The DIMACS/IBM Workshop on Data Mining in the Internet Age

Report

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

May 10, 2000

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

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 can not 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.