CCR/DIMACS Workshop/Tutorial on Mining Massive Data Sets and Streams: Mathematical Methods and Algorithms for Homeland Defense

Dates of Tutorial: June 17-19, 2002
Dates of Workshop: June 20-22, 2002
Location: Center for Communications Research (CCR), Princeton, NJ

Organizers:

Bob Grossman, chair
University of Illinois and Two Cultures Group
grossman@uic.edu

Paul Kantor
Rutgers University
kantorp@cs.rutgers.edu

Muthu Muthukrishnan
AT&T Labs - Research and Rutgers University
muthu@cs.rutgers.edu

Workshop Coordinator:
Dolores Koch, DJK@idaccr.org

Co-sponsored by the Center for Communications Research (CCR) and DIMACS.

The amount of data relevant to homeland defense is massive, distributed and growing rapidly through the addition of high volume data streams and feeds. This presents fundamentally new mathematical challenges. These relate to: 1) the real time and near real time detection of significant events in high volume data streams; 2) the forensic analysis of massive amounts of archived data to uncover patterns and events of interest; and 3) the mining of distributed data, which for a variety of reasons will never be centrally warehoused. To complicate matters further, homeland defense must concern itself with a variety of different data types, including signals, text, images, transaction data, streaming media, web data, and computer to computer traffic.

The event brought together researchers from a variety of fields for tutorials and specialized talks about these challenges. The tutorial, which ran from Monday to Wednesday, presented to non-experts or those wanting a coherent introduction to the field a variety of tools that are relevant to the topics described. The workshop, which ran from Thursday through Saturday, contained more specialized talks.

There were tutorials on text mining, parallel data mining, algorithmic issues in processing data streams, database support for data mining, on-line learning, forensic ring analysis, and data fusion, as well as a number of survey talks.

The workshop included talks on algorithmic issues in processing streaming data, text mining and classification, anomaly detection, outlier analysis, forensic data analysis, on-line learning, real-time data mining, parallel data mining, visualization and data mining, and mining graphical data.

Tutorial Program: June 17-19, 2002

Monday, June 17, 2002

7:30 - 9:45	Registration & Reception Breakfast.
9:45 - 10:00	Welcome & Greeting Fred Roberts, Director of DIMACS
10:00 - 12:00	Introduction Data Mining Andrew W. Moore, Carnegie Mellon University
12:00 - 1:30	Lunch
1:30 - 3:30	Classification and Mining of Text Data David D. Lewis, Independent Consultant
3:30 - 4:00	Break
4:00 - 5:30	An Introduction to High Performance Data Mining for Homeland Defense Robert Grossman, University of Illinois at Chicago

Tuesday, June 18, 2002

8:00 - 8:30	Registration & Breakfast
8:30 - 10:30	Streaming Data S. Muthu Muthukrishnan, AT&T Labs - Research and Rutgers University
10:30 - 11:00	Break
11:00 - 12:00	Mining Clickstream Data Stephen G. Eick, Visual Insights
12:00 - 1:30	Lunch
1:30 - 3:30	Web Mining for Hyperlinked Communities Gary William Flake, NEC Research Institute
3:30 - 4:00	Break
4:00 - 5:30	Machine Learning Algorithms for Classification Robert Schapire, AT&T

Wednesday, June 19, 2002

8:00 - 8:30	Registration & Breakfast
8:30 - 10:30	Database Suport for DM Raghu Ramakrishnan, University of Wisconsin
10:30 - 11:00	Break
11:00 - 12:00	Information Extraction Ronen Feldman, ClearForest Corporation
12:00 - 1:30	Lunch
1:30 - 3:30	On-line Learning Manfred Warmuth, University of California, Santa Cruz
3:30 - 4:00	Break
4:00 - 5:30	Data Fusion in Message Filtering Paul Kantor, Rutgers University

Tutorial Abstracts

Stephen G. Eick, Visual Insights

Title: *Mining Clickstream Data*

The growth of the Internet and the dot com revolution has established the e-channel as a critical component of how corporations interact with their customers. One of the unique aspects of this channel is the rich instrumentation where it is literally possible to capture every visit, click, page view, purchasing decision, and other fine-grained details describing visitor browsing patterns. The problem is that the huge volume of relevant data overwhelms conventional analysis tools. To overcome this problem, we have developed a sequence of analysis tools for understanding website structure, paths and flow through the site, website activity, and a click stream analysis tool called eBizinsights. The most innovative aspect of our tools is a rich visual interactive presentation layer.

