Databases Journal, Springer Verlag, Vol. 18, No. 1, July 2005, pp. 9-36.

Finally, how privacy issues can be handled semantically in Web services is addressed in the following publication:

- Tumer, A., Dogac, A., Toroslu, I. H., "A Semantic based Privacy Framework for Web Services", WWW'03 workshop on E-Services and the Semantic Web (ESSW 03), Budapest, Hungary, May 2003.

3. VIDEO DATABASES, SPATIO-TEMPORAL DATABASES

Content-based querying of multimedia data is a relatively new subject, which has arisen fast with the improvements in data processing and communication systems technologies. We present a spatio-temporal video data model that allows efficient and effective representation and querying of spatio-temporal objects in videos. The data model is focused on the semantic content of video streams. Objects, events and activities performed by objects, and temporal and spatial properties of objects are the main interests of the model. Spatial and temporal relationships between objects and events are dynamically calculated with the query processing methods introduced in this paper. The model is flexible enough to define new relationship types between objects without changing the data model and supports fuzzy querying of spatio-temporal objects in videos. Index structures are used for effective storage and retrieval of the mentioned properties. A prototype of the proposed model has been implemented. The prototype allows various spatio-temporal queries along with some fuzzy ones, and it is capable of implementing many other queries without any major changes in the data model.

- Koprulu, M., Cicekli, N. K., Yazici, A., "Spatio-temporal Querying in Video Databases", Information Sciences, Vol. 160, 2004, pp. 131-152.
- Koprulu, M., Cicekli, N. K., and Yazici, A., "Spatio-Temporal Querying in Video Databases", Proc. of the Sixth International Conf. on Flexible Query Answering Systems (FQAS'2002), Denmark, Oct 2002.
- Yazici, A., Yavuz, O., and George, R., "An MPEG-7 Based Video Database management System, Flexible Querying and Reasoning in Spatio-Temporal Databases: Theory and Applications", In Springer's Geo-sciences/Geoinformation series by Springer Verlag, 2004, pp. 181-210.

Depending on the proposed content-based spatio-temporal video data model, a natural language interface is implemented to query the video data. The queries, which are given as English sentences, are parsed using Link Parser, and the semantic representations of given queries are extracted from their syntactic structures using information extraction techniques. At the last step, the extracted semantic representations are used to invoke the related parts of the underlying spatio-temporal video data model to retrieve the results of the queries.

- Erozel, G., Cicekli, N. K., Cicekli, I., "Natural Language Interface on a Video Data Model", Proceedings of IASTED DBA 2005, Austria.

4. INTELLIGENT DATABASE SYSTEMS, FUZZY LOGIC AND DATABASE MODELING

Next generation information system applications require powerful and intelligent information management that necessitates an efficient interaction between database and knowledge base technologies. It is also important for these applications to incorporate uncertainty in data objects, integrity constraints, and applications.

Fuzzy relational database models generalize the classical relational database model by allowing uncertain and imprecise information to be represented and manipulated. We introduce fuzzy extensions to the relational database model. Within this framework of fuzzy data representation, similarity, conformance of tuples, the concept of fuzzy functional dependencies, and partial fuzzy functional dependencies are utilized to define the fuzzy key notion, transitive closures, and fuzzy normal forms. The fuzzy object-oriented data model is a fuzzy logic-based extension to the object-oriented database model, which permits uncertain data to be explicitly represented. The details are presented in the following publications:

- Koyuncu, M., and Yazici, A., "A Fuzzy Knowledge-Based System for Intelligent Retrieval", IEEE Transactions on Fuzzy Systems, to appear.

- Sozer, A., and Yazici, A., "Design and Implementation of Index structures for Fuzzy Spatial Databases", International Journal of Intelligent Systems, to appear.

- Bahar, O., and Yazici, A., "Normalization And Lossless Join Decomposition of Similarity-Based Fuzzy Relational Databases", International Journal of Intelligent Systems, Vol. 19, No. 10, October 2004, pp. 885-918.

- Aygun, R. S., and Yazici, A., "Modeling and Management of Fuzzy Information in Multimedia Database Applications", Multimedia Tools and Applications, Vol. 24, No.1, September 2004, pp. 29-56.

