The DIMACS/IBM Workshop on Data Mining in the Internet Age, May 1-2, 2000

Organizers
Rakesh Agrawal, IBM
Joan Feigenbaum, AT&T Labs
Prabhakar Raghavan, Verity Systems
Jeff Ullman, Stanford University

Author
Chen Li
Department of Computer Science
Stanford University, CA 94305
chenli@db.stanford.edu

1 Overview

The DIMACS/IBM Workshop on Data Mining in the Internet Age was held May 1-2, 2000, at IBM Almaden Research Center, San Jose, California. The objective of this workshop was to bring together pioneering researchers in the area of data mining and provide a forum for fundamental advances in data mining. Experts from various backgrounds – databases, statistics, algorithms, bioinformatics, and machine learning – gave 16 high-quality presentations to report the latest progress in data mining in various domains. The following is an overview of the 16 presentations:

Category	Presentations
The Web	*Google and the Importance of Search,* Sergey Brin, Google.com
	Combining Labeled and Unlabeled Data for Web Mining, Tom Mitchell, CMU and Whizbang! Labs
	Musings on Extraction of Structure from the Web, Rajeev Motwani, Stanford University
Genome Mining	*What have I learned at Surromed?* Shalom Tsur, Surromed
	Mining Biological Databases: From Gene Finding to Drug Discovery, Simon Kasif, Compaq Cambridge Research Laboratory and MIT Genome Center
	Association Rules, Boosting, and Genome Mining, Shinichi Morishita, University of Tokyo
Association Rules	*Cubegrades: Generalizing Association Rules*, Tomasz Imielinski, Rutgers University
	Mining Frequent Patterns without Candidate Generators, Jiawei Han, Simon Fraser University
E-commerce	*Mining E-Commerce Data: Challenges and Stories from the Trenches,* Ronny Kohavi, Blue Martini
	Integration of Data Mining and Relational Databases, Surajit Chaudhuri, Microsoft Research
Statistics	*Predictive Data Mining with Multiple Additive Regression Trees* Jerry Friedman, Stanford University
	Combining Combinatorial and Probabilistic Methods in Datamining, Heikki Mannila, Nokia and Helsinki University
Privacy	*Privacy-preserving Datamining,* Ramakrishnan Srikant, IBM Almaden Research Center
Astronomy	*Massive Data Sets in Astronomy,* Michael Vogeley, Drexel University
XML	*Using Datamining for XML Data Storage and Compression,* Dan Suciu, AT&T
Clustering	*Classification with Pairwise Relationships: Metric Labeling and Markov Random Fields,* Jon Kleinberg, Cornell University

2 Open Questions and New Directions

The following are some open questions and important new directions in the field covered by the workshop.

How to extract information from the Web?

Two approaches to extracting information from the Web were discussed in the workshop. The first one, presented by Tom Mitchell, uses machine learning techniques. It classifies web pages in different domains by labeling the pages. It then extracts data from the labeled pages and stores the data in a relational database. The second one, proposed by Sergey Brin and presented by Rajeev Motwani, extracts data that matches certain patterns, which are also generated from some seed data. It iteratively computes data/patterns from the web, until the final results stabilize.

The key observation of the two approaches is that there is a lot of information redundancy on the web, and people tend to use the same language to describe the information in certain domains, e.g., books, sports, etc. These two approaches start from some seed data (training data), and iteratively learn the patterns of the data, and then use the patterns to extract more data from the web. Both approaches show good accuracies of their experimental results. The following problems are still open:

1. In some applications, a very high accuracy of extracted data is desirable. For instance, comparison shopping companies need the detailed information about products such as prices, availabilities, and models, and they cannot tolerate even a small inaccuracy. The highest accuracy of the two proposed approaches is about 95%. How to modify their algorithms to get a higher accuracy?
2. The second approach scans the web pages multiple times until the final results stabilize, while each scan is expensive. How to modify its algorithm to make the computation converge faster?
3. The second approach works for only certain domains. For instance, starting from some baseball teams it can find many baseball teams, football teams, basketball teams, and hockey teams, but not all sports teams. One reason is that people use the same language to talk about these teams, but not for other sports teams in general. For example, people use different ways to talk about tennis teams and football teams. It would be interesting to analyze how the computation converges in different domains.
4. The algorithm in the second approach needs some seed data to generate patterns. What if the seed data has ambiguity, such as "Giants" can be either San Francisco Giants (a baseball team) or New York Giants (a football team)? Can the algorithm still get good results?

