National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM'07): Final Report

Report Committee: Tim Finin, João Gama, Robert Grossman, Diane Lambert, Huan Liu, Kun Liu, Olfa Nasraoui, Lisa Singh, Jaideep Srivastava, Wei Wang
Steering Committee: Rakesh Agrawal, Christos Faloutsos, Jiawei Han, Hillol Kargupta, Vipin Kumar, Rajeev Motwani, Philip S. Yu

Abstract

In this report, we review the events of the National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, which was held in Baltimore from October 10 to October 12, 2007. We discuss the key research issues identified by the participants and offer a set of recommendations that evolved out of the presentations and discussions in the symposium.

Index Terms

National Science Foundation symposium, NGDM, next generation data mining, discovery, innovation, scientific data mining, engineering, semantic web, social computing, web 2.0, social science, financial data analysis, bioinformatics, healthcare, security, surveillance, privacy protection, distributed data mining, high performance data mining

The authors are listed in alphabetical order.

CONTENTS

I. Introduction

The dramatic increase in the availability of massive, complex data from various sources is creating many fundamental challenges in computing, storage, communication, and human computer interaction issues for data mining. The 2007 National Science Foundation Symposium on Next Generation Data Mining and Cyber Enabled Discovery for Innovation (NGDM'07) brought together data mining researchers, scientists, engineers and domain experts from a diverse background to discuss various emerging problems that are relevant to Cyber Enabled Discovery for Innovation (CDI). The objective was to enhance the understanding of the challenges in front of the data mining and the CDI community.

NGDM'07 focused on the following areas:

1) Data Mining in e-Science and Engineering
2) The Web, Semantics, and Data Mining
3) Interfacing Data Mining with Social Science, Finance, Medicine, and Other Disciplines
4) Data Mining in Security, Surveillance, and Privacy Protection
5) Ubiquitous, Distributed, and High Performance Data Mining

The symposium was held over two and half days from October 10 to October 12, with 115 attendees. The symposium was divided into five sessions, with one each dedicated to the above areas. The technical program had three components: 1) 35 *individual presentations*; 2) 6 *panel discussions*; and 3) 20 *posters presentations*. In the sections that follow, we first give a grand summary of the recommendations from all sessions. Then, we review the presentations and discussions for each session in more detail, along with the key issues raised by the contributors.

II. Executive Summary

Table I gives a grand summary of the recommendations from all sessions.

III. Data Mining in e-Science and Engineering[1]

A. Introduction

The first session of NGDM'07, which consisted of ten talks and one panel, focused on Data Mining in e-Science and Engineering. In this section we describe the key contributions and recommendations made by each speaker, after which we make overall recommendations for research in this area.

B. Jiawei Han - Research Challenges for Data Mining in Science and Engineering

In the opening speech, Jiawei Han articulately presented nine research challenges for Data Mining in Science and Engineering with exemplar cases: (1)information network analysis, (2) discovery, understanding, and usage of patterns and knowledge, (3) stream data mining, (4) mining moving object data, RFID data, and data from sensor networks, (5) spatial, temporal, spatiotemporal, and multimedia data mining, (6) mining text, Web, and other unstructured data, (7) data cube-oriented multidimensional online analytical mining, (8) visual data mining, and (9) domain-specific data mining by integrating sophisticated scientific and engineering domain knowledge.

C. Saso Dzeroski - Inductive Databases and Queries for Computational Scientific Discovery

Saso Dzeroski systematically presented his vision of employing inductive databases and queries for computational scientific discovery. Computational scientific discovery aims to develop computer systems that automate or facilitate the various activities that humans perform in the process of scientific discovery. Inductive databases are an emerging research area at the intersection of data mining and databases. The creation of new taxonomies, laws, theories, and their revisions make up the bulk of scientific discovery, making them some of the key activities in science. Using the inductive databases (IDBs) framework to support computational scientific discovery presents new challenges: cross-over queries combining patterns and data to obtain new data, induction under constraints, and theory revision.

D. Chris Fischer, Kevin Tibbetts, Dane Morgan, Gerbrand Ceder - Machine Learning for the Computational Materials Scientist

Chris Fisher succinctly argued for the importance of machine learning to computational materials science, which encompasses the computational evaluation of materials before experimental synthesis. Both the size of the design space and large amounts of simulated data provide an arena in which machine learning techniques can focus attention on promising candidates. Two challenges are (1) prediction of the crystal structures that play a critical role in materials design and (2) high-throughput computing that makes available well-tested electronic structure codes and provides detailed views of nearly every known inorganic compound.

[1] This section was prepared by Huan Liu and Wei Wang. Huan Liu is with the Department of Computer Science and Engineering, Arizona State University. Wei Wang is with the Department of Computer Science, University of North Carolina at Chapel Hill.

Session	Key Issues and Recommendations
Data mining in e-science and engineering	• Science operates in a incremental and dynamic manner that acquires data continuously and revises models in response. We need to develop novel data-mining methods that support this cycle, rather than ones that produce single models from fixed data sets. • Work is needed to develop interactive discovery tools and visualization techniques that keep humans in the loop, and take advantage of human perceptual abilities. • Research is needed to develop integrated frameworks, such as inductive databases, that unify the processes of storing, retrieving, mining and managing scientific data and knowledge.
The web, semantics, and data mining	• Since approximately one third to one half of all new Web contents are generated by social media systems such as blogs, wikis and discussion forums, their reach and impact is significant. Research is needed to better understand the information ecology created by these new publication methods to make them and the information they provide more useful, trustworthy and reliable. • The Semantic Web offers languages and standards for specifying common ontologies with well defined meaning that are designed to support sharing and interoperability. Since these are grounded in Web technology, they are by nature compatible with sharing and linking data in a highly distributed and decentralized architecture. Researchers should be encouraged to specify their data models and schemas using semantic languages such as OWL, reusing appropriate published ontologies where possible and publishing any new ontologies developed. When feasible, they should also publish data on the web in RDF or as Web services or SPARQL endpoints, adding to the “Web of Data”. • Research is needed to address the privacy concerns when mining personal information on the web, query logs and click stream data.
Interfacing data mining with social science, finance, medicine, and other disciplines	• Next generation data mining must enable routine analysis of rich sources of data, like literary texts, video and user generated data like blogs, much more straightforward than it is today. • New methods are needed to include multiple sources of supporting data, such as external regulations, results of preclinical trials, past alerts, event histories, etc, in the data mining process and to provide results that are meaningful to the domain expert. • New methods are needed in distributed applications where there is a need to search for matching records and a way to rank the quality of the match at scale. • Domain experts need more advanced visual interfaces for exploratory data analysis. This has relevance in both traditional applications like omics studies and non-traditional applications like literary analysis. • A recurring theme is that any new data mining software, environment or methodology needs to include market-expected requirements as another kind of data and involve domain experts in the design and testing process.
Data mining in security, surveillance, and privacy protection	• Research is needed to develop metrics and models for privacy that bridge the gap between legal definitions and technological protections; and to understand the different implications of data mining on privacy in different domains. • New technology is needed to understand and maintain privacy as data mining expands to mobile (spatio/temporal) data and complex relationships (e.g., social networks.) • Work is needed to improve the performance/security tradeoffs in privacy-preserving data mining, including tasks with benchmarks simulating private data.
Ubiquitous, distributed, and high performance data mining	• Research is needed to develop new algorithms for mining very large distributed data that evolve in time and space in dynamic environments. • Promote multidisciplinary teams both at application level (with the inclusion of domain experts), and at platform level.