- Koyuncu, M., and Yazici, A., "IFOOD: An Intelligent Fuzzy Object-Oriented Database Architecture", IEEE Trans. on Knowledge and Data Engineering, September 2003, pp. 1137-1154.

- Sozat, M. I., Yazici, A., "A Complete Axiomatization for Fuzzy Functional and Multivalued Dependencies in Fuzzy Database Relations", Fuzzy Sets and Systems, Vol. 117/2, 2001, pp. 161-181

- Yazici, A., Zhu, Q., Sun, N., "Semantic Data Modeling of Spatiotemporal Database Applications", Int. Journal of Intell. Systems, Vol. 16, No. 7, July 2001, pp. 881-904

5. WORKFLOWS

The research on workflow management systems has concentrated on modeling and scheduling under resource allocation systems. In addition to the temporal constraints corresponding to the order of tasks, constraints related to resource allocation are also equally important. Workflow scheduling should include the decisions about which resources are allocated to which tasks, parallel to the task ordering decision. To solve this, two approaches have been studied. In the first one, we proposed an architecture to specify and schedule workflows under resource allocation constraints as well as temporal and causality constraints:

- Senkul, P., and Toroslu, I. H., "An architecture for workflow scheduling under resource allocation constraints", Information Systems, Vol. 30, Issue 5, July 2005, pp. 399-422.

In the second approach, we developed a new logical formalism, called Concurrent Constraint Transaction Logic (CCTR), which integrates Constraint Logic Programming (CLP) and Concurrent Transaction Logic, and a logic-based workflow scheduler that is based on this new formalism. The semantics of the CCTR modeling of a workflow represent a schedule that contains both an execution ordering that the specified workflow can execute, and a set of resource assignments to the tasks of the workflow satisfying the given constraints. The details of this work is presented in the following publication:

- Senkul, P., Kifer, M., Toroslu, I. H., "A Logical Framework for Scheduling Workflows under Resource Allocation Constraints", VLDB 2002, pp. 694-705.

As another research direction, we have proposed a logic-based framework for the specification and execution of workflows. The proposed approach is based on the Kowalski and Sergot's Event Calculus:

- Cicekli, N. K., and Cicekli, I., "Formalizing the Specification and Execution of Workflows using the Event Calculus", Information Sciences, to appear.

6. DATA MINING

Most research on data mining has focused on processing single relations. Recently, however, multi-relational data mining has started to attract interest. One of the earliest works on multi-relational data mining is the query flocks technique, which extends the concept of traditional association rule mining with a "generate-and-test" model for different kinds of patterns. One possible extension of the query flocks technique is the addition of view definitions including recursive views. Although in our system the query flock technique can be applied to a database schema including both the Intensional Database (IDB) or rules, and the Extensible Database (EDB) or tabled relations, we have designed an architecture to compile query flocks from datalog into SQL in order to be able to use commercially available Database Management Systems (DBMS) as an underlying engine of our system. Currently, we are extending our work on multi-relational data mining using inductive logic programming for discovering rules. This work is presented in the following publication:

- Toroslu, I. H., Yetisgen-Yildiz, M., "Data Mining in Deductive Databases Using Query Flocks", Expert Systems with Applications, Vol. 28, Issue 3, April 2005, pp. 395-407.

One of the most well-known data mining problems is the sequential pattern mining. The main drawback of sequential pattern mining is the large number of sequential patterns discovered, which makes it harder for the decision maker to interpret them. Thus, a new parameter is added to the definition of the sequential pattern mining problem that represents the repetitions of the sequences. In the following publication, this problem is introduced together with its possible applications and advantages over ordinary sequential pattern mining:

- Toroslu, I. H., "Repetition Support and Mining Cyclic Patterns", Expert Systems with Applications, Vol. 25, issue 3, october 2003, pp. 303-311.

