Research Activities in Database Management and Information Retrieval at University of Illinois at Chicago

Isabel Cruz, Ashfaq Khokhar, Bing Liu, Prasad Sistla, Ouri Wolfson and Clement Yu

Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607-7053

{ifc, ashfaq, liub, sistla, wolfson, yu}@cs.uic.edu

Introduction

There are 6 active researchers in the areas of database management, data mining and information retrieval at the Department of Computer Science, University of Illinois at Chicago. In this article, the works of Isabel Cruz, Ashfaq Khokhar, Bing Liu, Prasad Sistla, Ouri Wolfson and Clement Yu are sketched.

1. Metasearch Engines (Clement Yu & Weiyi Meng at SUNY Binghamton)

Today, millions of people employ powerful search engines such as Google to retrieve information from the Web on a daily basis. In spite of the success, there are problems associated with such powerful search engines. First, the number of pages which are captured by a single search engine is a few billion, while it has been reported that the entire Web has about 500 billion pages and is rapidly growing. Thus, the coverage of the Web by a single search engine is rather small. Second, an index database has to be built to contain the key information of the captured Web pages. This database is huge and takes substantial amount of time to refresh its contents. Thus, it is not surprising that substantial amount of information in the indexed database can be weeks out-of-date. Third, in order to retrieve information from the large database when there are a large number of queries, enormous hardware resources are needed. It has been reported that Google is utilizing many thousands of computers.

In order to alleviate the above problems, the metasearch approach is suggested. A metasearch engine is a system that connects to numerous search engines. Thus, the coverage of a metasearch engine is much higher than that of an individual search engine. When a query is received, the metasearch engine determines the best search engines (a very small number of search engines) to invoke for answering the query. After Web pages are retrieved by these selected search engines, they are ranked by the metasearch engine in descending order of desirability and then presented to the user. Since the actual searching of Web pages are carried out by the selected search engines and not by the metasearch engine, the hardware requirement imposed on the metasearch engine should be much smaller. If the metasearch engine chooses to connect search engines with relatively small databases, then these databases take less time to refresh. As a consequence, the data will be more up to date.

Our work in the metasearch engine area consists of

(a) optimal selection of search engines for any given query.
(b) selection of Web pages to retrieve from each chosen search engine.
(c) merging of Web pages of selected search engines.
(d) utilizing linkage information among Web pages to perform search engine selection.

Representative publications on these topics can be found at (http://opal.cs.binghamton.edu/~meng/metasearch.html)

Some of the problems that we are working on consist of: (a) automatic connection of a metasearch engine to numerous search engines; (b) personalized retrieval in which concepts are automatically associated with a user query, based on his/her retrieval preferences; (c) automatic extraction of URLs retrieved by search engines.

2. Image and Video Retrieval (Prasad Sistla, Clement Yu & Alp Aslandogan at U. of Texas at Arlington)

We assume that semantic objects (for example persons) and their relationships (spatial as well as relationships involving actions such as handshake) can be extracted from low level features and other information (such as tracking the movement of a person in a video or the use of audio or captions). This allows us to represent the contents of a picture as a set of objects and their relationships. A query for images is also expressed as a set of objects and their relationships. Similarity can be defined between an image and a query. Images having largest similarities with the query are retrieved. In videos, temporal relationships between objects in a sequence of frames are also captured. Queries for video retrieval can utilize temporal operators. Some representative publications are [16][17].

Another project is to retrieve images of a person from the Web, given his/her name. The name is utilized to retrieve Web pages containing the name. However, a Web page not containing an image or if the image does not contain the face of a person, then it is not shown to the user. Images are ranked based on the occurrences of the name in the Web pages containing the image. If the name appears in the caption of the image, in the ALT field of the image or in the name of the image file, or it appears in the Web page with high frequencies of occurrences, then the image is given higher similarity. Images are ranked in descending order of similarity. This usually yields about 90% of accuracy for the top 20 images, i.e., approximately 18 of them belong to the correct person.) One reason for the inaccuracy is that labels associated with images can be wrong. For example, when the name "Jay Leno" is submitted, images of performers who performed in the Jay Leno show are retrieved, because their images are labeled Jay Leno. To improve accuracy, clustering of images from the top 30 images is performed. Only images belonging to a cluster containing quite a few images are returned. The rationale is that images of the same person will form large clusters while images of different persons will form small clusters. This raises the accuracy of retrieval to beyond 95%. Some representative publications are [3][4].