How to mine e-commerce data?

There are many reasons that e-commerce data is suitable for data mining: (1) Clickstreams at many web sites provide large amounts of information for large data warehouses; (2) The data is collected electronically at the web sites, and it is clean without any legacy transformations; (3) If designed correctly, these web sites can assign many attributes to the content on their web pages, such as customers, products, and purchases.

However, most companies do not have the expertise to build data mining tools. In the workshop, speakers from two companies, Blue Martini and Microsoft, presented their systems that allow analysts and merchandisers to uncover relationships among products and relationships between products and markets. These systems share the same fundamental idea: to provide a system with the necessary basic functionality to do data mining. A data-mining manager can use the systems to find association rules. However, the systems are different in terms of the role a data-mining manager plays. The Blue-Martini data-mining system accepts explicit queries/investigations from the manager and operates on a database to uncover rules. In other words, data mining is done "outside" the database, and the system provides the basic tools. The Microsoft system "OLAP Services 2000" can iteratively mine training data sets to learn rules, and then apply the rules in large data set to obtain more predications. In other words, data mining can be done "inside" the database.

In general, how to help companies do data mining easily and efficiently? What functionalities should we supply to them? How to allow them to generate comprehensible models? How to allow users to specify the desirable algorithms to process their data? In addition, dates and times are very important attributes and they appear very frequently in e-commerce data. How to find simple but general rules to mine these attributes?

How to mine data without candidate generation?

Jiawei Han proposed a framework to find frequent patterns without candidate generation. The framework has two key ideas: (1) database compression using a structure called FP-tree; (2) database partition to use divide-and-conquer algorithms. It is known that existing algorithms are not efficient when there are too many frequent item sets with more than 3 items. Experiments show that the new framework has good efficiency in this case. Open problems include in what cases the framework can beat other approaches, how to implement the framework using SQL, and how to use the framework to mine other patterns.

How to mine data to incorporate privacy concerns?

Since the primary task in data mining is to develop models about aggregated data, how can we develop accurate models without access to precise information in individual data records? Rakesh Agrawal and Ramakrishnan Srikant showed how to do data mining while incorporating privacy concerns. There are a variety of research issues in privacy-preserving data mining.

How to mine structures from XML data?

XML is a markup language for documents containing structured information. It has been widely accepted as a standard for representing data on the Web. To store XML data in a relational database efficiently, we need to extract structures from the data to decide the schema of the database. How to find the structures using data mining techniques?

How to mine data in other applications?

Data mining can be used in various domains, such as gene finding in biological research, galaxy finding in astronomical research, and etc. How to apply data-mining techniques in these domains to solve domain-specific problems is still a challenging task.

3 Summaries of Talks (Speaker-Prepared)

Google and the Importance of Search
Sergey Brin, Google.com

Google is a large scale hypertextual web search engine. It uses data mining technologies to produce far more relevant search results than conventional search engines. Since its founding, Google has scaled to over 1 billion documents and 30 million searches per day. This has created a number of software and hardware challenges.

