Database Research at Bilkent University

Özgür Ulusoy
Bilkent University
Computer Engineering Department
Bilkent, Ankara, TURKEY
oulusoy@cs.bilkent.edu.tr

1. Introduction

This report provides a brief description of the research activities of the Database Research Group of Bilkent University. The current research of the group is mainly focused on the topics of *Multimedia Databases* (in particular video database management and content-based retrieval of document images), *Web Databases* (in particular the modeling and querying of Web resources and focused crawling), and *Mobile Computing* (in particular moving object processing and mobile data management).

An overview of each research topic investigated is provided together with the list of researchers, any external funding information, and a selection of associated publications. Interested readers should contact Özgür Ulusoy at oulusoy@cs.bilkent.edu.tr for further information. Most of the publications listed in this report are available at:
http://www.cs.bilkent.edu.tr/~oulusoy/~pubs.html.

2. Multimedia Databases

Researchers: Ö. Ulusoy, U. Güdükbay, E. Çetin, E. Şaykol, C. Alper, I. S. Altıngövde, T. Sevilmiş.
Past Researchers: M. E. Dönderler, U. Arslan, G. Ünel, A. K. Sinop.

Funding Sources: Scientific and Technical Research Council of Turkey (TÜBİTAK) under grant number EEEAG-199E025, Turkish State Planning Organization (DPT) under grant number 2004K120720, and European Commission 6th Framework Program, MUSCLE NoE Project, under grant number FP6-507752.

2.1 BilVideo: A Video Database Management System

We have developed a prototype video database management system, called *BilVideo* [1, 2, 3]. The architecture of *BilVideo* is original in that it provides full support for spatio-temporal queries that contain any combination of directional, topological, 3D-relation, object-appearance, trajectory-projection, and similarity-based object-trajectory conditions by a rule-based system built on a knowledge-base, while utilizing an object-relational database to respond to semantic (keyword, event/activity, and category-based), color, shape, and texture queries. The knowledge-base of *BilVideo* consists of a fact-base and a comprehensive set of rules implemented in Prolog. The rules in the knowledge-base significantly reduce the number of facts that need to be stored for spatio-temporal querying of video data [4].

To respond to user queries containing both spatio-temporal and semantic conditions, the query processor interacts with both the knowledge-base and a feature database, where the system stores fact-based and semantic metadata, respectively. Intermediate query results returned from these two system components are integrated seamlessly by the query processor, and final results are sent to Web clients. Raw video data and its features are stored in a separate database. The feature database contains video semantic properties to support keyword, event/activity, and category-based queries. The *video-annotator tool*, which we developed as a Java application, generates and maintains the features. The fact-base, which is a part of the knowledge-base, is populated by the *fact-extractor tool*, which is also a Java application [3].

BilVideo provides support for retrieving any segment of a video clip, where the given query conditions are satisfied, regardless of how video data is semantically partitioned. Object trajectories, object-appearance relations, and spatio-temporal relations between video objects are represented as Prolog facts in a knowledge-base, and they are not explicitly related to the semantic units of videos. Thus, precise answers can be returned for user queries, when requested, in terms of frame intervals.

BilVideo has a simple, yet very powerful SQL-like query language, which currently supports a broad range of spatio-temporal queries on video data [5]. We are currently working on integrating support for semantic and low-level (color, shape, and texture) video queries as well. We completed our work on semantic video modeling, which was reported in [6]. As for the low-level queries, our *Fact-Extractor* tool also extracts color, shape, and texture histograms of the salient objects in video keyframes. We have also developed a Web-based visual query interface for specifying video queries visually over the Internet. Furthermore, we have completed our work on the optimization of spatio-temporal video queries [7].

The *BilVideo* query language is designed to be used for any application that needs video query processing facilities. Hence, the language provides query support through external predicates for application-dependent data.

The Web-based Query Interface of BilVideo and its user manual are available at http://pcvideo.cs.bilkent.edu.tr/. A demo of the Web-based Query Interface can be seen at:
http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo/webclient.avi.