3. Scalable Content based Indexing and Retrieval for Audio and Image Data Archives

(Ashfaq Khokhar and colleagues) [1][2][10]

Research in the area of content-based indexing and retrieval (CBIR) has primarily focused on indexing and retrieval of visual information, such as image and video data. With the advent of Voice over IP services and fast packet switching networks, content based storage and retrieval of audio data is gaining tremendous interest. On the other hand, astounding advances in recent years in speech and audio recognition make it highly possible now to design efficient CBIR systems for audio data that will allow innovative audio applications involving manipulation of audio/voice data. Furthermore, content-based audio indexing will also improve efficiency and accuracy of integrated (voice-audio-video) multimedia information retrieval systems.

Audio analysis research in audio data management can be classified into three categories. One category is audio segmentation and classification. This category primarily addresses problems related to segmentation of music and speech data. The second category is audio indexing and retrieval. A variety of algorithms have been developed to perform indexing. One specific technique in content-based audio retrieval is query-by-example. Different approaches of audio retrieval have been proposed in recent years. The last category involves audio analysis for indexing and retrieval of video. In the following we provide a brief summary of our work in the context of audio/music data indexing and retrieval technologies. Our work on image indexing and retrieval can be found in [1].

Our indexing and retrieval framework for audio data is unique in the sense that instead of blindly computing indexing features, it identifies repetitive segments in the input samples. We argue that features that characterize a signal composed of repetitions of the same or similar units are different from those of an isolated original signal or "base" signal that composes it. This makes it difficult to automatically detect some similarity that is obvious to a human listener. We have developed a computationally efficient technique to automatically extract the base signal in a repetitive audio sample [1]. We have used the wavelet transform decomposition of the audio sample to analyze and index audio data. We use different statistical properties of the wavelet coefficients (number of zero crossings, STD, mean, etc.) at multiple levels of resolutions to index audio samples. This way, we obtain representation of the audio characteristics in multi-resolution time-frequency space, and then a set of statistical properties of wavelet coefficients in each wavelet subband is used to represent the audio clip. The multilevel wavelet decomposition of an audio signal highly resembles to its decomposition in sound octaves. A hierarchical indexing scheme has been developed based on the scale property of the wavelet transform [2].

4. Moving Objects Information Management

(Ouri Wolfson & Prasad Sistla) [14][15][18]

Miniaturization of computing devices, and advances in wireless communication and sensor technology are some of the forces that are propagating computing from the stationary desktop to the mobile outdoors. Some important classes of new applications that will be enabled by this revolutionary development include location-based services, tourist services, mobile electronic commerce, and digital battlefield. Some existing application classes that will benefit from the development include transportation and air traffic control, weather forecasting, emergency response, mobile resource management, and mobile workforce. Location management, i.e. the management of transient location information, is an enabling technology for all these applications. Location management is also a fundamental component of other technologies such as fly-through visualization, context awareness, augmented reality, cellular communication, and dynamic resource discovery. In this write-up we present our view of the important research issues in location management. These include modeling of location information, uncertainty management, spatio-temporal data access languages, indexing and scalability issues, data mining (including traffic and location prediction), location dissemination, privacy and security, location fusion and synchronization.

In 1996, the Federal Communications Commission (FCC) mandated that all wireless carriers offer a 911 service with the ability to pinpoint the location of callers making emergency requests. This requirement is forcing wireless operators to roll out costly new infrastructure that provides location data about mobile devices. In part to facilitate the rollout of these services, in May 2000, the U.S. government stopped jamming the signals from global positioning system (GPS) satellites for use in civilian applications, dramatically improving the accuracy of GPS-based location data to 5-50 meters.

As prices of basic enabling equipment like smart cell phones, hand helds, wireless modems, and GPS devices and services continue to drop rapidly, International Data Corp (IDC) predicts that the number of wireless subscribers worldwide will soar to 1.1 billion in 2003. Spurred by the combination of expensive new location-based infrastructure and an enormous market of mobile users, companies will roll out new wireless applications to re-coop their technology investments and increase customer loyalty and switching costs. These applications are collectively called location-based services.

Emerging commercial location-based services fall into one of the following two categories. First, Mobile Resource Management (MRM) applications that include systems for mobile workforce management, automatic vehicle location, fleet management, logistics, and transportation management and support (including air traffic control). These systems use location data combined with route schedules to track and manage service personnel or transportation systems. Call centers and dispatch operators can use these applications to notify customers of accurate arrival times, optimize personnel utilization, handle emergency requests, and adjust for external conditions like weather and traffic. Second, Location-aware Content Delivery services that use location data to tailor the information delivered to the mobile user in order to increase relevancy, for example delivering accurate driving directions, instant coupons to customers nearing a store, or nearest resource information like local restaurants, hospitals, ATM machines, or gas stations. Analyses Ltd. estimates that location based services will generate 18.5B in sales by 2006.