Combining Labeled and Unlabeled Data for Web Mining
Tom Mitchell, CMU and Whizbang! Labs

Most models of supervised learning consider only labeled training examples, and ignore the potential role of unlabeled data. This talk considers the question "when is it possible to use unlabeled data to increase accuracy in supervised learning?" This question was initially motivated by our research on algorithms for learning to classify web pages (e.g., as student home pages, faculty pages, etc.). The question of how to use unlabeled data is especially interesting for web page classification, given the hundreds of millions of easily accessible web pages. We present an algorithm and experimental results demonstrating that unlabeled data can significantly improve accuracy when learning to automatically classify web pages. We then identify a precise problem structure that is sufficient to assure unlabeled data will improve learning accuracy. Interestingly, the essential problem structure is found in many natural learning problems faced by humans, such as learning a semantic lexicon over noun phrases in natural language, and learning to recognize objects from multiple sensor inputs. These results suggest that our understanding of human and animal learning might also be improved by considering the potential role of unlabeled data in learning.

Musings on Extraction of Structure from the Web
Rajeev Motwani, Stanford University

One of the major challenges in dealing with the web is the unstructured or semistructured nature of the data. There are major benefits in extracting the structure implicit in the web or extracting a subset of the data that can be structured easily. I will give a personal view of some attempts in this direction. The focus will be on the problems and proposals for attacking these problems.

What Have I Learned at Surromed?
Shalom Tsur, Surromed

Bioinformatics is a modern term used to denote the computational aspects of biology. Traditionally, the term is used to denote computations involving genomic data DNA sequence data and the concepts derived from these such as genes or gene types. With the completion of the sequencing of the human genome, the field rapidly moves beyond the genomic domain into the expression products of genes such as proteins and in general, all of the derived information that eventually will lead to a complete knowledge framework explaining human development and disease. The pharmaceutical industry is among the prominent consumers of this knowledge, which it hopes, will be useful for modern drug development.

The presentation will focus on the question: to what extent can we apply the database methods that traditionally have been applied in the past, in a range of different domain data, to this new and rapidly evolving domain? in particular, the integration of diverse biological information sources and data mining. The author will provide an account of his own experiences in this respect, based on his involvement at SurroMed, Inc.

Mining Biological Databases: From Gene Finding to Gene Discovery
Simon Kasif, Compaq Cambridge Research Laboratory and MIT Genome Center

The current mode of biological exploration is changing from a single experiment where an individual investigator is exploring a single conjecture to hypothesis-free generation of data using revolutionary genomic technologies and new computatonal representations and algorithms for an "in-silico" exploration of the biology of living organisms.

Computational genomic data analysis has already inspired changes in theories of evolution in low-order organisms (prokaryotes); identified and mapped thousands of genes; provided order of magnitude improvements in the cost and speed of experiments; helped progress towards identifying minimal living organisms and suggested new methodologies for accurate diagnosis of subtle forms of cancer by clustering gene expression profiles.

A recurring theme in this research is the need to build perhaps initially naive models of biological systems that allow researchers to extract meaningful information from data. Without exception all work on computational gene finding makes use of assumptions about the organization of the genome and postulates basic gene models that are used as templates to extract putative genes. This paradigm has generated thousands of gene predictions that serve as potential targets for treating disease or other important bio-medical applications.

In this talk we review three techniques for gene identification in biological sequences, and discuss the application of learning systems and computer science algorithms to this scientific challenge. The three main techniques are based on gene identification by matching to biological databases, comparative cross-genomic analysis, and predictive data mining systems. Each approach generates deep algorithmic questions as well as many algorithm engineering challenges. Finally, we review the design of a predictive gene finding system Glimmer that identifies genes in microbial DNA with 98http://www.tigr.org/softlab/glimmer/glimmer.html

Association Rules, Boosting, and Gene Mining
Shinichi Morishita, University of Tokyo

In this talk, we discuss how to efficiently compute significant association rules according to common statistical measures such as a chi-squared value or correlation coefficient. For this purpose, one might consider to use of the Apriori algorithm, but the algorithm needs major conversion, because none of these statistical metrics are anti-monotone, and the use of higher support for reducing the search space cannot guarantee solutions in its the search space. We present a method of estimating a tight upper bound on the statistical metric associated with any superset of an itemset, as well as the novel use of the resulting information of upper bounds to prune unproductive supersets while traversing itemset lattices. The method is called AprioriSMP, and experimental tests demonstrate its efficiency.