2.2 Ottoman Archive Content-Based Retrieval System

The Ottoman Archive Content-Based Retrieval system is a Web-based program that provides electronic access to digitally stored

Ottoman document images. The Ottoman script is a connected script based on the Arabic alphabet. A typical word consists of compounded letters as in handwritten text. We have developed a framework for content-based retrieval of historical documents in the Ottoman Empire archives [8]. The documents are stored as textual images, which are compressed by constructing a library of symbols that occur in a document. The symbols in the original image are then replaced by the pointers into the codebook library. For symbol extraction, we use the features in wavelet and spatial domain based on angular and distance span shapes. Symbols are extracted from document images by a scale-invariant process. User can specify a query as a rectangular region in an input image, and content-based retrieval is achieved by applying the same symbol extraction process to the query region. The query is processed on the codebook of documents, and the resulting documents are ranked in the decreasing order of their similarity to the query image, which is determined by the total number of symbols matched with the query region. The resulting documents are presented by identifying the matched region of each document in a rectangle. The querying process does not require decompression of images.

The techniques we use in our work are not specific to documents with Ottoman script. They can easily be tailored to other domains of archives containing printed or handwritten documents.

The web site of the Ottoman Archive Content-Based Retrieval system is:
http://www.cs.bilkent.edu.tr/bilmdg/ottoman/webclient.html.

A step-by-step user manual describing how to access the Web query interface and specify queries is available at:
http://www.cs.bilkent.edu.tr/~bilmdg/ottoman/manual/manual.htm.

References

[1] M.E. Dönderler, E. Saykol, Ö. Ulusoy, U. Güdükbay. BilVideo: A Video Database Management System. *IEEE Multimedia*, 10, 5 (January/March 2003), 66-70.

[2] Ö. Ulusoy, U. Güdükbay, M.E. Dönderler, E. Saykol, C. Alper. BilVideo Video Database Management System (demo paper). In *Proceedings of the International Conference on Very Large Databases (VLDB'04)*. (Toronto, Canada, August-September 2004), 1373-1376.

[3] M.E. Dönderler, E. Saykol, U. Arslan, Ö. Ulusoy, U. Güdükbay. BilVideo: Design and Implementation of a Video Database Management System. To appear in *Multimedia Tools and Applications*, 2005.

[4] M.E. Dönderler, Ö. Ulusoy, U. Güdükbay. A Rule-based Video Database System Architecture. *Information Sciences*, 143, 1-4 (June 2002), 13-45.

[5] M.E. Dönderler, Ö. Ulusoy, U. Güdükbay. Rule-based Spatio-temporal Query Processing for Video Databases. *VLDB Journal*, 13, 1 (January 2004), 86-103.

[6] U. Arslan, M.E. Dönderler, E. Şaykol, Ö Ulusoy, U. Güdükbay. A Semi-Automatic Semantic Annotation Tool for Video Databases. In *Proceedings of the Workshop on Multimedia Semantics (SOFSEM'02)*. (Milovy, Czech Republic, November 2002), 1-10.

[7] G, Ünel, M.E. Dönderler, Ö. Ulusoy, U. Güdükbay. An Efficient Query Optimization Strategy for Spatio-Temporal Queries in Video Databases. *Journal of Systems and Software*, 73, 1 (September 2004), 113-131.

[8] E. Şaykol, A. K. Sinop, U. Güdükbay, Ö. Ulusoy, E. Çetin. Content-Based Retrieval of Historical Ottoman Documents Stored as Textual Images. *IEEE Transactions on Image Processing*, 13, 3 (March 2004), 314-325.

3. Web Databases

Researchers: Ö. Ulusoy, İ. S. Altıngövde, Ö. N. Subakan.
Past Researchers: S. A. Özel, M. Kutlutürk.
Collaborators: G. Özsoyoğlu, Z. M. Özsoyoğlu, A. Al-Hamdani (Case Western Reserve University).

