Database Research at the University of Illinois at Urbana-Champaign

M. Winslett, K. Chang, A. Doan, J. Han, C. Zhai, Y. Zhou
Department of Computer Science, 1304 W. Springfield Avenue, Urbana, IL 61801

July 11, 2002

The Department of Computer Science at the University of Illinois at Urbana-Champaign (UIUC) has identified the area of *information systems*, broadly construed, as one of three core areas for the department's future directions. In tandem with a mandate from the UIUC College of Engineering for the department to double its number of tenure-track faculty, this focus on information systems has resulted in a significant expansion of the number of faculty in the department in the database area over the past few years. Our roster currently includes Marianne Winslett, who joined the department in 1987; Kevin Chang and Jiawei Han, who joined us in 2000 and 2001, respectively; and AnHai Doan, Chengxiang Zhai, and Yuanyuan Zhou, who joined the department in 2002. In the near future, we plan to round out the information systems group with additional hires of senior faculty.

This short report describes the information systems research activities currently underway in our department. More information on these projects can be found through our web site, `http://drl.cs.uiuc.edu`.

1 Information Integration

Faculty: Kevin Chen-Chuan Chang, AnHai Doan, and Jiawei Han
Funding sources: NSF Career Award and UIUC startup funds

The rapid spread of communication networks has transformed our world into a vast information bazaar, with millions of sources providing data in every imaginable format and mode of interaction. Distributed information systems function as crucial middlemen in this chaotic market, by interacting with data sources, translating, and combining their data to obtain the information requested by users. However, today such systems are still very hard to build and costly to operate. They must be told in tedious detail how to interact with the data sources, and must be adjusted constantly to deal with the changes at the sources. Our research seeks to make such systems much easier to use with far less need for human intervention. The ultimate goal is to achieve the widespread use of online information processing systems that take only hours to be deployed (instead of weeks or months as is the case today), that require only minimal human coaching to rapidly reach and maintain competence, and that continuously improve over time, in terms of both performance and capabilities.

Toward this goal, a part of our research [3, 4] has focused on *schema matching*: finding the semantic correspondences between the schemas of data sources. This problem cuts across virtually all applications of distributed information management (e.g., data integration, knowledge-base construction, e-commerce, and the Semantic Web). Manual schema matching is labor intensive and simply does not scale up to a large number of sources. We developed an automatic solution to schema matching that can learn from past matching activities, that can work efficiently with a variety of data representations (including relational schemas, XML, and ontologies), that can discover both simple one-to-one and complex mappings, that is highly modular and easily customized to a new application domain, and that has been shown to achieve high accuracy in real-world domains. Our current work focuses on developing a theoretical foundation for schema matching, maintaining the correctness of mappings over time, and reasoning with approximate mappings when exact mappings are not available, as is common in practice.

Over the past few years, the Web has "deepened" dramatically—a significant and increasing amount of information is provided only behind the query interfaces of searchable databases. Because current crawlers and search engines cannot effectively query databases, such data remain largely inaccessible to users. In the *MetaQuerier* project, our main objective is to enable seamless and transparent access to queryable databases on the Web. These information sources are typically autonomous and heterogeneous, each with a different schema and native query constraints. We are working to build a "metaquery" system, to help users to find and query these databases in a uniform manner with richly expressive queries, extending our previous work on "static" query translation [1, 2]. To make this scenario possible, the key enabling technology in the *MetaQuerier* project is dynamic ad-hoc information integration: in contrast to a traditional static system, *MetaQuerier* is dynamic (new sources may be added anytime) and essentially requires ad-hoc integration, to dynamically select and bring together different sources to answer a query.

We are also developing solutions to other challenging problems in distributed information processing. Learning source characteristics such as source schema, coverage, query-processing capability, and overlap have been recognized as crucial to large-scale information management. Automatically detecting and adapting systems to changes in source characteristics is important to slash the high cost of system ownership. User feedback is critical to many tasks during system deployment (e.g., schema matching). However, requiring too much user feedback will seriously discourage users from deploying the system in the first place. Hence, we seek to develop techniques to minimize necessary user feedback while maximizing its impact. We combine techniques from several fields—most notably databases and A.I.—to address these problems. We validate our solutions by applying them to the construction of data integration systems on the Web and in specific application domains, such as astronomy and biology.