AprioriSMP is useful for classification tasks, because it has a strong connection with AdaBoost that combines a number of classifiers to make a highly accurate prediction. Finally we introduce several applications of AprioriSMP+AdaBoost for mining multiple sets of genes that are highly correlated with phenotype of interest.

Joint work with Jun Sese.

Cubegrades: Generalizing Association Rules
Tomasz Imielinski, Rutgers University

Cubegrades are generalization of association rules which represent how a set of measures (aggregates) is affected by modifying a cube through specialization (rolldown), generalization (rollup) and mutation (which is a change in one of the cube's dimensions). Cubegrades are significantly more expressive than association rules in capturing trends and patterns in data because they use arbitrary aggregate measures, not just COUNT, as association rules do. Cubegrades are atoms which can support sophisticated "what if" analysis tasks dealing with behavior of arbitrary aggregates over different database segments. As such, cubegrades can be useful in marketing, sales analysis, and other typical data mining applications in business.

In this paper we introduce the concept of cubegrades. We define them and give examples of their usage. We then describe in detail an important task for computing cubegrades: generation of significant cubes which is analogous to generating frequent sets.

We also demonstrate how to evaluate simple cubegrade queries and conclude with a number of open questions and possible extensions of the work.

Joint work with Leonid Khachiyan and Amin Abdulghani.

Mining Frequent Patterns without Candidate Generators
Jiawei Han, Simon Fraser University

Mining frequent patterns in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation-and-test is still costly, especially when there exist prolific patterns and/or long patterns.

Recently, we have proposed a frequent pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compact, crucial information about frequent patterns, and develop an efficient FP-tree-based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. Efficiency of mining is achieved with three techniques: (1) a large database is compressed into a condensed, smaller data structure, which avoids costly, repeated database scans, (2) our FP-tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of candidate sets, and (3) a partitioning-based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space. Our performance study shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent pattern mining methods. If time permits, we will also report our recent work on the extensions of this method to mining frequent closed itemsets and mining sequential patterns.

Mining E-Commerce Data: Challenges and Stories from the Trenches
Ronny Kohavi, Blue Martini

The e-commerce domain can provide all the right ingredients for successful data mining: large amounts of data (many records), rich data with many attributes (wide records), clean data, and the ability to take action and measure its effectiveness quickly.

In order to satisfy these requirements, however, merchandising and marketing sites need to design their collection and analysis architectures carefully, and must help simplify the knowledge discovery process.

Architectural decisions in the design of the data mining software made at Blue Martini Software will be motivated and discussed, raising interesting challenges for future research.

Stories and lessons from data mining customer data will be presented.

Integration of Data Mining and Relational Databases
Surajit Chaudhuri, Microsoft Research

In this talk, we discuss the key requirements and challenges in integrating data mining technology in the traditional relational database environment. We present the highlights of OLE-DB for Data Mining API that makes it possible to support data mining models and operations using mining models in the relational framework. OLE-DB for Data Mining is available as part of Microsoft SQL Server 2000 database product.

Predictive Data Mining and Multiple Additive Regression Trees
Jerry Friedman, Stanford University

Predicting future outcomes based on past observational data is a common application in data mining. The primary goal is usually predictive accuracy, with secondary goals being speed, ease of use, and interpretability of the resulting predictive model. New automated procedures for predictive data mining, based on "boosting" (CART) regression trees, are described. The goal is a class of fast "off-the-shelf" procedures for regression and classification that are competitive in accuracy with more customized approaches, while being fairly automatic to use (little tuning), and highly robust especially when applied to less than clean data. Tools are presented for interpreting and visualizing these multiple additive regression tree (MART) models.