Funding Sources: A joint grant of Scientific and Technical Research Council of Turkey (TÜBİTAK) under grant number 100U024 and National Science Foundation of the USA under grant number INT-9912229.

3.1 Metadata-Based Modeling of the Web

A recent approach to increase the quality of Web search is associating metadata to the resources on the Web. To this end, there are various standardization efforts and initiatives, such as the Dublin Core Framework, RDF, Semantic Web and Topic Maps. In [9, 10], we address the problem of modeling Web information resources using expert knowledge and personalized user information for improved Web searching capabilities. We propose a "Web information space" model, which is composed of Web-based information resources (HTML/XML documents on the Web), expert advice repositories (domain-expert-specified metadata for information resources), and personalized information about users (captured as user profiles that indicate users' preferences about experts as well as users' knowledge about topics).

Expert advice, the heart of the Web information space model, is specified using topics and relationships among topics (called metalinks), along the lines of the recently proposed topic maps. Topics and metalinks constitute metadata that describe the contents of the underlying HTML/XML Web resources. The metadata specification process is semi-automated. It exploits XML DTDs to allow domain-expert guided mapping of DTD elements to topics and metalinks. In particular, the domain expert specifies a mapping between the entities of our metadata model (topics and metalinks) and the XML DTDs in the corresponding domain(s). An agent then traverses the Web, extracts topics and metalinks for those XML files conformant with the input DTD, and stores them into a local object-relational database management system, which will then serve as an expert advice (metadata) repository for these visited Web resources.

To demonstrate the practicality and usability of the proposed Web information space model, we have created a prototype expert

advice repository of more than one million topics and metalinks for DBLP (Database and Logic Programming) Bibliography data set.

3.2 Querying Web Metadata: Native Score Management and Text Support in Databases

In this work, we discuss the issues involved in adding a native score management system to object-relational databases, to be used in querying web metadata (that describes the semantic content of web resources) [11, 12]. As described in the preceding section, the web metadata model is based on topics (representing entities), relationships among topics (metalinks), and importance scores (sideway values) of topics and metalinks. We extend database relations with scoring functions and importance scores. We add to SQL score-management clauses with well-defined semantics, and propose the sideway-value algebra (SVA), to evaluate the extended SQL queries.

SQL extensions include clauses for propagating input tuple importance scores to output tuples during query processing, clauses that specify query stopping conditions, threshold predicates—a type of approximate similarity predicates for text comparisons, and user-defined-function-based predicates. The propagated importance scores are then used to rank and return a small number of output tuples. The query stopping conditions are propagated to SVA operators during query processing. We show that our SQL extensions are well-defined, meaning that, given a database and a query Q, the output tuples of Q and their importance scores stay the same under any query processing scheme.

3.3 Focused Web Crawling

A preliminary step for extracting and querying metadata for Web resources is gathering all and only the relevant Web resources for a particular application domain or topic of interest. To this end, in [13] we propose a *rule-based focused crawling* strategy to crawl the Web and construct a repository of relevant Web pages. A focused crawler is an agent that concentrates on a particular target topic, and tries to visit and gather only relevant pages from the Web. In the literature, one of the approaches for focused crawling is using a canonical topic taxonomy and a text classifier to train the crawler so that those URLs that most probably point to on-topic pages will be identified and prioritized. Our research explores using simple rules derived from the linkage statistics among the topics of a taxonomy while deciding on the crawler's next move, i.e., to select the URL to be visited next. The rule based approach improves the harvest rate and coverage of the taxonomy-based focused crawler and also enhances it to support *tunneling*. More specifically, the rule based crawler can follow a path of off-topic pages that may at last lead to high quality on-topic pages. More information on our projects in *Web Databases* can be found through http://www.cs.bilkent.edu.tr/~bilweb.