Finally, we have been working on the development of Web structuring and Web mining technology, including extraction of semantics-based semi-structured data, schema generation, Web page classification, multi-dimensional and multi-layered Web warehouse construction, analysis of Web linkage and Web traversal patterns, and other related Web mining issues.

[1] K. C.-C. Chang and H. Garcia-Molina, Approximate Query Translation Across Heterogeneous Information Sources, Proceedings of the 2000 VLDB Conference, Cairo, Egypt, September 2000.

[2] K. C.-C. Chang and H. Garcia-Molina, Mind Your Vocabulary: Query Mapping Across Heterogeneous Information Sources. Proceedings of the 1999 ACM SIGMOD Conference, Philadelphia, Pa., June 1999.

[3] A. Doan, P. Domingos and A. Halevy, Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach, Proc. of the ACM SIGMOD Conf. on Management of Data (SIGMOD-2001), May 2001.

[4] A. Doan, J. Madhavan, P. Domingos and A. Halevy, Learning to Map between Ontologies on the Semantic Web, Proc. of the World-Wide Web Conf. (WWW-2002), 2002.

2 Security

Faculty: Marianne Winslett, Jiawei Han, Kevin Chen-Chuan Chang
Funding sources: DARPA

2.1 Security in Open Environments

We have been participating in a department-wide cyberterrorism initiative, focusing on the problem of providing security in open environments such as universities, airports, train stations, national monuments, and public buildings. These environments need to be open to access by strangers and must offer strong privacy guarantees, while at the same time detecting potential threats to security and reacting appropriately to them. Flexible means of establishing trust, as described in the following section, is an important aspect of such an environment. So is the ability to mine streams of events coming from sensors and other spatial, temporal, and multimedia data sources, to detect anomalies and outliers that may indicate threats to the environment. This mining must be robust in the face of noise and limited resources, and must respect the strong guarantees of privacy given to individuals who enter the environment. Finally, due to the dynamic nature of the environment (information sources may come and go) and to its privacy-preserving features, reaction to threats will require dynamic integration of information from multiple information sources. To address all these issues, we are applying the results from our data mining, information integration, and trust negotiation projects to the problem of security in open environments.

2.2 Trust Negotiation

Traditional approaches to protecting information relied on the fact that the data lived in a closed system, subject to centralized control. In the freewheeling world of e-commerce and in the dynamic nature of today's business and military relationships, centralized control has been replaced by autonomous systems, temporary partnerships, and interactions between strangers. In these situations, strangers need a means of establishing sufficient trust to feel comfortable carrying out their intended interaction, such as providing access to a confidential document, providing a bidder's card for an on-line auction, disclosing a credit card number, obtaining a student discount on a purchase, proving eligibility for an international adoption, voting in an election, etc. As part of the TrustBuilder project, we are developing an approach to establishing trust between strangers, called *trust negotiation*. Trust negotiation relies on public key infrastructure (PKI) for unforgeable, verifiable digital credentials that can be used to establish trust. With trust negotiation, each resource that a stranger might wish to access (e.g., a document or an on-line information source or service) must be protected by an access control policy that spells out exactly which digital credentials a stranger must disclose in order to obtain access to the resource. With digital credentials and access control policies, security agents can transparently negotiate trust on behalf of strangers, without human intervention.

TrustBuilder is a joint project with colleagues at Brigham Young University, who are developing scalable, reusable implementations of trust negotiation components. Our work at UIUC focuses on theoretical issues associated with trust negotiation, including strategies for negotiating trust, means of assuring individual autonomy during the process of trust negotiation [1], and privacy and security issues associated with trust negotiation.

[1] T. Yu, M. Winslett, and K. E. Seamons, Interoperable strategies in trust negotiation, ACM Conference on Computer and Communications Security, Philadelphia, November 2001.

3 Data Mining

Faculty: Jiawei Han, Kevin Chen-Chuan Chang, AnHai Doan, Chengxiang Zhai, Yuanyuan Zhou
Funding Sources: NSF, Microsoft Research, IBM Faculty Award, and UIUC startup funds

Our data mining projects focus on scalable and effective data mining methods, their system and architectural support, and their applications in stream data analysis [1], integration of data mining with data warehousing [2], Web databases [4], protection of homeland security, text data mining, and bio-data and other scientific data analysis [3]. The work on data mining applications and on system and architectural support for data mining is described in other sections of this document. In addition to that work, we are also pursuing the development of core data mining algorithms, as illustrated by the projects described below.