Gary William Flake, NEC Research Institute

Title: *Web Mining for Hyperlinked Communities*

No doubt about it, the web is big. So big, in fact, that many classical algorithms for databases and graphs cannot scale to the distributed multi-terabyte anarchy that is the web. How, then, do we best use, mine, and model this rich collection of data? Arguably, the best approach to developing scalable non-trivial applications and algorithms is to exploit the fact that, both as a graph and as a database, the web is highly non-random. Furthermore, since the web is mostly created and organized by humans, the graph structure (in the form of hyperlinks) encodes aspects of the content, and vice-versa.

This tutorial will survey the systematic regularities found within the link structure of the web and show how many new and popular algorithms (e.g., HITS, PageRank, and the Community Algorithm) exploit these regularities. We will pay particular attention to algorithms that identify communities or clusters of related web pages. We will further consider how many classical algorithms, which were formulated to minimize worst-case performance over all possible problem instances, can be adapted to the more regular structure of the web.

Robert Grossman, University of Illinois at Chicago and Two Cultures Group

Title: *An Introduction to High Performance Data Mining for Homeland Defense*

A fundamental problem in data mining is to develop data mining algorithms and systems which scale as the amount of data grows, as the dimension grows, and as the complexity of the data

grows. There are now parallel versions of some of the standard data mining algorithms, including tree-based classifiers, clustering algorithms, graph algorithms, and association rules. We will cover these algorithms in detail, as well as provide an overview of some of the general techniques and approaches which have been used developing parallel and high performance version of these and other algorithms data mining algorithms.

Paul Kantor, Rutgers University

Title: *Data Fusion in Message Filtering*

When messages are to be assigned to classes that assignment is based upon features of the message y, which are assumed stochastically dependent on the classification. Let $f(y|c)$ be the distribution of the feature vector y, when the message belongs to class c. When there are two different features, y and y' then study of either feature alone can lead to an optimal decision rule. Data fusion is the effort to exploit multiple sets of features, y,y', y'' and the rules corresponding to each of them separately. In the most general case, it is an effort to estimate the joint distributions $f(y,y',y''|c)$ from some knowledge of the marginals. Historically the problem is approached parametrically.

Various approaches can be unified by considering an abstract space in which the axes represent the likelihood ratios of the several classifications, as computed using the several sets of features. With this as a foundation we can explore Boolean, Fuzzy, linear, non-linear and non-parametric approaches to data fusion. Arguments of simplicity, monotonicity, and disorder have all beem applied to solve the underconstrained problem of data fusion. The ultimate test, however, lies with the effectiveness of the result meta-classifier,

Pointers to the relevant literature on both theory and evaluation will be included. This tutorial presumes a basic knowledge of probability (say, Feller, volume 1) and some mathematical sophistication (early graduate student).

David D. Lewis, Independent Consultant

Title: *Classification and Mining of Text Data*

Natural language text violates every precept of good data modeling: it is ambiguous, redundant, high-dimensional, and has complex and poorly defined structure and semantics. This tutorial will present methods for organizing and analyzing text data in the face of these problems.

You will learn about techniques for classifying natural language of all sorts and finding patterns within and among textual items. Examples will be presented from a variety of real-world text classification and mining applications. The emphasis will be on statistical and machine learning techniques, with some discussion of the role of linguistic processing.

Andrew Moore, Carnegie Mellon University

Title: *Introduction to Data Mining*

This tutorial will survey the fundamental ideas from computer science and statistics that combine together to give data mining. We will see examples of some of the most popular data mining algorithms in action and we will briefly review the key points that make them work. We'll visit decision trees, association rules, non-linear regression, Support Vector Machines and Bayesian approaches. We will then pay careful attention to some classic pitfalls of careless data mining, including overfitting, counfounding causation with correlation, and multiple testing. We will finish with some examples of the kind of work currently taking place in academic and industrial data mining research, including a survey of approaches used to scale statistical data mining to very large data sets, and a couple of exotic models being used in current Homeland Defense data mining applications (Surveillance for bioweapon use, and terrorist network analysis).

S. Muthu Muthukrishnan, AT&T Labs - Research and Rutgers University

Title: *Streaming Data*

Say you go fishing. There are many different types of fish and you catch a bounty. At the end of the day, how can you quickly estimate the number of fish types that are only a few in your catch or those that are the most abundant? Suppose now that your colleague goes fishing and also gets a bountiful catch: how do you quickly check if the fish types you have caught are similar? These problems are interesting when you can not remember all the fish types in your catch, and do not have the time to sort through them for answering the questions above.

