Summer School Report: DIMACS Summer School Tutorial on New Frontiers in Data Mining, Aug. 13-17, 2001

Dimitrios Gunopulos
Dept. of Computer Science and Eng.
Univ. of California, Riverside
Riverside, CA 92521
dg@cs.ucr.edu

Nick Koudas
AT&T Research
Florham Park, NJ
koudas@research.att.com

1. INTRODUCTION

The summer school tutorial program on New Frontiers in Data Mining was presented under the auspices of DIMACS, the Center for Theoretical Computer Science at Rutgers University. It took place from August 13, 2001 to August 17, 2001 at the DIMACS center at Rutgers. Dimitrios Gunopulos and Nick Koudas were the organizers of the tutorial that included a large list of invited speakers and presenters. The National Science Foundation provided funding for this event. Fred Roberts, the DIMACS director, provided guidance during the course of this work.

Data mining is an exciting new field of computer science research, encompassing several diverse techniques for analyzing large datasets. The goal of data mining is to obtain new, interesting, and actionable pieces of information. Data mining is an interdisciplinary field, building on a synergy of techniques from theory, statistics, databases and machine learning. As the field grows beyond traditional applications in business and commerce, new techniques and issues emerge, targeting novel applications.

2. TUTORIAL FOCUS

This tutorial program was aimed at providing background, vocabulary, and theoretical methodology to non-specialists in data mining and others who wish to explore this field and at bringing together students, postdocs, and researchers working on algorithms for data mining with those working in related areas, including bioinformatics, networking, and the Web. More specifically, the aim was to introduce the attendees to the fundamental theoretical/algorithmic issues that arise in data mining and its new applications.

The tutorial focused in the following subject areas:

- **Networking Applications:** Networking and telecommunications applications produce large amounts of data that can be mined for various properties of interest. Time series data prevail in such domains and algorithms for time series matching and sequential pattern identification are of great interest. In this context, stream (incremental, one pass) algorithms are of specific interest.
- **Web:** The Web has emerged as a vast data store, containing diverse pieces of information. Recent approaches to mine information on the World Wide Web, include efficient Web searching and Web site personalization efforts.
- **Bioinformatics:** Biological research is undergoing a major revolution as new technologies, such as high-throughput DNA sequencing and DNA micro arrays, create large amounts of data. New techniques in analyzing such data are important in the understanding of biological processes. Many bioinformatics problems can be formulated as generalized searching problems in a large search space.

3. TUTORIAL PROGRAM

3.1 Day 1 and Day 2

The first two days of the tutorial were devoted to an introduction to the general data mining process and several introductory lectures on general data mining tasks, presented by the organizers. These included lectures on techniques for addressing the problems of clustering, classification, time series similarity, mining association rules and sequential patterns, and on dimensionality reduction techniques, and their applications to data mining problems.

The first two days also included a number of lectures on more specific data mining topics. Ramakrishnan Srikant (IBM Almaden) gave a talk on new data mining techniques that are designed to preserve the privacy of individuals. The problem is very general and interesting, and he addressed the specific problem 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, he proposed a novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, it is possible to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data.

Umesh Dayal (HP Labs) gave a lecture on how data mining can be applied to E-business, and what kinds of problems come up in the area. In the first wave of Internet-based e-commerce, data mining has been used primarily in a business-to-consumer (B2C) setting, e.g., for making product recommendations, or for personalized delivery of content. The next wave is widely expected to see the increasing automation of business functions and business-to-business (B2B) interactions. The richer e-business scenarios present numerous opportunities for applying data mining technologies. These range from making business processes such as customer relationship management and supply chain management smarter, to content management for portals and trading hubs.

In addition, Johannes Gehrke (Cornell U.) gave a tutorial lecture on how to construct and use decision tree classifiers; Piotr Indyk (MIT) presented an introduction on external memory structures for solving nearest neighbor problems and Carlotta Domeniconi (UCR) presented an overview of locally adaptive nearest neighbor classification techniques.

3.2 Day 3: Networking

The main topic of the third day was to provide a description of the data mining problems in the field of networking and to discuss specific applications.

In networking, data arrive faster than we are able to query, process and mine. In some cases (e.g., large networks), merely storing all the data produced would be prohibitively expensive. To avoid wasting this data, we must switch from the traditional querying and data mining approach to systems that are able to query and mine continuous, high-volume, open-ended data streams as they arrive. In this context, Sudipto Guha (AT&T and U. Penn) presented an introduction of the stream model of computation, and an overview of algorithmic techniques that have been proposed in this model. Pedro Domingos (U. Washington) described a general method for transforming batch data mining algorithms into data-stream versions and its application to decision tree induction, k-means clustering, and the EM algorithm.

Balachander Krishnamurthy (AT&T) considered the problem of identifying groups of clients that are responsible for a significant portion of a Web site's requests. Such information can be helpful to both the Web site and the clients because it is beneficial to move content closer to groups of clients that are responsible for large subsets of requests to a server.

Matt Grossglauser (AT&T Labs) considered the problem of how to observe the spatial flow of traffic through an operator's domain, i.e., the paths followed by packets between any ingress and egress point in a network. Such knowledge is useful for traffic engineering, capacity planning, and troubleshooting. He proposed a method that allows the direct observation of traffic flows through a domain by observing the trajectories of a subset of all packets traversing the network.

Munir Cochinwala (Telcordia) presented an introduction on techniques for data quality assurance in network databases. Operation Support Systems (OSS), which support a telecommunication carrier's network operations, usually maintain large databases that model physical networks and their components. The issue of data quality is to ensure that the data are correct, current and consistent in the databases. This is vital to the efficiency and effectiveness of the operations.