Combining Combinatorial and Probabalistic Methods in Datamining
Heikki Mannila, Nokia ans Helsinki University

Data mining methods typically use either purely combinatorial information (e.g., association rules) or purely probabilistic descriptions (e.g., belief networks). Combining these two techniques might yield useful new methods. We discuss some abstract problems arising from such combination approach. We also describe one particular combination, using maximum entropy approach on top of frequent set information for query selectivity estimation.

Privacy-Preserving Datamining
Ramakrishnan Srikant, IBM Almaden Research Center

A fruitful direction for future data mining research will be the development of techniques that incorporate privacy concerns. Specifically, we address the following question. Since the primary task in data mining is the development of models about aggregated data, can we develop accurate models without access to precise information in individual data records?

We consider the concrete case of building a decision-tree classifier ¿from training data in which the values of individual records have been perturbed. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. While it is not possible to accurately estimate original values in individual data records, we propose a novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, we are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data.

Joint work with Rakesh Agrawal.

Massive Data Sets in Astronomy
Michael Vogeley, Drexel University

Sky surveys from ultraviolet to radio wavelengths now underway promise to yield Terabyte-class data sets that will revolutionize the mode in which astronomical research is conducted. These multiwavelength data, including images, spectra, and measured properties of tens of millions of galaxies, stars, and quasars, will be seamlessly archived in object-oriented databases that may be queried by astronomers from around the world. Development of sophisticated query mechanisms and data mining tools will be important to allow complex searches in this vast multi-parameter space. I will focus on results and prospects for the Sloan Digital Sky Survey, a project to digitally map half of the northern sky from ultraviolet to near-infrared wavelengths and to obtain spectra of the one million brightest galaxies.

Using Datamining for XML Data Storage and Compression
Dan Suciu, AT&T

In this talk I will describe two applications of data mining. The first is XML data storage: given a large XML data instance, derive a relational schema for it, then store the XML data in a relational data database system. The advantage is that one can use performant database engines to process the XML data. The challenge however is to find the "best" relational schema. While this leads to a interesting optimization problem, the complexity of the associated decision problem is NP complete in the size of the XML data instance. As a practical substitute, we applied data mining to the XML data instance to derive a "good" relational schema, even if suboptimal.

The second application deals with XML data compression. In XMill we achieve about twice the compression ration of gzip, at about the same speed, when compressing XML data. XMill gets the best compression when the user provides it with "hints" about the data: which fields are integers, or dates, or IP addresses, etc. A natural next step would be to use data mining to derive these hints automatically. While part of the problem relies in just recognizing a few simple datatypes, we observed that the best compression improvements are achieved when "interesting" facts about certain CDATA values are exploited. This leads naturally to an interesting data mining problem on strings, that I will describe.

Classification and Pairwise Relationships: Metric Labeling and Markov Random Fields
Jon Kleinberg, Cornell University

In a traditional classification problem, we wish to assign one of k labels (or classes) to each of n objects, in a way that is consistent with some observed data that we have about the problem. An active line of research in this area is concerned with classification when one has information about pairwise relationships among the objects to be classified; this issue is one of the principal motivations for the framework of Markov random fields, and it arises in fields such as image processing, biometry, and hypertext document analysis. In its most basic form, this style of analysis seeks to find a classification that optimizes a combinatorial function consisting of assignment costs – based on the individual choice of label we make for each object – and separation costs – based on the pair of choices we make for two related objects.

We formulate a general classification problem of this type, the "metric labeling problem"; we show that it contains as special cases a number of standard classification frameworks, including several arising from the theory of Markov random fields. From the perspective of combinatorial optimization, our problem can be viewed as a substantial generalization of the multiway cut problem, and equivalent to a type of uncapacitated quadratic assignment problem. We develop approximation algorithms with provable performance guarantees for the metric labeling problem, including an algorithm that can be applied to an arbitrary metric on a set of k labels, and an arbitrary set of pairwise relationships over the objects to be classified. For the special case in which the labels are endowed with the uniform metric – all distances are the same – our methods provide an approximation algorithm to within a factor of two.

This talk will be based on joint work with Eva Tardos.