Multi-dimensional stream data analysis. A fundamental difference in the analysis of stream data, as opposed to relational and warehouse data, is that the stream data is generated in huge volumes (often with very detailed information), flowing in and out dynamically, and changing rapidly. Due to limited memory, disk space and processing power to handle such huge volumes of data, most data streams can only be examined in a single pass. However, stream data applications often require multi-dimensional analysis, with real-time response. We have been working on multi-dimensional on-line mining of unusual patterns in stream data, including stream data cubing, clustering, classification, and comparison of multiple data streams for mining unusual patterns, and multi-dimensional regression analysis of time-series data streams, such as [1].

Scalable methods for mining frequent, sequential and structured patterns. With the successful development of the FPgrowth, H-mine, and PrefixSpan algorithms for mining frequent patterns and sequential patterns, we have been working on constraint-based, scalable methods for mining max and closed sequential patterns, tree patterns, and graph patterns in a noisy environment (such as [5, 6]), as well as their applications in classification, Web structure mining, and bio-medical data analysis, such as [3,4].

Integration of data warehousing and data mining. We have been working on semantic compression of data cubes [2] as well as intelligent mining and exploration of data warehouses.

[1] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang, Multi-Dimensional Regression Analysis of Time-Series Data Streams, VLDB'02, Hong Kong, Aug. 2002.

[2] L. V. S. Lakshmanan, J. Pei, and J. Han, Quotient Cube: How to Summarize the Semantics of a Data Cube, VLDB'02, Hong Kong, Aug. 2002.

[3] J. Han, R. B. Altman, V. Kumar, H. Mannila and D. Pregibon, Emerging Scientific Applications in Data Mining, Communications of the ACM, Aug. 2002.

[4] H. Yu, J. Han, and K. C.-C. Chang, PEBL: Positive Example Based Learning for Web Page Classification Using SVM, KDD'02, Edmonton, Canada, July 2002.

[5] J. Liu, Y. Pan, K. Wang, and J. Han, Mining Frequent Item Sets by Opportunistic Projection, KDD'02, Edmonton, Canada, July 2002.

[6] J. Yang, P. S. Yu, W. Wang, and J. Han, Mining Long Sequential Patterns in a Noisy Environment, SIGMOD'02, Madison, WI, June 2002.

4 System and Architecture Support for Databases, Data Mining, and Storage Systems

Faculty: Yuanyuan Zhou
Funding Sources: UIUC startup funds

Most database servers run on general purpose operating systems and hardware architectures. The substantial synergy between databases and underlying systems indicates that it is important to consider the interaction between these two in order to provide high performance, scalability, reliability and availability in databases. Motivated by this, we are conducting interdisciplinary research between databases and systems to address problems in database systems such as I/O problems and availability problems.

Storage I/O has been one of the major performance bottlenecks in databases. This bottleneck will become more and more serious due to the widening processor-disk performance gap. To alleviate this bottleneck, we have investigated techniques to improve I/O rate and reduce I/O overhead.

Improve I/O rate. We have proposed an effective way to manage the storage buffer cache in a multi-tier storage infrastructure with three levels of caches: client cache, database server cache and storage server cache [1]. Our research studied the database I/O patterns with several commercial databases and workloads and derived a multi-queue algorithm for storage server caching. Our experimental results with simulation and implementation show that this algorithm is quite effective.

Reduce I/O overhead. I/O related processor overhead is another important factor that limits database performance because it can take away useful CPU cycles from real transaction processing. To minimize I/O related processor overhead, we have proposed an I/O architecture that allows databases to access back-end storage systems directly from the user level, bypassing the operating system [2]. This method can significantly reduce I/O related overhead and ultimately improve database transaction rates.

