Research Statement – Ruoming Jin

My research interests are in data mining, databases, and high performance computing, more recently, as they apply to problems in bio-informatics. The focus of my research to date has been on developing techniques to provide (near) interactive-speed evaluation of ad-hoc data mining and OLAP queries. In particular, I have studied the following issues: 1) how to use parallel computers, including both distributed-memory and shared-memory machines, to speedup data mining and OLAP processing; 2) how to use samples to approximate OLAP queries with accurate error-bound estimation (confidence interval), and how to derive accurate mining results from sampling results; 3) how to decompose a complex mining query into basic mining operators, and generate efficient query plans to evaluate them; and 4) how multiple complex mining queries can be optimized together, and how a knowledgeable cache can be used to further speedup such evaluation. In the following, I summarize each of these efforts in turn, and give an outline of my plan for future.

1 Scalable Data Mining and Parallel OLAP

With the advances of computer technology, parallel computers, especially clusters and relatively large SMP machines, have become very cost-effective to speedup the mining and online analytic processing of large datasets. However, implementing a scalable data mining algorithm on a parallel machine can be very time-consuming and challenging. In order to deal with this problem, I have developed a middleware, called FREERIDE (Framework for Rapid Implementation of Data mining Engines), for efficiently parallelizing data mining algorithms. I have also designed new parallel algorithms for efficient data cube construction.

FREERIDE for Scalable Data Mining [5, 6, 7, 8, 16] FREERIDE targets both distributed-memory and shared-memory parallelization of mining algorithms, and supports efficiently processing large out-of-core datasets by asynchronous I/O. It provides a simple interface for implementing a scalable mining algorithm from its sequential version. A variety of scalable mining algorithms, including Apriori, FP-tree, Decision Tree Construction, K-means, and EM clustering, have been implemented over FREERIDE.

One of the key challenges in this middleware is to perform efficient shared-memory parallelization of a common canonical loop, shared by many mining algorithms. However, traditional techniques, such as static partition and the inspector/executor method, can not be applicable. To meet this challenge, I have proposed a set of new techniques that use optimized memory layout to avoid race condition and reduce the synchronization costs. In particular, two new techniques, which interleave the reduction elements and lock in each cache line, are very efficient in terms of memory cost, locking overhead, and cache misses. I have further developed a detailed analytic model, considering the locking cost, L1/L2/coherence cache misses, TLB misses, and false sharing, to compare the performance of proposed techniques. Experiments showed the difference between the measured performance and the predicted performance is within 20% in most of the cases. This prediction work was published in SIGMETRICS’02.

Recently, FREERIDE has been used by my colleagues, Leo Glimcher and Xuan Zhang, to parallelize scientific data mining applications, including Vortex Detection and Defect Detection. Their research works have been published in IPDPS'04 and IPDPS'05 [1, 3]. Another Ph.D. student in our research group, Xiaogang Li, has studied language and compiler support for data mining applications. FREERIDE has served as a compiler target in his research [18].

Communication and Memory Efficient Parallel Data Cube Construction [14, 15, 17, 20] The construction of data a cube is very compute and data intensive task. Our initial experiments [20] showed promising results in accelerating this process by distributed-memory parallelization. However, we also realized that many factors, such as communication volume/frequency, memory and I/O cost, can affect the speedup. But previous research did not address these issues adequately. Therefore, I investigated the optimal conditions to build MOLAP in terms of communication volume, memory, and I/O cost. In particular, I derived the closed formulas for communication volume, and found a linear deterministic procedure to minimize the communication volume. Further, I developed a tiling-based approach to handle the situations where the datasets or the data
cube cannot be held in the main memory. The (near) optimized tiling condition is provided to reduce the I/O cost.

The experiments further validate the theoretical results. Besides being published in several parallel computing conferences, a comprehensive paper describing this research has recently been accepted in IEEE transaction on Parallel and Distributed System.

2 Sampling Techniques for OLAP and Data Mining

Even with the use of the state-of-art mining algorithms and OLAP evaluation techniques, many data mining and OLAP queries can still require hours or days to complete execution. In order to provide interactive response times, many approximation techniques, such as histograms and wavelets, have been developed. However, though experimentally such techniques showed good approximation, little information is provided about the accuracy of such estimation, i.e. no meaningful confidence bound is associated with this kind of estimation. To address these issues, I studied the following two sampling-based approaches. The first one incorporate additional summary with random samples to improve the accuracy of the confidence interval for the OLAP queries, such as Sum and Count. The second approach considers deriving accurate mining results from the approximate results based on sampling.

Approximate OLAP Queries by random sampling and pre-aggregation [11] This work is done in collaboration with Prof. Chris Germaine in University of Florida. He first proposed the idea of utilizing additional summary information with a random sample to improve estimation. However, his original work does not provide confidence bounds. To address this limitation, I proposed a new ensemble-based approach, which utilizes the pre-aggregation of the low dimensional facts. The basic idea is to linearly combine multiple unbiased estimators, derived through the pre-aggregation results. This new approach can provide very accurate confidence intervals associated its estimation by optimizing the linear parameters for the estimation (utilizing the Lagrange’s Multiplier). The experiments on real datasets also validate the theoretical results—the new approach significantly reduces the width of confidence interval and simultaneously boosts the accuracy of confidence level. In particular, because the summary information usually costs only a couple of kilobytes, and can be computed easily by a single pass of the datasets, this approach can be potentially applicable in the streaming environment as well.