References

[9] I. S. Altingovde, S. A. Özel, Ö. Ulusoy, G. Özsoyoğlu, Z. M. Özsoyoğlu. Topic-Centric Querying of Web Information Resources. In *Proceedings of the Database and Expert Systems Applications (DEXA'01)*, *Lecture Notes in Computer Science (Springer Verlag)*, vol.2113, (Munich, Germany, September 2001) 699-711.

[10] S. A. Özel, I. S. Altingovde, Ö. Ulusoy, G. Özsoyoğlu, Z. M. Ösoyoglu. Metadata-Based Modeling of Information Resources on the Web. *Journal of the American Society for Information Science and Technology (JASIST)*, 55, 2 (January 2004), 97-110.

[11] G. Özsoyoğlu, Abdullah Al-Hamdani, I. S. Altingovde, S. A. Özel, Ö. Ulusoy, Z. M. Özsoyoğlu. Sideway Value Algebra for Object-Relational Databases. In *Proceedings of the International Conference on Very Large Databases (VLDB'02)*. (Hong Kong, August 2002), 59-70.

[12] G. Özsoyoğlu, I. S. Altingovde, A. Al-Hamdani, S. A. Özel, Ö. Ulusoy, Z. M. Özsoyoğlu. Querying Web Metadata: Native Score Management and Text Support in Databases. *ACM Transactions on Database Systems*, 29, 4 (December 2004), 581-634.

[13] I. S. Altingovde, Ö. Ulusoy. Exploiting Interclass Rules for Focused Crawling**.** *IEEE Intelligent Systems*, 19, 6 (November-December 2004), 66-73.

4. Mobile Computing

Researchers: Ö. Ulusoy, M. Karakaya, İ. Körpeoğlu, S. Çıracı.
Past Researchers: , Y. Saygın, E. Kayan, J. Tayeb, G. Gök, I. Yoncacı, G. Yavaş.
Collaborators: O. Wolfson (University of Illinois at Chicago), K. Y. Lam (City University of Hong Kong), D. Katsaros, Y. Manolopoulos (Aristotle University), A. K. Elmagarmid (Purdue University).

Funding Sources: Scientific and Technical Research Council of Turkey (TÜBİTAK) under grant number EEEAG-246, NATO Collaborative Research Program under grant number CRG 960648, and the bilateral program of scientific cooperation between Turkey and Greece (TÜBİTAK grant number 102E021 and *Γ.Γ.Ε.Τ.*).

4.1 Moving Object Processing

Our earlier work on moving object processing includes indexing locations of moving objects. In [14], we propose an indexing technique for moving objects based on a variant of the quadtree data structure in which the indexing directory is in primary memory and the indexed data resides in secondary storage. The method is useful in any application that involves dynamic attributes whose values vary continuously according to a given function of time. Our approach is based on the key idea of using a linear function of time for each dynamic attribute that allows us to predict its value in the future. We contribute an algorithm for regenerating the quadtree-based index periodically to minimize CPU and disk access cost.

Another important issue in moving object database management is to provide support for processing location-dependent queries, where the answer to a query depends on the current location of

the user who issued the query. A location-dependent query can become more difficult to process when it is submitted as a continuous query for which the answer changes as the user moves. In [15], we present an efficient method to monitor the locations of moving objects so that timely and accurate results can be returned to location dependent continuous queries, while minimizing the location update cost. Location-dependent queries from mobile clients may also be associated with timing constraints on their response times. In [16], we propose a method to monitor the locations of moving users/objects based on the real-time criticality of location dependent continuous queries, so that higher levels of accuracy can be achieved for the results returned to the queries categorized with higher criticality. Another related issue in processing location dependent continuous queries is the transmission of query results to the users. In [17], various methods that can be used to determine the transmission time of query results are investigated with the goal of minimizing data transmission costs.

In [18], we present a data mining algorithm for the prediction of user movements in a mobile computing system. The algorithm proposed is based on mining the mobility patterns of users, forming mobility rules from these patterns, and finally predicting a mobile user's future movements by using these rules.