Many recent efforts have focused on building fault-tolerant database systems using clusters of commodity components so when a node fails another node in the same cluster can take over. However, the failover time in most existing fault-tolerant clusters is unacceptable for many mission-critical database applications. The long failover time is mainly caused by periodically checkpointing the memory state into external shared storage. We have proposed a novel way of using virtual memory-mapped communication (VMMC) to reduce the failover time on clusters [3]. In this model, a database's virtual address space can be efficiently mirrored onto remote memory automatically. When a machine fails, the database can restart from the most recent checkpoint on the failover node, with minimal memory copying and disk I/O overheads.

In the future, our research will continue to focus on system and architecture support for databases and data mining applications. Currently we are exploring a more cooperative approach between databases and systems. For example, we are extending the interface of an underlying system to globally manage some resources such as the multi-tier buffer cache hierarchy. We are also investigating architectural extensions to better support database query processing. Another direction of our interdisciplinary research is to study the effects and issues of databases on modern architectures such as CC-NUMA machines or clusters of SMPs. On a different but related spectrum, we are also looking at how to apply some techniques from database and data mining in systems.

[1] Y. Zhou, J. F. Philbin and K. Li, The Multi-Queue Replacement Algorithm for Second Level Buffer Caches, USENIX Technical Conference, June 2001.

[2] Y. Zhou, A. Bilas, S. Jagannathan, C. Dubnicki, J. F. Philbin and K. Li, Experiences with VI Communication for Database Storage, 29th International Symposium on Computer Architecture (ISCA), May 2002.

[3] Y. Zhou, P. M. Chen and K. Li, Fast Cluster Failover Using Virtual Memory-Mapped Communication, 13th ACM International Conference on Supercomputing, June, 1999.

5 Text Retrieval and Mining

Faculty: Chengxiang Zhai
Funding sources: UIUC startup funds

With the dramatic increase in online information in recent years, management of textual information is becoming increasingly important. Our research is driven by the following two major challenges in managing *large amounts* of text:

- *Text retrieval.* How do we find information that satisfies a user's information needs?

- *Text mining.* How do we exploit a large amount of text to discover any meaningful global patterns or regularities?

Unlike the *structured* information managed by a traditional database system, textual information is *unstructured*, and carries contents that are generally vague and ambiguous. Thus, both text retrieval and text mining require the capability to handle uncertainty in the meaning or content of a piece of text. Statistical language models (i.e., probabilistic models of text) are powerful tools for quantifying and reasoning with such uncertainty. We emphasize the use of language models in all text management tasks, and have been exploring many different language models for improving text retrieval.

We have proposed a general probabilistic risk minimization framework for text retrieval based on Bayesian decision theory, which unifies several existing retrieval models as well as facilitates the development of new principled retrieval approaches based on statistical language models [1]. We have been exploring several interesting special cases of this framework. In the case of a two-stage language model, we demonstrated that with appropriate smoothing of language models and with statistical estimation techniques, we could achieve excellent retrieval performance without any ad hoc parameter tuning [2]; this is in contrast with traditional retrieval models, which all rely heavily on ad hoc parameter tuning to achieve satisfactory retrieval performance. We are currently exploring the potential of the risk minimization retrieval framework to go beyond the traditional notion of topical relevance, and developing language modeling methods that can rank documents in terms of both relevance and sub-topic diversity.

The lack of learning ability and personalization in existing retrieval systems creates a "ceiling" for their retrieval performance. For example, a web search engine typically "sees" a user only through the text query entered by the user, and generates search results based solely on the very few words in the query. However, the same query may be entered by different users with quite different information needs. For example, the acronym "CD" in the query "CD buying tips" could mean either "Compact Disc" or "Certificate of Deposit"; thus a perfect retrieval performance on this query is just inherently impossible without knowing the intended meaning of "CD". However, resolving such ambiguity is often possible if we exploit other information of the user. (Imagine if we know that the user's previous query was "greatest hits of the year".) Thus, to break this performance ceiling, the system must learn from its interactions with the user and see the user as more than a single query. Some feedback information about whether a user likes a particular document is often very useful for improving performance. We have proposed and studied algorithms for learning from such feedback information in an online information filtering system [3]. We are currently studying the problem of learning and personalization formally using the risk minimization framework described above, and intend to develop an advanced web information management system that learns over time to improve its capability to help a user satisfy both "long-standing" and "short-term" information needs.