Efficient and Accurate Data Mining based on Sampling Techniques [9, 12] Sampling is a popular approach used in data mining for data reduction. However, the results in the sampling can be misleading because of the perturbation of the samples. Further, due to the complexity of mining algorithms, reasonable bounds can be hardly derived to estimate the accuracy of the final mining results. In many situations, users still prefer to have the final accurate mining results based on the complete datasets. But the complete datasets are usually very large, and are disk-resident. In the meantime, many mining algorithms are data intensive. From this perspective, I have considered the following question: “Can the sampling results be helpful to speedup the process to derive the final accurate results from the complete datasets?”

The basic idea of the approach I have taken is as follows. First, we mine approximate results on a sampled dataset. Utilizing Hoeffding Inequality or other estimation techniques, we can deduce the accuracy of the key computation in the mining algorithms. Then, we perform one pass on the complete dataset to collect the necessary information (sufficient statistics) in order to derive the final accurate results. Finally, utilizing these sufficient statistics, we can derive the correct final results and verify the correctness of our previous estimation. The incorrect estimation may trigger another pass of above procedure, but by as we use the statistical inference, it rarely happens.

Unlike the previous approaches to reduce the number of passes on the datasets, this approach is guaranteed to find the exact results. Based on this idea, I have designed two new mining algorithms, which are FEKMK (Fast and Exact K-means Clustering) for clustering disk-resident datasets and SPIES (Statistical Pruning Intervals for Enhanced Scalability) for parallel decision tree construction.

3 Mining Complex Patterns

In many real applications, especially in the biological domain, users need to find relatively complex patterns instead of the simple frequent patterns. Mining simple patterns is referred as finding all (or a subset, i.e.
constraint mining) of the frequent patterns, such as itemsets, sequences, subtrees, subgraphs, from a single dataset for a given support level. Fast response to the queries on the complex pattern mining can be very crucial to speedup the entire knowledge discovering process. In this study, I focus on efficiently discovering two types of complex patterns. The first one is to mine “interesting” patterns from multiple datasets. The second one is to mine frequent “topological” patterns from the structure datasets, such as graph or tree datasets.

Mining Multiple Datasets [4] In many situations, users need to find interesting patterns across multiple datasets. For example, the following queries are particularly useful and important for a biologist: “find the subsequences frequently occurring in both human and chicken protein primary structures” or “find the frequent motifs shared by human and mouse but infrequent for chicken”. I found the key issue to efficiently mine such complex patterns is to utilize the intermediate mining results. For instance, if a biologist first discovered the frequent patterns in the human gene, instead of independently discovering the frequent patterns in the chicken gene, he/she should use the previous mining result to guide his/her search for the frequent patterns in chicken gene. I formalized this operation as a basic mining operator, the constraint mining operator. It can find the frequent patterns in a given set of patterns on a single dataset. Based on such observations, I formulated a set of mining operators to mine such complex patterns across multiple datasets.

However, many evaluation plans can be used for a given mining task. One of the challenges is how to enumerate such plans systematically. I developed a novel tabular-based approach to serve this purpose. Further, assuming the cost for each mining operator is known, I proved the problem to find optimal evaluation plan is NP-hard. Therefore, I proposed several heuristic algorithms for finding the efficient query plan. My algorithms on real and synthetic datasets showed a performance gain up to an order of magnitude for mining such complex patterns. Finally, I also studied how to extend SQL and relational algebra to express this kind of query, and demonstrated the potential applicability of this work in the real relational database systems.

Mining Frequent Topological Patterns from Graph Datasets [2] This research is motivated by the need to discover similar structures shared by many graphs. For example, in the secondary structure of a group of proteins, the particular shapes formed by α-helix and β-sheet in the 3-dimension space can be a very important feature to understand their functionalities. However, the current graph mining tools target connected subgraphs which most likely capture the common fragments, not the common structures. Inspired from a well-established mathematical concept: topological minor, which can compress a large graph and still preserve the essential structure of the graph, I defined a two-level framework to discover frequently occurring structures in the graph datasets. The first level of this framework allows users focusing on the frequent topological structures shared by graphs; the second level allows users to drill-down to find the similar subgraphs containing the topological structures. This is an on-going effort. I have developed an efficient algorithm to complete the first level of this framework, i.e. finding frequent topological patterns. With the help from my colleagues in bio-informatics, I am currently investigating the applicability of such patterns in the biology datasets.