Fishing puzzles above are examples of problems that arise when one mines streaming data logs. The Theoretical Computer Science community has recently produced models, algorithms and complexity results to reason about processing streaming data, be it for estimating statistical parameters, clustering or accurate summarization of the streaming signals. In this tutorial, I will present an overview of this emerging theory. Problems become hard when one throws some of the catch back into the pond!

Robert Schapire, AT&T

Title: *Machine Learning Algorithms for Classification*

Machine learning studies the design of computer algorithms that automatically make predictions about the unknown based on past observations. Often, the goal is to learn to categorize objects into one of a relatively small set of classes. This tutorial will introduce some of the main state-of-the-art machine learning techniques for solving such classification problems, namely, decision trees, boosting and support-vector machines. The tutorial will also discuss some of the key issues in classifier design, including avoidance of overfitting.

Manfred Warmuth, University of California, Santa Cruz

Title: *On-line Learning*

We consider learning from examples. We start with a parameterized model class and a loss function that assigns each example and model a non-negative loss.

The on-line algorithm sees one example at a time and incurs a loss on the current example based on its current model. This model (hypothesis) is updated on-line as more examples are seen by the learner. The best fixed model is chosen off-line. It is the model in the class with the smallest (total) loss on all examples.

The loss of the on-line algorithm on a sequence of examples is typically larger than the loss of the best off-line model. However, the goal of the on-line learner is to minimize the additional loss of the on-line algorithm over the loss of the best off-line model. Thus the off-line model serves as a comparator. Bounds relating the on-line loss to the best off-line loss are called relative loss bounds. Such bounds quantify the price of hiding the future examples from the learner.

We will review several methods for deriving on-line algorithms and proving relative loss bounds for them.

Finally we discuss the case when the off-line comparator is allowed to ``shift'' over time. Now the off-line algorithm is allowed to partition the data stream into a small number section and choose the best hypothesis for each section. In contrast, the on-line algorithm does not know the best partition of the data. The algorithms is faced with a dilemma: Should it learn from the recent examples or should it hold on to its old hypothesis.

We show how to resolve this dilemma in various ways and give algorithms with good bounds on the additional loss of the on-line algorithm over the loss of the best ``shifting'' off-line model.

Various applications of these methods will be discussed.

Workshop Program

Thursday, June 20, 2002

8:30 - 9:25	Registration & Breakfast
9:25 - 9:30	Welcome & Greeting Fred Roberts, Director of DIMACS

9:30 - 10:30 Distributed Mining and Monitoring
Johannes Gehrke, Cornell University

10:30 - 11:00 Break

11:00 - 12:00 Massive Multi-Digraphs
James Abello, AT&T and DIMACS

12:00 - 2:00 Lunch

2:00 - 3:00 Semantic Information Processing of Spoken Language in Dialog
Allen Gorin, AT&T

3:15 - 4:00 Learning Mixture of Markov Chains
Sudipto Guha, University of Pennsylvania

4:00 - 4:30 Break

4:30 - 5:30 Content-Sensitive Fingerprinting and its Applications
Rafail Ostrovsky, Telcordia Technologies

Friday, June 21, 2002

8:30 - 9:30 Registration & Breakfast

8:30 - 10:30 Graph Mining: Discovery in Large Networks
Daryl Pregibon, AT&T

10:30 - 11:00 Break

11:00 - 12:00 Change Detection: A Tutorial Overview
Vincent Poor, Princeton University

12:00 - 2:00 Lunch

2:00 - 3:00 Protecting privacy in data-mining applications
Rebecca Wright, DIMACS

3:00 - 4:00 Algorithmic Techniques for Clustering in the Streaming Data Model
Moses S. Charikar, Princeton University

4:00 - 4:30 Break

4:30 - 5:30 Algorithmic Embedding for Comparing Large Text Streams
Graham Cormode, University of Warwick

Saturday, June 22, 2002

8:30 - 9:30	Registration & Breakfast
9:30 - 10:30	Merging of High Bandwidth Data Streams Marco Mazzucco, University of Illinois at Chicago
10:30 - 11:00	Break
11:00 - 12:00	A General Framework for Mining Very Large Databases and Data Streams Geoff Hulten, University of Washington

---

Workshop Abstracts

James Abello, AT&T and DIMACS

Title: *Massive Multi-Digraphs*

A variety of massive data sets exhibit an underlying structure that can be modeled as dynamic weighted multi-digraphs. Their sizes range from tens of gigabytes to petabytes. The sheer volume of these graphs brings with it an interesting series of computational and visualization challenges.

We will discuss external memory algorithms for connectivity, minimum spanning trees and maximal matchings together with heuristics for quasiclique finding and diameter computations. From the visualization store we will describe an external memory hierarchy that allow us to use computer graphics techniques like dynamic visibility to provide navigation control. We will present experimental results with graphs having on the order of 200 million vertices and we will point out some mathematical problems that have surfaced along the way.