TABLE I

GRAND SUMMARY OF THE RECOMMENDATIONS FROM ALL SESSIONS.

E. Vipin Kumar - Discovery of Patterns in the Global Climate System using Data Mining

Vipin Kumar discussed the potential and challenge of mining patterns from observations of the Earth’s climate. Remote sensing data from satellites, combined with results from ecosystem models, offers an unprecedented opportunity for understanding the Earth’s biosphere. Data include measurements of various atmospheric, land, and ocean variables such as sea surface temperature, pressure, precipitation, and Net Primary Production. However, the spatiotemporal nature of Earth Science data, along with its size and high dimensionality, makes pattern mining a difficult task. Nevertheless, he reported impressive results that included detection of ecosystem disturbances and relations between changes in ocean variables and events on land.

F. Haym Hirsh - From Data to Knowledge for Cyber-Enabled Discovery and Innovation

Haym Hirsh enthusiastically presented the background of NSF's new initiative in Cyber-Enabled Discovery and Innovation (CDI). The advent of computing has changed how science and engineering occurs: computers not only provide infrastructure but also offer new metaphors for thinking. CDI aims to promote multi-disciplinary research that takes advantage of computational concepts, methods, models, algorithms, and tools. The initiative seeks transformative research via innovations in and/or innovative use of computational thinking by (1) creating new knowledge from the growing abundance of heterogeneous digital data, (2) understanding complexity in natural, built, and social systems, and (3) advancing our ability to build and leverage the cyberinfrastructure that links people and resources across institutional, geographic, cultural, and temporal boundaries. Taken together, these promise to change both the practice and outcomes of science and engineering research.

G. Steven Salzberg - Computational Gene Finding in the Human Genome: How Many Genes Do We Have?

Steven Salzberg presented the research journey for finding genes in human genome, a challenge that has faced biologists for decades and that remains unsolved. The estimated number of genes has gone from an initial 100,000 to about 20,000 in recent reports. Although there is still uncertainty over the human gene count, we have gained valuable insights and developed many computational methods that let us approach the answer one step at a time. We have also recognized that the accumulation of genome sequences is only the start of understanding biology, and it has made clear that computational thinking is a necessary and crucial component of research.

H. Wei Wang - Data Mining for Enabling Genomic-Wide Computing

Wei Wang shared her vision of current data-mining challenges in biomedical domains. Systems biology has given rise to large-scale, complex, and heterogeneous data at multiple resolutions. These can support studies by the larger scientific community that incorporate genetic, environmental, and developmental variables into comprehensive computational models. However, a number of challenges lie ahead: (1) The dimensionality is high since the data contain massive amounts of information on (relatively) few subjects, and they involve complex correlations and causal relationships among variables; (2) The data are comprised of disparate measurements which include continuous and discrete variables that may not be directly comparable; (3) The data are not static, but grow as new samples and measurements become available; (4) Individual items may be contaminated, noisy, or missing, which makes relationships hard to detect and interpret; (5) The number of unknowns far exceeds the number of knowns, making it intractable to evaluate every potential hypothesis. We need to develop novel and scalable data management and mining techniques that enable high-throughput analysis of genetic networks, real-time genome-wide exploratory analysis, and interactive visualization. This requires new methods to support rapid access and computation of user-specified regions, along with fast and accurate correlation, calculation, and retrieval of loci with high linkage disequilibria.

I. D. Quick and Margaret Dunham - TCGR: A Novel DNA/RNA Visualization Technique

Donya Quick and Margaret Dunham discussed the importance of visualization in science and illustrated this idea through their own experiences with visualizing structural patterns in DNA/RNA sequences. They argued that visual patterns can greatly enhance the identification of motifs, and that it is robust to the presence of insertions, deletions, and single nucleotide polymorphisms. We should develop novel methods that integrate visualizations with sequence alignment algorithms and Markov models, so as to generate more accurate data representations.

J. Kirk Borne - A Machine Learning Classification Broker for Petascale Mining of Large-scale Astronomy Sky Survey Databases

Kirk Borne presented his vision for the emerging field of astroinformatics. Rapid advances in telescopic, detection, and computational techniques have led to increasing amounts of data. To meet the resulting challenges, he reported a global interoperable virtual data system that aims to revolutionize both observational and theoretical research in astrophysics. The project will support the discovery of new knowledge, models, and scientific understanding from the data flowing from large sky surveys. One challenge involves the need for efficient data-mining methods that are more user-centric and collaborative in their operation. They must also handle both known and unknown objects and events in a dynamic fashion, as well as support distributed sharing of data.

K. Larry Hall and Kevin Bowyer - Finding Lookmarks for Extreme-scale Simulation and Scientific Data

Larry Hall and Kevin Bowyer reported their experiences with methods for finding regions of interest in the results of extreme-scale scientific simulations. This task is essential yet challenging in many science and engineering fields, as the data are often highly skewed, with only a small fraction of instances in the most interesting class. A number of data mining and machine learning methods are useful for automatically separating these instances for presentation to scientists. In particular, semi-supervised techniques are needed because there are large data sets, only a portion of which is labeled. Another challenge arises from the heterogeneous and distributed nature of these data, which motivates the need for developing ensembles of classifiers. Their success in developing "lookmarks" demonstrates the promise and importance of this approach.

L. Pat Langley and Alok Choudhary - Panel on Future Research Challenges and Needed Resources for Data Mining Research in e-Science and Engineering

The two panelists for this session were Alok Choudhary and Pat Langley. Alok Choudhary shared his views about large-scale scientific knowledge discovery, noting that the data sets generated in scientific domains often have massive sizes and high dimensionality, as well as being dynamic and distributed. These characteristics require the design and development of new data management and mining methods that include, but are not limited to, techniques that scale well with respect to both data volume and dimensionality, that can handle dynamic data, that support visualization and user interaction, and that one can parallelize and run in a distributed fashion.