In addition to commercial systems, management of moving objects in location based systems arises in the military, in the context of the digital battlefield. In a military application one would like to ask queries such as "retrieve the helicopters that are scheduled to enter region R within the next 10 minutes".

Location management, i.e. the management of transient location information, is an enabling technology for all these applications. Location management is also a fundamental component of other technologies such as fly-through visualization (the visualized terrain changes continuously with the location of the user), context awareness (location of the user determines the content, format, or timing of information delivered), augmented reality (location of both the viewer and the viewed object determines the type of information delivered to viewer), and cellular communication.

Location management has been studied extensively in the cellular architecture context. The problem is as follows. In order to complete the connection to a cellular user u, the network has to know the cell id of u. Thus the network maintains a database of location records (key, cell-id), and it needs to support two types of operations: (1) Point query when a cellular user needs to be located in order to complete a call or send a message, e.g., find the current location (cell) of moving object with key 707-476-2276, and (2) Point update when a cellular user moves beyond the boundary of its current cell, e.g., update the current location (cell) of moving object with key 707-476-2276. The question addressed in the literature is how to distribute, replicate, and cache the database of location records, such that the two types of operations are executed as efficiently as possible. Related questions are how frequently to update, and how to search the database. Many papers have addressed this question.

However, the location management problem is much broader. The main limitations of the cellular work are that the only relevant operations are point queries and updates that pertain to the current time, and they are only concerned with cell-resolution locations. For the applications we discussed, queries are often set oriented, location of a finer resolution is necessary, queries may pertain to the future or the past, and triggers are often more important than queries. Some examples of queries/triggers are: during the past year, how many times was bus#5 late by more than 10 minutes at some station (past query); send me message when a helicopter is in a given geographic area (trigger); retrieve the trucks that will reach their destination within the next 20 minutes (set oriented future query).

In terms of MRM software development, the current approach is to build a separate, independent location management component for each application. However, this results in significant complexity and duplication of efforts, in the same sense that that data management functionality was duplicated before the development of Database Management Systems. To continue the analogy, we need to develop location management technology that addresses the common requirements, and serves as a development platform in the same sense that DBMS technology extracted concurrency control, recovery, query language and query processing, and serves as a platform for inventory and personnel application development. And this is the objective of our DOMINO project, and our venture-funded startup company called Mobitrac (www.mobitrac.com). In this project we address the main research challenges in building a general purpose location management system.

These include modeling and spatio-temporal operators, uncertainty in location management, location management in a distributed/mobile environment, location and traffic prediction by data mining, indexing, and other issues, particularly, privacy and location fusion.

5. Visual Query Languages, Data Integration and Visualization (Isabel Cruz & colleagues)

Visual query languages take advantage of the user's visual perception to convey information efficiently. The successful implementation of visual query languages will provide database systems with fundamental capabilities not currently available, such as the ability to specify the format in which the retrieved information is displayed.

The visual query language, DOODLE [5], has as its underlying principle that users can *display and query the database with arbitrary pictures* created from simple graphical primitives and constraints. DOODLE uses the technology of deductive query languages for object-oriented databases supporting recursion, being powerful enough to compute the topological numbering of a DAG. The expressive power of DOODLE has been demonstrated in graph drawing.

Graph drawing addresses the problem of automating the construction of geometric representations of graphs and networks. Our graph drawing work has focused on a declarative approach based on constraint satisfaction techniques. Constraints allow users to specify properties that the drawings must satisfy: for example, that a hub in a communication network must occupy a central position, or that the tasks in a decision support graph must be drawn left to right depending on their times of completion. Salient features of this work include the ability to specify the drawing of important classes of graphs and drawings, and that our algorithms are optimal [7].

The DOODLE language is the foundation of two different systems: Delaunay and Delaunay Multimedia (Delaunay^{MM}). Key components of Delaunay are its efficient constraint solver, the interface modules supporting advanced visualization techniques, and the expressiveness and effectiveness principles it incorporates. Our successful implementation demonstrates the feasibility of a powerful visualization system that allows for domain-independent and tailorable visualizations to be specified using a visual query language.

Delaunay^{MM} is an authoring, querying, and visualization framework for multimedia information retrieved from distributed Web repositories. Users compose virtual documents by constructing visual templates that contain both layout information and query specification. To reduce information overload, the user interface supports pre- and post-query refinement capabilities. Extensive usability studies were performed on the user interface [8].