(some of this work has been done in cooperation with A. Buschbaum, J. Korn, M. Kreuseler, S. Krishnan, P. Pardalos, M. Resende, S. Sudarsky, A. Ucko, and J. Westbrook)

Moses S. Charikar, Princeton University

Title: *Algorithmic Techniques for Clustering in the Streaming Data Model*

Clustering can be viewed as an optimization problem where the goal is to cluster the data so as to optimize the quality of the clustering measured by an objective function. Such problems have traditionally been studied in the offline model assuming arbitrary access to the data to be clustered. We will present some recent results on clustering in the streaming data model. The

goal here is to produce an implicit description of the clusters in one pass over the data using a small amount of storage so that the corresponding clustering solution is approximately optimal.

To facilitate clustering in a streaming fashion, our model assumes that data items are represented in some form (e.g. as sets or vectors) so that the similarity (or distance) between pairs of items can be computed from their representations. A closely related issue is the design of compact representation schemes that allow possibly complicated data items (e.g. large documents or images) to be represented in a small amount of space so that the similarity between data items can be estimated from their compact representations. We will briefly discuss some recently developed schemes to achieve this.

Graham Cormode, University of Warwick

Title: *Algorithmic Embedding for Comparing Large Text Streams*

Texts are ubiquitous in daily life, varying in size from small (SMS and email) to potentially immense (automatically generated reports, biological sequences). When scanning for similarities between new data and previously stored examples, we need a model that takes account of how texts are changed: pieces are inserted or deleted, and sections are moved around. With a large number of large texts, trying to compare all pairs is not feasible, and for the largest texts we may not even be able to hold the whole of any text in memory at once. We describe an approach to approximately comparing large texts quickly and in sublinear space. It relies on finding combinatorial structures present in any string of characters, and generating a vector representation. This allows rapid comparison of sequences based on a succinct representation, and the application of clustering, nearest neighbor searching, and other data mining techniques.

Allen Gorin, AT&T

Title: *Semantic Information Processing of Spoken Language in Dialog*

The next generation of voice-based user interface technology enables easy-to-use automation of new and existing communication services, achieving a more natural human-machine interaction. By natural, we mean that the machine understands what people actually say, in contrast to what a system designer expects them to say. This approach is in contrast with menu-driven or strongly prompted systems, where many users are unable or unwilling to navigate such highly structured interactions. AT&T's 'How May I Help You?' (tm) technology shifts the burden from human to machine, wherein the system adapts to peoples' language, as contrasted with forcing users to learn the machine's jargon. This technology has been nationally deployed by AT&T for customer care, handling millions of calls each month. In this talk I will describe methods for mining and modeling of speech, language and dialog information from this natural spoken dialog system.

Johannes Gehrke, Cornell University

Title: *Distributed Mining and Monitoring*

We are witnessing the emergence of a new class of applications that is best characterized as monitoring applications. Examples are homeland security services, environmental observation, surveillance and tracking, network management, and sensor data management. Monitoring applications are heavily network-centric, and need to process high-speed data streams in real time with a potentially huge number of complex continuous queries. The current generation of data management architectures and data mining systems are completely inadequate for monitoring applications.

In this talk I will address several technical challenges in building monitoring applications with an emphasis on (1) distributed processing of high-speed data streams, and (2) a framework for quantifying and defining the difference between two datasets.

Sudipto Guha, University of Pennsylvania

Title: *Learning Mixture of Markov Chains*

We consider the problem of inferring a ``mixture of Markov Chains'' based on observing a stream of interleaved outputs from these chains. In the first variant we consider, the mixing process chooses one of the Markov chains independently according to a fixed set of probabilities at each point, and only that chain makes a transition and outputs its new state. For this case, when the individual Markov Chains have disjoint state sets we show that a polynomially-long stream of observations is sufficient to infer arbitrarily good approximations to the correct chains. We also explore a generalizations.

Geoff Hulten, University of Washington

Title: *A General Framework for Mining Very Large Databases and Data Streams*

Much work in KDD has focused on scaling machine learning and statistical algorithms to large databases. The goal is generally to obtain algorithms whose running time is linear (or near-linear) in the size of the database, and that only access the data sequentially. So far this has been done mainly for one algorithm at a time, in a slow and laborious process. In this talk we propose a scaling-up method that is applicable to essentially any induction algorithm based on discrete search. The result of applying the method to an algorithm is that its running time becomes independent of the size of the database, while the decisions made are essentially identical to those that would be made given infinite data. The method works within pre-specified memory limits and, as long as the data is iid, only requires accessing it sequentially. It gives anytime results, and can be used to produce batch, stream, time-changing and active-learning versions of an algorithm. We have applied the method to learning decision trees and Bayesian networks, developing algorithms that are substantially faster than previous ones, while achieving

essentially the same predictive performance. We observe these gains on a series of large databases generated synthetically, from benchmark networks, and on a Web logs containing 100 million requests.