Both text mining and structured data mining have been studied extensively, but independently. An integration of the two approaches may potentially lead to more powerful mining techniques. We are most interested in exploring such an integration in the bioinformatics domain, with an emphasis on biological sequence mining. Our main strategy is to extract structural information from biological literature, which is then used to enrich the existing sequence data, providing more evidence for sequence mining. We are currently exploring the possibility of predicting functions of unknown amino acid patterns by mining the Gene Ontology protein sequence annotations as well as the related biological literature.

[1] J. Lafferty and C. Zhai, Document language models, query models, and risk minimization for information retrieval, Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, 2001.

[2] C. Zhai and J. Lafferty, Two-Stage Language Models for Information Retrieval, Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[3] C. Zhai, P. Jansen, and D. A. Evans, Exploration of a heuristic approach to threshold learning in adaptive filtering, Proceedings of the 23th ACM SIGIR Conference on Research and Development in Information Retrieval, 2000.

6 Database Support for High-Performance Computing

Faculty: Marianne Winslett

Funding Sources: Department of Energy, NSF

Scientists' traditional theoretical and empirical investigative paradigms have recently been supplemented with a third paradigm, due to the ability to simulate physical phenomena on high-performance massively parallel computers. The fruits of these simulations include, for example, everyday weather prediction, global climate change prediction, simulation of nuclear explosions (replacing empirical testing), evaluation of new airplane wing designs, and new efforts to understand protein folding. At run time, these simulations need high-performance data reorganization and movement facilities—from a platform that may have thousands of processors, to the platform's local disks, to a shared file system, and to remote storage and visualization facilities. In the Panda project, our goal is to provide these data movement and reorganization facilities in an easy-to-use manner that provides portable and scalable high performance. Due to the idiosyncracies of common parallel platforms, we believe that in order to be easy to use and to provide portable high performance, a parallel I/O system must also be self-tuning [1].

Though both can be viewed as giant optimization problems, the I/O needs of high-performance scientific simulation codes have little in common with the needs of traditional database applications. However, the flavor of the problem and of the solution approaches that we have developed for simulations has much in common with the flavor of the problems found in enterprise storage management. Further, we have found that a database perspective is invaluable in devising principled approaches to simulations' I/O needs.

The Panda project began in the early 1990s, and has reached the stage where technology transfer is as important as the production of new research results. Our current research emphases are new approaches to buffering, alternative approaches for making parallel I/O systems self-tuning, end-to-end data movement facilities that transport data across the internet during a simulation run, and I/O-aware compilation. We are very involved with technology transfer of our previous research results to HDF and ROMIO, which are data management and parallel I/O libraries that are extremely popular with scientists. We also provide the I/O facilities for the flagship code of the Center for the Simulation of Advanced Rockets at UIUC, and participate in the Center for Scalable Parallel Programming Models, headquartered at Argonne National Laboratory.

[1] Y. Chen, M. Winslett, S. Kuo, and Y. Cho, Automatic parallel I/O performance optimization in Panda, IEEE Transactions on Software Engineering, April 2000.

7 Top-k Queries

Faculty: Kevin Chen-Chuan Chang
Funding sources: NSF Career Award and UIUC startup funds

The *MPro* project addresses the problem of evaluating ranked top-k queries with *expensive predicates*. As all major DBMSs now support expensive user-defined predicates for Boolean queries, we believe such support for ranked queries will be essential. In particular, our notion of expensive predicates provides a unified abstraction for (1) *user-defined functions* (for modeling user-specific concepts or preference), (2) *external predicates* (for integrating autonomous sources), and *fuzzy joins* (for dynamically associating multiple relations). These predicates, being dynamically defined or externally accessed, cannot rely on index mechanisms to provide zero-time sorted output, and must instead require per-object probes to evaluate. To minimize expensive probes, we have studied the formal principle of "necessary probes," and developed Algorithm *MPro*, which is provably optimal with minimal probe cost [1]. We will continue to extend the framework for handling preference-based search in large databases, by using top-k queries with dynamically-defined preference criteria.

[1] K. C.-C. Chang and S.-W. Hwang, Minimal Probing: Supporting Expensive Predicates for Top-k Queries, Proceedings of the 2002 ACM SIGMOD Conference, Madison, Wisconsin, June 2002.