The work on visual query languages has been supported by an NSF CAREER Award since 1996. In 2000, our paper that describes the design and implementation of Delaunay has received the best paper award at the IEEE Visual Languages Symposium [9].

Some of our most recent work has been on Semantic Web issues and on Geographical Information Systems for data integration and for visualization. In particular, we have been doing research on effective visualization techniques for decision support. We extract relevant geospatial information for mission planning from a variety of databases, relate that information, and present it to the users using visual techniques that emphasize the correlation of disparate events across the globe. We establish semantic relationships between heterogeneous information and carry those semantic relationships and metadata to the visualization layer. Likewise, the result of the user's interaction with the data and with the user interface is stored and managed to allow for adaptation to the user and/or to the task.

Our work on data integration for Geographic Information Systems addresses research issues that arise in concrete applications in the context of the State of Wisconsin Land Information System (WLIS). It is our goal to incorporate Web-based search and query techniques in WLIS. The concept of WLIS as a statewide access mechanism for land related data has been actively under development for several years by a working group that includes various levels of government and the University of Wisconsin, with whom we collaborate. However, there are several technical challenges that require extensive state-of-the-art database research to achieve their realization. Our main focus has been on interoperability issues to achieve data integration. In WLIS, data is stored in XML using independently maintained local schemas. However, answers to many queries must span several schemas. Our approach is based on the existence of a central ontology and on declarative transformations between the local schemas and that central ontology. Using our approach, users can seamless query and aggregate semantically related data that is available throughout the state [6].

6. Text Classification with only Positive and Unlabeled Data - *Partially Supervised Learning*

(Bing Liu and colleagues) [11]

Text classification commonly described as follows: Given a set of labeled training documents of *n* classes, the system uses this training set to build a classifier, which is then used to classify new documents into the *n* classes. Although this traditional model is very useful, in practice we also encounter another problem. That is, one has a set of documents of a particular topic or class *P*, and is given a large set *M* of mixed documents that contains documents from class *P* and also other types of documents. One wants to identify the documents from class *P* in *M*, or to classify the documents in *M* into documents from *P* and documents not from *P*. The key feature of this problem is that there is no labeled non-*P* document, which makes traditional techniques inapplicable, as they all need labeled documents of every class. We call this problem *partially supervised classification*. In this work, we showed that this problem can be posed as a constrained optimization problem and proved that under appropriate conditions, solutions to the constrained optimization problem will give good solutions to the partially supervised classification problem. We also proposed a novel technique to solve the problem and demonstrate its effectiveness through extensive experimentation.

7. Handling Model Misfit in Text Categorization (Bing Liu and colleagues) [19]

Text classification is the automated assigning of text documents to pre-defined classes based on their contents. So far, many effective techniques have been proposed. However, most techniques are based on some underlying models and/or assumptions. When the data fits the model well, the classification accuracy will be high. However, when the data does not fit the model well, the classification accuracy can be very low. In this work, we propose to use successive refinements of classification on the training data to correct the model misfit. We apply the technique to improve the classification performance of two simple and efficient text classifiers, the Rocchio classifier and the naïve Bayesian classifier. Extensive experiments on two benchmark document corpuses demonstrate that the proposed technique can improve text classification accuracy of the two techniques dramatically, and also consistently outperforms the popular SVM algorithm.

8. Web Site Comparisons (Bing Liu & colleagues) [13]

The Web is increasingly becoming an important channel for conducting businesses, disseminating information, and communicating with people on a global scale. More and more companies and individuals are publishing their information on the Web. With all this information publicly available, naturally companies and individuals want to find useful information from these Web pages. As an example, companies always want to know what their competitors are doing and what products and services they are offering. Knowing such information, the companies can learn from their competitors and/or design counter measures to improve their own competitiveness. In his work, we proposed a novel visualization technique to help the user find useful information from his/her competitors' Web site easily and quickly. It involves visualizing the comparison of the user's Web site and the competitor's Web site to find similarities and differences between the sites. The visualization is such that with a single glance, the user is able to see the key similarities and differences of the two sites. He/she can then quickly focus on those interesting clusters and pages to browse the details.

9. Classification Using Association Rules (Bing Liu and colleagues) [12]

Existing classification algorithms in machine learning mainly use heuristic/greedy search to find a subset of regularities (e.g., a decision tree or a set of rules) in data for classification. In the past few years, extensive research was done in the database community on learning rules using exhaustive search under the name of association rule mining. The objective there is to find all rules in data that satisfy the user-specified minimum support and minimum confidence. Although the whole set of rules may not be used directly for accurate classification, effective and efficient classifiers can be built using the rules. We proposed the first algorithm that use association rules for classification [12]. We showed that on average it outperform the state-of-the-art existing machine learning techniques, i.e., the decision tree technique, the naïve Bayesian technique, and the covering technique.