Pat Langley argued that most work on scientific data mining is overly narrow and suggested that we should redirect attention toward the broader range of discovery tasks that actually arise in scientific fields. He also offered five challenges for the field: (1) discovering knowledge cast in communicable forms that scientists find familiar; (2) ensuring that discoveries are consistent with background knowledge and thus more plausible to scientists; (3) discovering scientific knowledge from the small data sets that still occur in many fields; (4) generating scientific models that explain data in terms of familiar concepts, rather than merely describing observations; and providing interactive computational aids that support human scientists rather than automated methods that replace them.

M. Overall Recommendations

In summary, science and engineering are fertile lands for data mining, providing both promising applications and challenges that drive development of new methods. Some fields, like astronomy, provide very large data sets that fit traditional assumptions well. Other disciplines, like biology, involve high-dimensional and heterogeneous data that raise new issues. Still others, like ecology, remain data poor but involve search through large spaces of candidate models. Despite their differences, computational methods for data mining and knowledge discovery have key roles to play in each type of discipline.

Data mining and related techniques have proven that they can aid scientists and engineers in addressing the complexities that arise in modern research. But they can provide even greater benefits if we respond to some remaining challenges:

- Science operates in a incremental and dynamic manner that acquires data continuously and revises models in response. We need to develop novel data-mining methods that support this cycle, rather than ones that produce single models from fixed data sets.
- Scientific and engineering data and models are complex in nature, involving variations over time and space, heterogeneous types of measurements, and information at multiple scales. We need more research on methods that address these issues.
- Data mining alone will not satisfy the needs of scientists and engineers, who also require computational tools for storing, retrieving, and managing data and knowledge. We must develop integrated frameworks, such as inductive databases, that unify the processes of mining and managing scientific data.
- Computational discovery in science and engineering is not limited to inducing descriptive patterns from large data sets. There are also important roles for using knowledge to guide search through hypothesis spaces, constructing models that explain observations in familiar terms, and detecting anomalous events of scientific interest.
- Scientists and engineers desire computational aids that empower them and increase their productivity, rather than ones that aim to replace them. We should put greater emphasis on developing interactive discovery tools that keep humans in the loop, rather than focusing on completely automated, standalone algorithms.
- Visualization is one interactive technique that holds great promise, in that it lets scientists comprehend their data, models, and how they relate, as well as communicate their results to other researchers. We need more work on visualization methods that take advantage of human perceptual abilities.

These and other challenges are best addressed in the context of specific problems that arise in science and engineering. Only if we collaborate closely with domain experts and end users will the methods we develop have a substantial impact on the scientific enterprise.

IV. The Web, Semantics, and Data Mining[2]

A. Introduction

The second NGDM session focused on data mining and the web, both in its current form and looking forward to the impact of new models and technologies such as the Semantic Web. The session was chaired by Professor Doina Caragea of Kansas State University and included six talks and a short panel session.

[2] This section was prepared by Tim Finin and Jaideep Srivastava. Tim Finin is with the Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County. Jaideep Srivastava is with the Department of Computer Science and Engineering, University of Minnesota.

B. Raghu Ramakrishnan - Web Data Management: Powering the New Web

Raghu Ramakrishnan of Yahoo Research and University of Wisconsin made a presentation that illustrated how the existing data management infrastructure is not sufficient to address the emerging, and often even present, needs of large scale web data. In a detailed talk, which gave a number of specific examples, he illustrated this thesis, identified the challenges, and proposed approaches to addressing them. He also described the ongoing collaboration he has with AnHai Doan of the University of Wisconsin. Specific concepts discussed were:

- Task centricity: People search the Web in a certain context, i.e. the search is part of some task they are performing. Today's search engines focus on the query alone, and not the context around it. Future of search must try to understand the task the individual is trying to perform to answer the query much better.
- Content interpretation: Interpretation of content is key to being able to understand the context and task being performed. Improved techniques in this are needed
- Social search: Social search is the leveraging of knowledge of people on the Web to provide "answers to questions" vs. the "URLs of pages potentially relevant to keywords" that current search engines provide. Specific techniques for doing so were described.
- PeopleWeb: Leveraging the contributory power of the people on the Web to make the Web content more "semantic" in nature. Potentially another approach to the Semantic Web.
- Community information management: A web service that hosts a variety of extremely large, globally distributed communities faces some unique challenges in data management. These were described for Yahoo! community services. A number of research areas arise. The PNUTS project at Yahoo! is addressing them.
- Massive, distributed data management: A host of new challenges in this area. Age-old problems, when scaled to the Web level, turn into new technical challenges, since the existing solutions do not work.

In summary, Ramakrishnan's talk was a challenge to the research community to reconsider data management issues for web scale on-line services.

C. Danny Weitzner - Information Accountability

Danny Weitzner of the MIT Distributed Information Group and the World-Wide Web Consortium talked about their work on "information accountability" as a new approach to privacy-preserving data mining. According to the authors, a key observation is that current technical approaches are not compatible with the current legal system. In addition, the techniques are not scalable to Web-scale information systems. Moreover, both technology and law are in a state of flux in this area. For example, today most security breaches are from unauthorized inferences that can be drawn from authorized data collection, while the laws are geared towards preventing unauthorized data collection. The concepts proposed include:

- Privacy-safe zone: A secured area where data mining is done, and if laws are respected, privacy risks for the individual are eliminated. Challenges include ensuring that the privacy-safe zone is maintained over time, across an institution, and across the Web. Also, it is unclear what laws apply outside the privacy-safe zone.
- Information accountability: When information has been used, it should to possible to determine what happened, and to pinpoint use that is inappropriate. Rules and law should govern how information is used, and interactions with data are logged in order to provide possibility of machine-assisted human-driven accountability.
- Regulatory patterns for large scale information flows: These should be governed by regulation such as Fair Credit Reporting Act and Securities Laws.
- Accountability architecture: Access control should be through Decentralized Authentication Proofs, based on access rules expressed over data semantics. There must be transparent data usage logging for real-time compliance hints and a posteriori accountability. And, the system should be engineered as Web architecture components.
- AIR: An RDF based policy language for specifying usage rules.

Research questions posed by Weitzner include (i) What kind of language should be used to express information usage rules?, (ii) What tools and techniques will help encourage rule-following and spot non-compliance?, (iii) Can we apply statistically-based anomaly detection and pattern recognition to spot privacy intrusions?, and (iv) What kinds of legal systems will encourage the development of accountable information systems?