3.3 Day 4: Web Mining

The main theme of day 4 was the applications of Data Mining to the Web. Using material prepared by Soumen Chakrabarti (IIT Bombay) the organizers presented a general introduction on how data mining techniques have been applied to Web data. Topics covered included the use of hubs and authorities to discover hyper-linked communities, automatic classification of Web documents, Web page clustering, and similar page retrieval.

Prabhakar Raghavan (Verity Inc.) introduced the topic of social networks. Social Networks have been recognized for some time as key mechanisms for information sharing and dissemination. He presented an overview of classical and recent (Web-derived) mining and knowledge discovery algorithms, viewing them in the context of social networks. A recurrent phenomenon in these settings is the presence of power-law distributions. He gave a stochastic model for such distribution, and presented empirical results on text frequency distributions that suggest new methods for mining text associations.

Rajeev Rastogi (Bell Labs) presented an overview of applications of data mining techniques for the Internet. Modern Web search tools are plagued by a number of problems, including result abundance, limited coverage, restricted query interfaces, and limited customization. He discussed how data mining techniques can be used to organize Web content in a more structured fashion and also to improve the quality of search. He also discussed the use of mining techniques for semantic compression and fast query processing of Web data.

Sridhar Rajagopalan (IBM Almaden) talked about graph problems that come up in the analysis of Web data, and gave applications of graph algorithms in this setting.

Panagiotis Ipeirotis (Columbia U.) considered the problem of Categorizing Hidden-Web Databases. The contents of many valuable Web-accessible databases are only accessible through search interfaces and are hence invisible to traditional Web "crawlers." Recent studies have estimated the size of this "hidden Web" to be 500 billion pages, while the size of the "crawlable" Web is only estimated at two billion pages. Recently, commercial Web sites have started to manually organize Web-accessible databases into Yahoo!-like hierarchical classification schemes. He presented a method for automating this classification process by using a small number of query probes.

3.4 Day 5: Bioinformatics

The focus of the fifth day was applications of data mining in the analysis of biological data.

Isidoros Rigoutsos (IBM Watson) presented an overview of pattern analysis problems and techniques in computation biology. One of the many interesting problems in computational biology is the discovery of sub-sequences (patterns) that are common to a given collection of related streams. The streams could be composed of amino acids, nucleotides, gene marker expression levels, natural words, etc. The discovered patterns typically reveal correlated elements with an associated functional, structural or other significance. He also presented some of the bioinformatics-related pattern discovery work done in his group. The applications that he described included a wide range of problems: from unsupervised motif discovery and multiple sequence alignment, to tandem-repeat finding, gene expression analysis, functional annotation, and local structure characterization.

Rahul Singh (Exelixis) gave an overview of computational knowledge discovery and pattern analysis problems in contemporary drug discovery and design. Exploring the relationship between the structure of a molecule and its biochemical properties constitutes the basis of drug discovery. State-of-the-art approaches to this investigation involve a combination of techniques that include physical enumeration and tests based on combinatorial chemistry and high-throughput screening as well as rational pharmaceutical design based on geometric and chemical characteristics of molecule-molecule interaction. Furthermore, understanding and optimizing factors like the effect of the compound on the body and the effect of the body on the compound are essential in developing a drug. Given the exploratory nature of drug discovery, the data volume, and the multiple data modalities, it is therefore, not surprising that the area is rich in algorithmic problems related to knowledge discovery, pattern analysis, and efficient computability.

Dennis Shasha (Courant I., NYU) presented a new approach for mining biological data, called Activist Data Mining. Historically, data miners, like astronomers, wait for data to appear. The implicit model is that cash registers, telephone switches, or other information sources disgorge data that the data miner consumes to derive valuable information. This model, however, lacks a key ingredient: experimentation. He illustrated genomic applications of an alternate "activist" methodology:

1. perform data mining on given data (e.g. wild type species, current store setup),
2. arrive at a hypothesis
3. suggest an experiment to test hypothesis
4. if satisfied then stop else return to 2.

Igor Jurisica (U. of Toronto) discussed the effects that the size and characteristics of the data in bioinformatics, have on the design of mining techniques. Significant advances in molecular technology have caused a data explosion and resulting lower signal to noise ratio for epidemiologists and oncologists attempting to link clinical information to genomic data from patient samples. None of the existing approaches alone is sufficient to find all the relevant patterns in the expression data and only few will scale to highly multidimensional problem space. The analysis of this data is further complicated by many missing values and the fact that no systematic scientific approach is uniformly used for background thresholding and normalization strategies is not optimal. Integrating statistical approaches to gene expression data analysis with heuristic ones provides faster way to identify patterns (but only stronger patterns can be found); integrating heuristics with statistics provides more systematic identification of all patterns. Applying and/or combining diverse computational techniques to analysis of large, multidimensional and multimodal data set of gene expressions combined with clinical descriptors provides increased confidence in discovered patterns and will enlarge the set of discovered patterns. Applying unsupervised learning methods enables us to discover previously unknown cancer subtypes. Addressing important clinical questions in cancer research needs systematic knowledge management and mining of the large amount of gene expression information. This will then help to improve prevention, early diagnosis, cancer classification, prognostics and treatment planning, and to discover useful patterns. He showed how a case-based reasoning (CBR) system can be used for knowledge management in biology and medicine. CBR is an important paradigm for knowledge management because: (a) it is similar to human problem solving and thus it can complement the user; (b) it supports evolving domain models and thus can help to increase domain understanding; and (c) it diminishes the problem of exceptions and over-generalizations.

More information about the summer school tutorial, as well as electronic copies of the tutorial slides, appear in the tutorial Web page at:

`http://www.dimacs.rutgers.edu/Workshops/MiningTutorial`