10. Interestingness and Rule Query Language (Bing Liu and colleagues)

It is well known that many existing data mining techniques often produce a large number of rules or patterns, which makes it very difficult for manual inspection of the rules to identify those interesting ones. This problem represents a major gap between the results of data mining and the understanding and use of the mining results. In the past few years, we proposed a number of techniques to deal with the problem to reduce the burden of the user in analyzing the discovered rules. These techniques includes: (1) finding unexpected rules by asking the user to give is/her existing knowledge; (2) summarizing the discovered rules so that the user can see the general patterns in the domain easily; (3) organizing rules in such a fashion that the user can browse the rules easily; and (4) designing a rule query language. We have published a number of papers on these topics. Please refer to my Web page http://www.cs.uic.edu/~liub for papers from *IEEE TKDE* (1999), *AAAI* (1996, 2000), and *KDD* (1997-2002).

References

[1]. E. Albuz, E. Kocalar, and A. Khokhar, "Scalable indexing and retrieval system using vector wavelets." *IEEE Transactions on Knowledge and Data Engineering*, May, 2001.

[2]. M. Aslam, A. Khokhar, R. Ansari, and B. De Ballion, "Predominant pitch contour extraction from audio signals,'' To appear in *IEEE International Conference on Multimedia and Expo* (*ICME*), 2002.

[3]. A. Aslandogan A. and C. Yu. "Multiple evidence combination in image retrieval: Diogenes searches for people on the web." *SIGIR-00*, 2000.

[4]. A. Aslandogan A. and C. Yu. "Evaluating strategies and systems for content based indexing of person images on the Web." *ACM Multimedia*, 2000.

[5]. I. F. Cruz. "A visual language for object-oriented databases." *SIGMOD-92*, 1992

[6]. I. F. Cruz and P. W. Calnan. "Object interoperability for geospatial applications: a case study." In I. Cruz, S. Decker, J. Euzenat, and D. McGuinness, editors, *The Emerging Semantic Web*. IOS Press, 2002.

[7]. I. F. Cruz and A. Garg. "Drawing graphs by example efficiently: trees and planar acyclic digraphs." *Graph Drawing '94*, 1995.

[8]. I. F. Cruz and K. M. James. "A user interface for distributed multimedia database querying with mediator supported refinement." *International*

Database Engineering and Applications Symposium (*IDEAS '99*), pages 433-441, 1999.

[9]. I. F. Cruz and P. S. Leveille. “Implementation of a Constraint-based Visualization System.” *IEEE Symposium on Visual Languages* (*VL'00*), 2000.

[10]. G. Li and A. Khokhar, “Content-based indexing and retrieval of audio data using wavelet.’’ *International Conference on Multimedia and Expo* (*ICME*), 2000.

[11]. B. Liu, W. Lee, P. Yu & X. Li. “Partially supervised classification of text documents.” *ICML-2002*, 2002.

[12]. B. Liu, W. Hsu, Y. Ma. “Integrating classification and association rule mining.” *KDD-98,* 1998.

[13]. B. Liu, K. Zhao, and L. Yi. “Visualizing Web site comparisons.” *WWW-2002*, 2002.

[14]. A. P. Sistla, O. Wolfson. “Minimization of communication cost through caching in mobile environments.” *IEEE Transactions on Parallel and Distributed Systems*, 9(4), April 1998, pp 378-390.

[15]. A. P. Sistla, O. Wolfson. “Modeling and Querying moving objects.” *ICDE-97,* 1997.

[16]. A. P. Sistla, C. Yu, C. Liu and K. Liu. “Similarity based retrieval of pictures using indices on spatial relationships.” *VLDB-95*, 1995.

[17]. A. P. Sistla, C. Yu, and R. Vankatasubrahmanian. “Video retrieval.” *ICDE-97*, 1997.

[18]. O. Wolfson. “Moving objects information management: The database challenge.” *Proceedings of the 5th Workshop on Next Generation Information Technologies and Systems* (*NGITS-2002*), Israel, 2002. Springer LNCS 2382, pp. 75-89.

[19]. H. Wu, T. Phang, B. Liu, X. Li. “A refinement approach to handling model misfit in text categorization.” *KDD-2002*, 2002.