D. D. Nitin Agarwal, Huan Liu, John Salerno and Philip Yu - Searching for "Familiar Strangers" on Blogosphere: Problems and Challenges

Huan Liu of Arizona State University presented their paper on Searching for "Familiar Strangers" on Blogosphere: Problems and Challenges. The basic idea is that on the blogsphere, often there are two people who have a lot of interest and habits in common, and could benefit from exchanging information, yet they are unable to do so because they are disconnected on the blogosphere. In a sense, this is like two people who check out the same set of books from the library, but have never met, and hence cannot share their thoughts. Unlike the library, our activities on the blogosphere are logged at the micro level, and its analysis can yield very useful insights into people's behavior. The concepts proposed include:

- Long tail of blogoshpere: This paper shows that the blogoshere has a very long tail, and hence bloggers with similar interests and habits not knowing about each other is the norm rather than the anomaly.
- Familiar strangers: The concept of "familiar strangers" is introduced to capture the phenomenon described. Refinements of it, such as "partial strangers" and "total strangers" are also defined.
- Levels of consideration: The analysis of familiar strangers can be done at the level of a single community, e.g., the community called "A group of those who love history" in MySpace, at the level of a networking site, e.g., all of MySpace, or at the level of the entire blogoshpere.
- Challenges: A number of technical challenges are described, including link analysis, searching via blog posts, the use of context in helping with the search, and ascertaining the ground truth for evaluation purposes.

In conclusion, finding "familiar strangers" on the blogoshpere would be useful, but faces a number of challenges. The long tail of the distribution of blog posts presents the need for finding "familiar strangers" on the blogosphere, and aggregating posts in the long tail demands substantial novel research. A workshop related to this topic will be held in April 2008 (see http://www.public.asu.edu/˜huanliu/sbp08).

E. Vasant Honavar, Doina Caragea - Towards Semantics-Enabled Infrastructure for Knowledge Acquisition of Distributed Data

Vasant Honavar and Doina Caragea, of Iowa State University and Kansas State University, respectively, discussed the issues in knowledge discovery in an environment when data is distributed across different machines/sites, and it is not possible, due to technical or policy reasons, to collect all of it in one place. Following concepts were discussed in this talk:

- Many applications in bioinformatics and computational biology present challenges in mining distributed, semantically heterogeneous data. For example, researchers working on genome annotation often need to access and analyze distributed data sets - e.g., to annotate newly sequenced genomes based on previously annotated genomes.
- Data analysis challenges in a distributed environment: The challenges faced by data analysts and model builders in a distributed environment were described. Specific issues include memory/bandwidth/computing limitations, access and privacy constraints, homonym and synonym problems in data naming, and heterogeneity in attribute domains and measurement units.
- Distributed approach for learning classification trees: Key observation is that as long as certain sufficient statistics are available, access to more detailed data is not needed. Examples of this were provided. An analogy was drawn from statistical parameter estimation.
- Evaluation of distributed learning: The main question here is "what are we missing when we do not have full access to all data?" Examination of this question leads to two evaluation criteria for models generated by distributed learning. First is "exactness", i.e. guarantee that the learned hypothesis is the same as or equivalent to that obtained by the batch counterpart. Second is "approximation", i.e. guarantee that the learned hypothesis is an approximation (in a quantifiable sense) of the hypothesis obtained in the batch setting.
- Self describing and ontology-extending data sources: Not being able to access detailed data from a source can limit what can be done with it. It is shown that if each data source in the distributed environment is "self describing" and "ontology extending", this problem can be solved for the task of data mining. Specific techniques for these are described. The problem of semantic heterogeneity between schemas is proposed to be addressed by using ontology mapping techniques.

In summary, this project draws upon a suite of tools, from distributed data mining, ontology mapping, and data integration to address the problem of data mining in the bioinformatics domain, and specifically in a distributed setting. A number of techniques are presented, and the existing challenges are described.

F. Anupam Joshi, Tim Finin, and the Ebiquity Group - Web 2.0 Mining: Analyzing Social Media

Professor Anupam Joshi of University of Maryland, Baltimore County talked about work with the goal of developing a model of information flow, influence and trust in the Blogosphere. The approach is based on a system that extracts and represents semantic information from blogs, including their respective topical communities, their expressed sentiment and opinions, and their relationships to other blogs and Web resources. The model is evaluated using both live data and synthetic data generated from a new blog graph model that simulates the interactions of blog reading and writing. The following concepts were discussed:

- The game theoretic web: The Web is a competition, in the game theoretic sense, between those who want to develop mechanisms that are fair, and those who want to create asymmetry of information and take advantage of it. Since both sides are interacting intimately on the Web, this game is becoming very interesting and challenging to study.
- Social influence: How the media and social interactions influence our choices and actions.
- Influence spread models: Spread of influence can be modeled as the spread of an epidemic. A challenging issue is identifying the sources of influence that matter the most.
- Authority and popularity: Authority and popularity of sources can be measured by counting links in an appropriate manner. While authority almost always implies influence, popularity does not necessarily do so. The concept of "link polarity" is introduced, which describes whether a link represents a positive or a negative endorsement.

- Influence propagation: Guha et al's model of trust propagation is used for modeling influence propagation.

This talk presented a number of interesting ideas about identifying key influencers in the blogosphere, understanding the impact of their endorsements - both positive and negative, and how it is propagated. It also identified the challenges of working with real blogs, especially the noise factor, and the fact that people are often trying to game the system.

G. Philip Yu, Xifeng Yan, Jiawei Han, Hong Cheng, Feida Zhu - Approximate Frequent Pattern Mining

Philip Yu of the IBM T.J. Watson Research Center presented joint work with Jaiwei Han and others at the University of Illinois, Urbana-Champaign on issues involving data mining using "frequent patterns" – mining for frequent itemsets. This work starts with the observation that while the traditional exact model for frequent pattern requires that every item occurs in each supporting transaction, real application data is usually subject to random noise or measurement error, which poses new challenges for the efficient discovery of frequent pattern from the noisy data. Mining approximate frequent pattern in the presence of noise thus involves addressing two important issues: the definition of a noise-tolerant mining model and the design of an efficient mining algorithm.

H. Panel and Discussion

The session ended with a panel and short discussion. The panel included comments from Gary Strong of Johns Hopkins University and Tim Finin of the University of Maryland, Baltimore County.