Marco Mazzucco, University of Illinois at Chicago

Title: *Merging of High Bandwidth Data Streams*

Merging two data streams on their keys is one of the fundamental problems in data mining streams. We present practical results for merging streams over optical networks. We examine two algorithms for merging and measure their real world performance. We discuss the limitations of each algorithm and discuss which algorithm is best suited for which data stream enviroment.

Rafail Ostrovsky, Telcordia Technologies

Title: *Content-Sensitive Fingerprinting and its Applications*

A standard notion of a hash-function is the one that maps all documents to short "random-looking" outputs. That is, if two documents differ only in a few bits, a "classical" hash function is geared to output "unrelated" results. In many settings, there is a need for hash-functions which produce "similar" short outputs for "similar" documents. This is especially relevant when searching for a "related" documents in a large database, where one must find answers faster then searching through the entire database, with applications to information distillation and data mining algorithms. In this talk, I will show how to construct such a hash function (for a number of different metric spaces) and describe the underling ideas of the algorithm. I will then show applications of such a hash function to approximate searching, dimension-reduction, clustering, approximate matching, facility location, nearest neighbor search and other approximation problems. The talk will be self-contained.

H. Vincent Poor, Princeton University

Title: *Change Detection: A Tutorial Overview*

Many problems in data mining involve detecting the sudden onset of anomolous behavior in time series data. After a brief discussion of several motivating applications in this area, the fundamentals underlying algorithms for this purpose will be reviewed.

Daryl Pregibon, AT&T

Title: *Graph Mining: Discovery in Large Networks*

Large financial and telecommunication networks provide a rich source of problems for the data mining community. The problems are inherently quite distinct from traditional data mining in that the data records, representing transactions between pairs of entities, are not independent. Indeed, it is often the linkages between entities that are of primary interest. A second factor, network dynamics, induces further challenges as new nodes and edges are introduced through time while old edges and nodes disappear.

We discuss our approach to representing and mining large sparse graphs. Several applications in telecommunications fraud detection are used to illustrate the benefits our approach.

Rebecca Wright, DIMACS

Title: *Protecting Privacy in Data-mining Applications*

Data mining is concerned with revealing information contained in some data. However, in some data mining applications, it is also desirable to protect information other than the computed output at the same time. Doing so can provide protection of personal information, protection of sensitive information, and foster collaboration between different agencies. For example, two different governmental agencies may be more willing to work together to compute outputs of their joint data if they need not reveal that data to each other in order to do so. In this talk, I will discuss some recent work in two areas: secure multiparty computation of approximations and privacy-preserving statistics computations.

Secure multiparty approximations. Approximations are often useful to reduce computation and communication costs in a distributed setting where the inputs are held by different parties and are extremely large. Secure multiparty computation of approximations addresses applications in which the parties want to cooperate to compute a function of their inputs without revealing more information than they have to. Secure multiparty computation is a well-studied tool for computing exact functions without revealing unnecessary information, but these definitions can not be applied directly to approximation functions without a potential loss of privacy. We extend these definitions to apply to secure multiparty approximate computations, and we present an efficient, sublinear-communication, private approximate computation for the Hamming distance; we also give an efficient, polylogarithmic-communication solution for the L2 distance in a relaxed model.

Privacy-preserving statistics: Suppose a client wishes to compute some aggregate statistics on a privately-owned data base. The data owner wants to protect the privacy of the personal information in the data base, while the client does not want to reveal his selection criteria. Privacy-protecting statistical analysis allows the client and data owner to interact in such a way that the client learns the desired aggregate statistics, but does not learn anything further about the data; the data owner leans nothing about the client's query. Motivated by this application, we

consider the more general problem of "selective private function evaluation," in which a client can privately compute an arbitrary function over a database. We present various approaches for constructing efficient selective private function evaluation protocols, both for the general problem and for privacy-protecting statistical analysis.

(This talk includes joint work with Ran Canetti, Joan Feigenbaum, Yuval Ishai, Ravi Kumar, Tal Malkin, Kobbi Nissim, Michael Reiter, Ronitt Rubinfeld, and Martin Strauss.)