7. QUERY PROCESSING

The "Multiple Query Optimization" (MQO) problem, a well-known query-processing problem in databases, has been studied in the database literature since the 1980s. MQO tries to reduce the execution cost of a group of queries by performing common tasks only once. In our work, we assume that, at the beginning of the optimization, all promising alternative plans have been generated and shared tasks are identified. Our algorithm finds an optimal solution to the MQO problem. This form of MQO problem has been formulated as an NP-complete optimization problem where several heuristic functions are used to guide an A* search. In the following work, we propose the first dynamic programming approach to this problem:

- Toroslu, I. H., and Cosar, A., "Dynamic Programming Solution for Multiple Query Optimization Problem", Information Processing Letters, Vol. 92, Issue 3, 15 November 2004, pp. 149-155.

There has been several other previous works done by the members of the database group at METU focused on other aspects of multiple query optimization and semantic query optimization. Our most recent work on semantic query optimization is as follows:

- Polat, F., Cosar, A., Alhajj, R., "The Semantic Information-based Alternative Plan Generation for Multiple Query Optimization", Information Sciences, Elsevier, Vol. 137/1-4, 2001, pp. 103-133.

8. OBJECT-ORIENTED DATABASES

Research on object-oriented databases is primarily based on extensibility and dynamic schema evolution. We have benefited from having an object algebra maintaining closure that makes it possible to have output from a query persistent in the hierarchy. We automate the process of properly placing new classes in the class hierarchy. We are able to map a view, which can easily be defined in our model, that is intended to be persistent into a class. The object algebra is utilized to handle basic schema evolution functions without requiring any special set of built-in functions.

- Alhajj, R., Polat, F., and Yilmaz, C., "Views as First-Class Citizens in Object-Oriented Databases", *The VLDB Journal*, Vol. 14, No. 2, April 2005, pp. 155-169.

- Alhajj, R., and Polat, F., "Rule-Based Schema Evolution in Object-Oriented Databases", *Knowledge-Based Systems*, Vol. 16, No. 1, Jan. 2003, pp. 47-57.

- Alhajj, R., and Polat, F., "Reengineering of Relational Databases into Object-Oriented: Constructing the Class Hierarchy and Migrating the Data", *Proceedings of the IEEE Working Conference on Data Reverse Engineering WCRE 2001*, IEEE Press, Stuttgart, Germany, Oct. 2001, pp. 335-344.

- Alhajj, R., and Polat, F., "Transferring Database Contents from a Conventional Information System to a Corresponding Existing Object Oriented Information System", *Proceedings of the ACM Annual Symposium on Applied Computing*, ACM Press, Las Vegas, USA, Mar. 2001, pp. 220-224.

9. BIOINFORMATICS

We worked on the problem of identifying Differentially Expressed Genes (DEG) and improved the power of PaGE by estimating prior probability used in the confidence computation. We developed two methods based on the *q-values* and *maximum likehood* approaches to find DEG on a dataset of microarray experiments for pattern generation. We formulate the control problem for dynamic discrete regulatory networks and defined various control-monitor strategies. The Multiobjective control problem is further studied and a case study is done for proof of correctness.

- Abul, O., Alhajj, R., Polat, F., and Barker, K., "Finding Differentially Expressed Genes: Pattern Generation", *Bioinformatics*, Vol. 21, No. 4, Feb 2005, pp. 445-450.

- Abul, O., Alhajj, R., and Polat, F., "Importance of Monitoring in Developing Optimal Control Policies for Probabilistic Boolean Genetic Networks", *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (IPMU2004), Perugia, Italy, July 2004, pp. 1145-1152.

- Abul, O., Alhajj, R., and Polat, F., "Markov Decision Processes Based Optimal Control Policies for Probabilistic Boolean Network", *Proc. of IEEE Symposium on Bioinformatics and Bioengineering* (BIBE2004), Taichung, Taiwan, May 19-21, 2004, pp. 337-344.

- Abul, O., Alhajj, R., Polat, F., and Barker, K., "Finding Differentially Expressed Genes: Pattern Generation using q-values", *Proceedings of the ACM Annual Symposium on Applied Computing (SAC)*, Cyprus, Mar. 2004, pp. 138-142.

- Abul, O., Lo, A., Alhajj, R., Polat, F., and Barker, K., "Cluster Validity Analysis Using Subsampling", *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 2, Washington DC, Oct. 2003, pp. 1434-1440.