Gary Strong made some observations from the perspective of the social sciences and drawing on an internal research project being carried out at MITRE. The talk pointed out some limitations that "social network analysis" has in some important situations. Strong motivated his sort presentation with a paper by Simson Garfinkle on *Leaderless Resistance* (Garfinkel, S.L., *Leaderless resistance today*, First Monday, 3, 2003).

> "Leaderless Resistance is a strategy in which small groups (cells) and individuals fight an entrenched power through independent acts of violence and mayhem. The cells do not have any central coordination, 'they are leaderless', and they do not have explicit communications with one another."

The problem is that people identify with a group and take on the identity of the group along with its values, opinions and beliefs. They may do this even without having direct communication with other members of the group. A challenge for data mining, then, is to discover the existence of such a group in the absence of the usual social networking clues.

Tim Finin talked about some of the challenges that new developments in Web technology bring to data mining. These new Web developments include the "Web 2.0" technologies, social media systems, the Semantic Web and what he characterized as the "game theoretic Web". Signature features of Web 2.0 systems include a focus on mashups and dynamic data exchange, folksonomies and metadata, and richer interactions between clients and servers. These raise many issues for programs that try to mine data out of the resulting Web services and pages. A second trend is the rise of the "Social Web" in systems like blogs, YouTube, Wikipedia and Facebook. Analyzing these dynamic and sometimes faddish applications can uncover serious privacy concerns and also challenge our ability to model and reason about their underlying, sometimes enormous, social graphs and extract and exploit information bound up in text and images. The Semantic Web is another area that is expanding and evolving, with new paradigms ("linked data"), languages (RDFa), and approaches (Freebase), all of which share critical unresolved issues, such as reasoning (how much, where, what kind), search and trust. Finally, Finin identified the "game theoretic web" as what results when web miners and their targets interact to optimize their payoffs. A simple example today is the dance that Google and SEO optimizers do around the algorithms for ranking pages and recognizing spam.

I. Recommendation and Resource Needs

The Web is undoubtedly the most important medium for publishing and accessing data, information, knowledge and services today. Virtually everyone who uses a computer engages the Web, not only to access information, but increasingly to publish information as well. As a medium, its languages, software infrastructure, protocols and modes of use are evolving continuously and at a rapid pace. Mining Web resources and services will be increasingly important and will touch the lives of almost all of the population.

The techniques and research issues involved in mining the Web are grounded in the same principles as mining databases, natural language text and other forms of information. There are differences of various kinds and degrees, however, that deserve attention and present opportunities to foster data mining research in this important context. We mention several of them in the paragraphs below.

a) Social Computing: Web-based social media systems such as blogs, Wikis and message forums are an important new way to publish information, engage in discussions and form communities on the Internet. Since approximately one third to one half of all new Web contents are generated by social media systems, their reach and impact is significant. Governments, corporations, traditional media companies and NGOs are pursuing more effective uses of social media while adapting to this new environment. Citizens, both young and old, are also discovering how social media technology can improve both their lives and their ability to be heard. We must better understand the information ecology created by these new publication methods to make them and the information they provide more useful, trustworthy and reliable.

b) Semantic Web: The Semantic web effort as defined by Tim-Berners Lee and supported by the World-Wide Web Consortium fills a critical enabling role for data mining. It offers languages and standards for specifying common ontologies with well defined meaning that are designed to support sharing and interoperability. Since these are grounded in Web technology, they are by nature compatible with sharing and linking data in a highly distributed and decentralized architecture. New components, such as the SPARQL query language for Semantic Web stores and the RDFa framework for embedding semantic annotations into XHTML are addressing known problems and support practical use cases. Researchers should be encouraged to specify their data models and schemas using semantic languages such as OWL, reusing appropriate published ontologies where possible and publishing any new ontologies developed. When feasible, they should also publish data on the web in RDF or as Web services or SPARQL endpoints, adding to the "Web of Data".

c) Privacy: Mining live data from the Web has the potential for immediate privacy concerns even though the data is publicly available. Collecting and integrating information about individuals can reveal fact that many will find very intrusive. Some information, such as query logs and click stream data, is difficult to anonymize. We need to develop new approaches to thinking about access control and privacy that can address the new challenges. For example, some groups are exploring how to specify constraints or policies that control and constrain how information can be used and techniques that can detect misuse. Making properly anonymized and vetted datasets such as query log and click streams available for researchers will foster research in this area.

V. Interfacing Data Mining with Social Science, Finance, Medicine, and Other Disciplines[3]

A. *Introduction*

Advances in computing and data mining are having a profound effect on fields as diverse as social science, medicine, finance and literary analysis. The session *Interfacing Data Mining with Social Science, Finance, Medicine, and Other Disciplines* brought together data mining and domain experts to discuss the data mining challenges and resource needs that must be met to enable progress in these domains. This section summarizes the talks in this session and gives a set of overall recommendations based on a synthesis of the talks.

B. *V. Hristidis, F. Farfan, R. Burke, A. Rossi and J. White - Information Discovery on Electronics Medical Records*

Using medical records to track the care of particular patients and to understand disease trends and population health has a long history. The introduction of standardized electronic medical records (EMRs) promises to provide much deeper and broader insights into both patient care and population health, but, as Hristidis discussed, there are major barriers to success. One barrier is defining a standard EMR format that is ideal for record storage and retrieval and yet meets the physician's needs for managing patient care and the medical researcher's needs for analyzing population-level data. The xml-based format developed by the Clinical Document Architecture (CDA) of the Health Level 7 group is popular because it has a rich ontology and provides a tree view of the data (patient and doctor at the head, with a body of medications, procedures, vital signs and other health indicators) that is natural to medical practitioners, but it is not ideal. Hristidis noted the following research needs.

- A standard record format that is rich enough to incorporate domain-specific semantics and references to external information sources like medical dictionaries and yet not so complex that it is unmanageable must be designed and tested.
- There are graphs of patient records, reference materials, and medical research reports. What is the right way to merge these graphs?
- New algorithms are needed for gathering and ranking results from EMRs in a medically meaningful way. Algorithms are needed for searching at the level of the patient, the medical center, and the population. Result ranking must incorporate semantic information and be medically meaningful; standard metrics like inverse document weighting are inadequate.
- Record retrieval needs to be personalized, in the sense that rankings, views of relationships, and edge and node weightings depend on the user's role.
- There is a need for a standardized format for results as well as a record format. A result can be a record, a subtree of the XML structure, a path through multiple databases, or summarized information from many records.
- Given the sensitivity of patient records, there must be authentication and secure exchange systems that protect the security and limit the access of different users differently.
- New algorithms will be needed to locate medically similar records through simple queries, where the notion of similar may not be precise.