4 Multiple Query Optimization and Caching for Data Mining

As the amount of data available for analysis in both scientific and commercial domains is increasing dramatically, the volume of data mining tasks imposed on these datasets also grows quickly. We envision that the scenarios in databases and data warehouses will happen in data mining: A user may analyze one or more datasets by issuing a sequence of related complex mining queries, and several users may be analyzing a set of datasets concurrently, and may issue related complex queries. To mine in such scenarios, we need to not only evaluate each single mining query efficiently, but also need to optimize multiple queries simultaneously. Furthermore, we need to be able to utilize mining results from past queries in a systematic fashion.

In this study [13], I designed a system architecture to deal with such a query intensive environment. In particular, I proposed new algorithms to perform multiple-query optimization for frequent pattern mining queries which involve multiple datasets. I also designed a knowledgeable cache which can store the past query results from queries, and enable the use of these results to further optimize multiple queries. Together with another Ph.D. student, Kaushik Sinhak, we have built such a system and evaluated it with both real and synthetic datasets. Our experimental results have demonstrated a speedup of up to a factor of 9.
5 Future Research Agenda

A number of interesting issues are raised in my current research and I am planning to address these issues in the future work. Specifically, I will list both the short term goals and the long term goals.

5.1 Short Term Goals

The short term goals are to apply or extend the techniques I developed in the new sub-fields, such as mining data streams and sensor networks, understanding their limitations as well as their relations with other techniques, and studying them in application domains, such as biology, networking, and XML.

**Topological Structures in Graph Mining**  The concept of topological structures has shown many potential advantages to analyze graphs. So far, I have developed an efficient algorithm to discover frequent topological structures in a graph dataset. In the future work, I plan to further investigate both theoretical and application aspects of this concept. Furthermore, this concept naturally brings the question of comparing the difference between two graphs. How can we formalize such differences and define a new graph edit distance? What is the advantage of this method comparing with the traditional maximal common graph approach? In addition, how frequent topological structures can be used to classify graph datasets? Finally, besides working on biology datasets, I would like to look at other domains, including social networks and networking topology, where topological structures arise naturally.

**A Mining System for Biological Data in the Grid**  Huge amounts of biological data are distributed in the Internet. Accessing and analyzing these data have become common for a number of biologists, statisticians, bio-informatics researchers, and others around the world. However, the pre-processing and mining tasks, such as BLAST and Frequent pattern mining, can be tedious and often take a very long time. Especially, significant amounts of overlap or repetition exist among these tasks issued on these datasets. Therefore, in order to speedup the mining process, I believe we need to utilize the mining results from other mining queries. In [4, 13], I have studied how such optimization can be done for frequent pattern mining on datasets stored in a single machine. The challenge for biological datasets is that they are distributed in the Grid, and mining queries can be issued over the grid as well. I plan to build a mining system on the grid to automatically cache, locate, and finally utilize the (intermediate) mining results from other queries.

**Mining Correlation Patterns and Its Application in Database Selectivity Estimation**  In statistical modeling research for database optimization, such as selectivity estimation, we often rely on assumptions about uniformity of datasets and independence of attributes. Without considering the correlation factor, these assumptions are likely to be inappropriate for real datasets, and consequently, the estimation accuracy can become poor. Such observations have been confirmed in two recent studies [11, 19] I have been involved in. To address these issues, my short term goal will focus on studying the correlation factor in structured data, such as trees (XML) and graphs (Biology Domains). In particular, from the data mining perspective, I will study how to formalize the correlation patterns and mine them efficiently (such patterns can be very useful to understand the key properties of the data), from the database perspective, I will study how to efficiently test the uniformity and independence assumptions for a given real dataset, and utilize the correlation patterns to improve the estimation accuracy if the assumption does not hold.

**New Techniques for Mining Data Streams, Distributed Grids, and Sensor Network**  Sampling and building sufficient statistics are two popular approaches for data reduction, and are the keys for mining in scenarios with resource constraints, such as mining data streams. My approach on scalable mining algorithms is a combination of the previous approaches and involving using results from samples to collect the sufficient statistics for mining the complete datasets. A natural question is how my approach can be used in the resource constraint environments. Some initial results have shown potential applicability. In [10] and [12], I demonstrated its applications in decision tree construction on streaming data, and k-means clustering in distributed grid, respectively. I plan to further investigate its applicability in mining large scale sensor networks.
5.2 Long Term Goals

Besides developing efficient mining algorithms and new techniques for estimation problems in database system and OLAP, my research in the long run will focus on the following three aspects: 1) theoretical foundations of data mining; 2) fundamental issues in building a Knowledge and Data Mining System (KDDMS); and 3) application driven data mining.

In the first aspect, I am interested in understanding the underlying connections among data mining, statistics, and machine learning. Especially, I would like to consider a generalized framework which can uniformly treat clustering and classification problems. My long-term goal in the second aspect is to build a KDDMS system, which can support the entire data mining process in an interactive manner. Some fundamental issues include the appropriate language and algebra for such a system, and how to construct a small set of basic mining operators to be able to express/compose more complex mining tasks. Finally, I would like to motivate my research from the real world problems, especially, in the biology domain and the fields of computer science itself, such as computer architecture, software engineering, and Internet.

References


[2] Ruoming Jin et al. Mining Frequent Topological Patterns in Graph Datasets. in preparation.