4.2 Mobile Data Management

Dissemination of data by broadcasting may induce high access latency in case the number of broadcast data items is large. In [19], we propose two methods to reduce client access latency for broadcast data. Our methods are based on analyzing the broadcast history (i.e., the chronological sequence of items that have been requested by clients) using data mining techniques. The proposed methods are implemented on a Web log, and it is shown that the proposed rule-based methods are effective in improving the system performance in terms of average latency as well as cache hit ratio of mobile clients. Our other work on broadcast scheduling considers a pull-based data delivery environment [20]. We propose a variant of the *Longest Wait First* heuristic in scheduling data broadcast, which provides a practical implementation of the heuristic by avoiding the decision overhead.

One of the features that a mobile computer should provide is disconnected operation, which is performed by hoarding. The process of hoarding can be described as loading the data items needed in the future to the client cache prior to disconnection. Automated hoarding is the process of predicting the hoard set without any user intervention. In [21], we describe a generic, application-independent technique for determining what should be hoarded prior to disconnection. Our method utilizes association rules that are extracted by data mining techniques for determining the set of items that should be hoarded to a mobile computer prior to disconnection.

While there has been much research interest in mobile computing issues, an issue that has not received much attention is the management of the database of a mobile system under timing constraints. In [22], we present a mobile database system model that takes into account the timing requirements of applications supported by mobile computing systems. We provide a transaction execution model with alternative execution strategies for mobile transactions and evaluate the performance of the system under various mobile system characteristics, such as the number of mobile hosts in the system, the handoff process, disconnection, coordinator site relocation, and wireless link failure.

References

[14] J. Tayeb, Ö. Ulusoy, O. Wolfson, A Quadtree Based Dynamic Attribute Indexing Method**.** *The Computer Journal*, 41, 3 (1998) 185-200.

[15] K. Y. Lam, Ö. Ulusoy, et al., An Efficient Method for Generating Location Updates for Processing of Location-Dependent Continuous Queries**.** In *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA'01)*. (Hong Kong, April 2001) 218-225.

[16] Ö. Ulusoy, I. Yoncacı, K. Y. Lam. Evaluation of a Criticality-Based Method for Generating Location Updates. In *Proceedings of Workshop on Database Mechanisms for Mobile Applications, Lecture Notes in Informatics*, vol.43, (Karlsruhe, Germany, April 2003) 94-105.

[17] G. Gök, Ö. Ulusoy. Transmission of Continuous Query Results in Mobile Computing Systems. *Information Sciences*, 125, 1-4 (2000) 37-63.

[18] G. Yavaş, D. Katsaros, Ö. Ulusoy, Y. Manolopoulos. A Data Mining Approach for Location Prediction in Mobile Environment. *Data and Knowledge Engineering*, 54, 2 (2005) 121-146.

[19] Y. Saygın, Ö. Ulusoy. Exploiting Data Mining Techniques for Broadcasting Data in Mobile Computing Environments. *IEEE Transactions on Knowledge and Data Engineering*, 14, 6 (November/December 2002) 1387-1399.

[20] M. Karakaya, Ö. Ulusoy. Evaluation of a Broadcast Scheduling Algorithm. In *Proceedings of Advances in Databases and Information Systems (ADBIS'01), Lecture Notes in Computer Science (Springer Verlag)*, vol.2151, (Vilnius, Lithuania, September 2001) 182-195.

[21] Y. Saygın, Ö. Ulusoy, A. K. Elmagarmid. Association Rules for Supporting Hoarding in Mobile Computing Environments. In *Proceedings of IEEE International Workshop on Research Issues on Data Engineering (RIDE'00)*, (San Diego, CA, USA, February 2000) 71-78.

[22] E. Kayan, Ö. Ulusoy. An Evaluation of Real-Time Transaction Management Issues in Mobile Database Systems. *The Computer Journal,* 42, 6 (November 1999) 501-510.