C. *David Covell - Omics-based Discovery Strategies: Collection, Mining and Analysis*

Covell discussed the challenges in organizing the enormous volumes of information from omics-based studies in a way that can be used to probe disease mechanisms, to identify novel therapeutic targets, and to propose markers for designing

[3] This section was prepared by Olfa Nasraoui and Diane Lambert. Olfa Nasraoui is with the Department of Computer Science and Computer Engineering, University of Louisville. Diane Lambert is with Google Research.

and monitoring therapeutic studies. He emphasized the success that data mining has had in finding locations of activity and mechanisms of action, but argued that much is left to be done. Mining omics data alone provides only a limited glimpse of the complexities inherent in therapeutic systems. There is a crucial need to fuse multiple sources of data, such as cytoxicity studies, mRNA, and pathway information, and to have better ways to organize and visualize data. He also described the need for making hypothesis generation easier and the (conflicting) need to limit the dangers of high false positive rates when looking at many hypotheses. He described the rich and diverse data generation and analysis processes for the National Cancer Institute's screening databases, and presented a strategy to mine this data based on an "Interactive WEB" that fuses data at different levels of detail, context and scales (individual to group). Covell identified the following specific challenges in CDI for omics data.

- There is a need for methods that go beyond sorting hypotheses into true and false and cluster them into similar sets.
- Visual interfaces and user-friendly software for analyzing omics data are lacking. This software must incorporate background knowledge, such as information about possible pathways.
- Decision trees and random forests that can control false positive rates are needed for very high dimensional data.
- Scientists need ways to share chemical, gene expression, mutation and other data from different sources and must have access to negative results from previous studies.
- Tools for meta-analyses that blend pre-clinical data and data from clinical trials are lacking.

D. Olfa Nasraoui - Market Based Decentralized Profile Infrastructure: Giving Back to the User

E-commerce companies maintain a personal profile containing information such as name, address, products viewed, and purchase and payment history on each of their customers. The company can use the profile for targeted marketing or it can sell the profile to other companies, typically without renumeration to the customer. Customers cannot control or benefit from the use of their profile. Nasraoui proposed a market-based profile infrastructure that stands between the server (the business) and the client (the user) that would allow users to retain control of their profiles and allow them to profit from its use. She described the following research challenges.

- A hardware network for storing and sharing repositories of profiles will be huge and continually changing, and hence challenging to design. The infrastructure could be built as a peer-to-peer network, an internet service platform, or a social networking site, for example.
- Logging and communication using the user's computer or mobile device or RFID reader will be needed. New peer-to-peer logging and communication protocols may be needed.
- The mechanism for micro-payments and e-wallets at this large scale needs to be worked out.
- A market (e.g. online automated auctions) between consumers and e-businesses could optimize the sale of profiles. Research in game theory or economics is needed to determine the equilibria of such markets.
- Specifications for privacy policies that encode what can be logged in the user profile and what parts of the profiles can be sold to which buyers must be determined.
- The effects of privacy policies on profile value and network complexity should be studied.
- Privacy preserving data mining algorithms for collaborative filtering will be needed to discover patterns in the exchange of profiles, for example.
- A random walk or network-based graph model may be used to implement distributed collaborative filtering confined to local subnets. Research on the nature of collaborative filtering in a profile market is needed.
- Profile repository growth must be modeled to predict infrastructure and communication needs.

E. David Goldberg - Three Lessons of Ancient and Modern Philosophy for Creative Human-Centered Computation

David Goldberg argued that we need to create new categories of knowledge, not just enhance known categories, and that that requires new data mining paradigms. In the pre-Internet world, humans created errors and computers eliminated errors to build knowledge. In the Internet world, humans provide knowledge that computers find and synthesize. This shift to integrate humans into friendly and effective cyber-environments, sometimes known as human-based computing, brings a new set of design problems that the usual data mining toolkit may not be fully satisfactory for solving. To make his case, Goldberg made recommendations based on lessons from modern and ancient philosophy.

- As Socrates suggested, we must continually ask new questions. We need ways to generate hypotheses that if answered will generate more hypotheses and not just an end result.
- As Aristotle suggested, we need to have better ways to categorize knowledge, and to answer questions such as what do these items have in common and in what ways are they different? In addition to helping to answer these questions, can data mining add more dimensions to the space to extend our knowledge in a particular area?
- Research is needed to understand how human-centered computing can help the user to consider unfamiliar issues starting from a near-blank slate.
- The designer of human-centered systems needs formal ways to understand the nature of the "institutional artifacts" that underlie knowledge.

- The web is a massively interconnected set of institutional facts and hardware that we accept without question. Understanding the nature of the web can lead to a basis for a new discipline of "postmodern systems engineering".

F. Andrew Kusiak - Data Mining and Innovation Science

What does data mining contribute to innovation? That is the question addressed by Kusiak. He stressed that innovation is the development of new knowledge, and it is fundamental to progress in engineering, business, natural, social and behavioral sciences. He supported the idea that the most important steps in innovation are the generation of new ideas and their evaluation. Data mining is essential because people have limited ability to generate and evaluate a large number of hypotheses. Kusiak's recommendations include the following.

- A new data mining framework should include market-expected requirements elicited from multiple sources, not just data.
- Evolutionary computation is a key to data mining for innovation.
- Kusiak proposes the creation of a Living Innovation Laboratory in which data and requirements are collected and analyzed with new tools to showcase the role of data mining in innovation.

G. James Gentle - Challenges in Computational Finance and Financial Data Analysis

James Gentle talked about the challenges in mining financial data to predict price movements, to understand market behavior, or to identify anomalous or fraudulent behavior such as insider trading and market manipulation. Each of these applications requires reliable models of price movements, but the stochastic behavior of price movements is complex. Gentle sees the following research issues in mining financial data.

- Price movements follow non-Gaussian distributions with heavy tails, asymmetry, correlated changes of volatility, quasi long range dependence and seasonality under normal conditions.
- External data about companies that may have bearing on the overall economy must be incorporated into the model. This changes the problem from analysis of a single time series to analysis of time series with event history data of disparate types.
- Relating exogenous events to price movement requires a reliable database of current information, which does not yet exist.

H. Matthew Kirschenbaum - The Remaking of Reading: Data Mining and the Digital Humanities

While data miners have focused mainly on scientific, business and social data, Kirschenbaum focused on data mining in the humanities. The unit of analysis for a humanity scholar is a text. Kirschenbaum showed that treating the text as data can lead to new insights into its meanings. In particular, mapping the text or representing it as a graph can uncover patterns that are difficult for even experts to penetrate. For example, Tanya Clement, a graduate student at the University of Maryland, applied several data visualization tools to *The Making of Americans* by Gertrude Stein to uncover a previously undetected ring structure that she related to the literary themes of the text. Kirschenbaum sees several ways to extend the reach of data mining into the humanities.

- There are tens of millions of books, and a new book is published every 30 seconds. There are far too many books for each to be analyzed manually. On the other hand, many books are now available online through *Google Books* and the *Open Content Alliance*. The development of a standard suite of data mining tools for text analysis would allow scholars to extend the reach of their analyses.
- Close reading of text is best understood in terms of identifying textual elements that provoke the reader or denote a significant change. Distant reading is the search for patterns across an entire text or corpus. Data mining tools in the digital arts must support both.
- There is a need for data mining tools that incorporate morphological and syntactic metadata, such as parts of speech, named entities, or source documents. The domain expert needs to be involved in developing tools, not just testing, to ensure that the tools are effective.
- Training data for comparative classification and analysis are needed.
- We cannot hold an entire text of a novel, short story or long poem within our visual field,just as we cannot hold an entire data set in our visual field. Just as visualizations have been developed in the context of particular scientific applications, new kinds of visualizations are needed for text analysis in the humanities, both for matching across texts and finding structure within a text. There is also a need for visualizations that incorporate metadata about the texts.

I. Panel and Open Discussion: "Future Research Challenges and Needed Resources for Integrating Data Mining with Social Science, Finance, and Medicine"

1) Diane Lambert (Google) : Lambert started by noting that there is a huge amount of social science data on the Web, if we think of social science broadly as the study of how people interact with each other and their environment. The data range from official statistics that are rigorously tested and cleaned, to social network graphs at sites like Facebook and Myspace, bidding patterns at EBay, and informal surveys and votes at sites like YouTube. The web is also a virtual social science observatory, where people and groups gladly self-report, providing data on their opinions, attitudes and behavior that would be expensive to obtain otherwise. The web is also democratic, in the sense that anyone with an internet connection can participate and contribute user generated data, but only a highly non-random sample of people do. The web, then, changes the sources, complexity and scale of data available to social scientists. It also changes the kinds of data that are easily available. Yet the analysis goals are often the same as in other settings: highlight the typical, find patterns, match similarities, describe structure, detect trends, identify anomalies and find outlying episodes. Clearly, scaling data mining tools up for routine analysis of web data is a huge challenge, but there are many other barriers to analysis as well.

- Any analysis requires deciding which data to use (sample) and which to ignore. The gold standard in social science studies is random sampling. We need to understand what randomly sampling web data means, and then provide tools for sampling the web.
- Social science has a long history of testing hypotheses and providing measures of uncertainty about estimates. Social scientists will need data mining tools that report not just anomalies or patterns found but also the strength of the supporting evidence and easily understood conventional measures of bias and uncertainty. These measures will need to take into account the complexity of the data, including the fact that observational data are not random.
- Cluster computing, like the IBM-Google academic cloud computing initiative, dramatically increases the scale of data that can be analyzed. But it is essential to determine the software tools that will make highly parallel distributed data mining accessible to social scientists.

2) Henry Goldberg (Financial Industry Regulatory Authority): Goldberg discussed the challenges and needs of data mining in an ongoing regulatory environment. The major goals of a regulatory agency are to identify anomalies indicative of fraudulent or deceptive behavior and provide a continual feed of alerts to regulators. Usually, alerts must be manually coded and ranked according to risk of harm and supported by data that can be presented as evidence to regulatory agencies and in court. The problem is further complicated by its scale (there are over 400 million transactions per trading day), periods of missing data, the need to link asynchronous data streams, domain volatility, and the adversarial nature and complexity of some market players. Goldberg indicated several research challenges for data mining.

- Regulations are complex and difficult to translate to technical system requirements. More research is needed into how knowledge from experts is best acquired and encoded. Formalizing capital market regulations using semantic and rule-based technologies should also be studied.
- More research is needed on mining alerts to identify meaningful events.
- Online mining of data streams should be further studied to develop methods that are not sensitive to gaps and can find anomalous subsequences, subgraphs or subsets.
- Methods for automatically translating alarms into regulatory evidence should be developed.
- Data should be made available to the research community to spur progress. This requires a system that can flexibly provision the data to meet the needs of different researchers, automatically generate metadata that explains the data structure, and provide background about the problem domain. This will require funding for hardware and software (such as knowledge management tools) and to support regulatory analysts who can contribute financial data models.

J. Floor Discussion

Much of the floor discussion focused on enabling university researchers to work in the intersection of data mining and the social sciences or humanities. In particular, some in the audience questioned whether these disciplines should be the focus of academic data mining at all, given that the data are observational and that there is often no underlying theoretical principles in the sense that physics has, for example. Not everyone, including the National Science Foundation directors, supported this view. However, many university researchers described the difficulties of becoming involved in interdisciplinary work. The costs of learning a new subject area well enough to contribute to its advance or to use it to advance data mining itself are high, and some university departments do not reward research outside core computer science. These departments may even actively discourage any interdisciplinary work, especially by untenured faculty. Others, including some newly tenured professors, argued that their own departments have changed and actively encourage and reward interdisciplinary work. And, of course, the prestige of NSF funding influences what departments consider to be good work.

K. Recommendations and Resource Needs

While each topic has its own specific recommendations, there are several broad themes in the different sets of recommendations; these are given below.

- Next generation data mining must enable routine analysis of rich sources of data, like literary texts, video and user-generated data like blogs, much more straightforward than it is today.
- New methods are needed to include multiple sources of supporting data, such as external regulations, results of preclinical trials, past alerts, and event histories, in the data mining process.

- There is a need for tools or methodologies that make generating multiple hypotheses or theories easier, and also a need for new ways to rank their importance and uncertainty, to cluster them in ways that are sensible in the context of the application, and to limit the dangers of high false positive rates.
- A related issue is the availability of training data and simple tools for extracting representative data from complex, non-random sources. Methods that control for differences in the available data and the population of interest should also be supported.
- More broadly, there is a need for data mining tools that use external information to provide results that are meaningful to the domain expert.
- In distributed applications, whether they be current medical record systems or future infrastructure for profile exchange, there is a need to search for matching records and a way to rank the quality of the match at scale. New methodology and metrics that are more appropriate for the application of interest than current search algorithms is needed.
- Next generation data mining tools should make it easier for domain experts who are not computer scientists to analyze massive distributed data.
- Many data mining applications with distributed data require secure computing and ways to protect the anonymity of data that are beyond current technology.
- Domain experts need visual interfaces to data for exploratory analysis. There is exciting new work in both traditional applications like omics studies and non-traditional applications like literary analysis. There is also creative visualization of online sampled data (e.g., http://www.nytimes.com/2007/10/25/arts/design/25vide.html) but more effort is needed to enlarge the repertoire of graphics available to domain experts.
- A recurring theme is that any new data mining software, environment or methodology needs to include market-expected requirements as another kind of data and involve domain experts in the design and testing process.
- As data mining systems increase in scale and complexity, there is a need for a new kind of systems engineering.

Resources are needed to support both researchers working on data mining fundamentals, including distributed data analysis and visualization and interdisciplinary teams of data mining experts working with domain experts on particular areas of applications.

VI. Data mining in Security, Surveillance, and Privacy Protection[4]

A. *Introduction*

While literature within the field of privacy-preserving data mining (PPDM) has been around for seven years, understanding the role of privacy in this context is still very much in its infancy. Most of the algorithms and approaches that have been introduced are very ad-hoc, computationally expensive, and do not support a formal theoretical approach similar to those that exist in security or databases. As a result, the applications of PPDM are quite rare in industry although there have been a plentiful publications in academia. This lack of standards and applications was echoed throughout the session. Therefore, as we consider issues and directions for the next generation of PPDM, we must preface this discussion with the need to establish a clear definition and theoretical framework for privacy. We must also bear in mind that this framework should be practically feasible in general as well as in the context of data mining applications.

We begin by describing the key contributions of each talk, identifying the recommendations made by each speaker. Please note that some of the talks in this session addressed areas outside of privacy. In this section, we only focus on the components of the talks related to privacy, security, surveillance and the like. We then integrate them and present a comprehensive set of recommendations, discussing each of them in more detail.

B. *Chris Clifton - Is Privacy Still an Issue for Data Mining?*

1) Talk overview: The speaker first gave a brief review of the history of PPDM. He pointed out that although PPDM research has been active in academia, there are still no practical applications in industry. One reason for this is the lack of understanding about what the privacy related problems are and how they relate to data mining. Understanding the problem is critical for marketing the technology. The real problem, emphasized by the speaker, is the misuse of data. For example, card systems usually save customers' data for analysis; however, without data mining, storing those data long term is not necessary. Therefore, data mining is a cause of data misuse and PPDM can help address this problem. As a result, the speaker suggested marketing PPDM as a means of protection against misuse. The speaker also discussed the possibility of marketing PPDM as a collaboration technology, e.g., secure supply chain management. Finally, the speaker identified some key issues for the next generation of PPDM (to be described in the next subsection).

[4] This section was prepared by Lisa Singh and Kun Liu. Lisa Singh is with the Department of Computer Science, Georgetown University. Kun Liu is with IBM Almaden Research.

2) Recommendations:

- Develop a formal and practical definition of privacy. It is not only associated with individually identifiable data.
- Develop PPDM techniques that support profitable usage, e.g., controlling disclosure risk/cost, optimizing supply chain without losing competitive advantage, etc.
- Understand the benefits of data mining. How do we measure the confidence in data mining results? How do we limit an adversary's learning ability? Can privacy be incentive based? For example, are people willing to give better data if privacy is protected?

C. Alessandro Acquisti and Ralph Gross - Privacy Risks for Mining Online Social Networks

1) Talk overview: The research presented focuses on privacy risks associated with information sharing in online social networks. Online social networks including Facebook, Friendster, and MySpace have grown exponentially in recent years. However, because participants reveal vast amounts of personal and sometimes sensitive information, these computer-mediated social interactions raise a number of privacy concerns. In an effort to quantify the privacy risk associated with these networks, the authors combined online social network data and other publicly available data sets in order to estimate whether it is possible to re-identify PII (personally identifying information) from simple PI (personal information). This research supports the claim that large amounts of private information are available publicly.

2) Recommendations:

- Identify ways to quantify the degrees of privacy associated with publicly available data and information shared in online social networks.
- Develop efficient mitigation strategies that can enhance privacy while preserving valuable online interactions.

D. Jaideep Srivastava - Extraction and Analysis of Cognitive Networks from Electronic Communication

1) Talk overview: Social network analysis focuses on understanding social relationships and interactions within a group of individuals. Cognitive analysis of social networks focuses on understanding what an individual's perception is about other individuals in the network. The speaker began by modeling cognitive social networks and presenting quantitative measures for perception and belief. He then illustrated the usefulness of these ideas using the Enron email communication network and also attempts to identify concealed relationships. The speaker then discussed the problem of modeling and analyzing group dynamics in a social network. A new domain for analyzing group dynamics is massively multi-player online games that include tens of thousands of players who work together in groups to accomplish tasks within the game. While this data set, extracted from web logs, is well-suited for understanding the dynamics of group behavior, data collection, appropriate mining algorithms, and scalability are large issues. Further, the theoretical framework for group behavior is still in its infancy, particularly for ad hoc groups.

2) Recommendations:

- Support interdisciplinary research that will advance computer science as well as other disciplines.
- Develop new, scalable approaches for data access and data cleaning.
- Encourage use of large, real world data sets to validate new data mining algorithms.
- While security is a necessity, a balance between PPDM and data analysis is necessary. Future research need to consider information flow during data analysis.

E. Lisa Singh - Exploring Graph Mining Approaches for Dynamic Heterogeneous Networks

1) Talk overview: Much graph mining research to date focuses on simple network models containing a single node type and a single edge type. In this talk, the speaker discussed the need to develop hidden community identification, spread of influence, and group formation mining algorithms for graphs involving many different node and edge types. Because these graphs are large, graph approximations are necessary to adequately tackle different graph mining problems. The speaker described different approximations and abstractions of complex networks for prediction, visualization, and privacy in the context of observational scientific data. Questions under investigation include: when should we use attributes vs. link structure when building predictive models, how can we use visualization to enhance the quality of mining results, and can we use the same abstraction for different mining applications?. In the context of privacy, the speaker discussed the need to formally define what constitutes a privacy breach within a graph. To date, researchers have proposed conflicting definitions. She then discussed the need to understand network topology in order to effectively determine when certain abstractions of the graph are more private than others.

2) Recommendations:

- Promote developing graph mining algorithms for complex, dynamic networks with multiple node and edge types.
- Define privacy breaches in the graphs. What constitutes a breach?
- Develop metrics for understanding the topology of graphs? This structure can then be used to measure the level of anonymity in the network.
- Consider the privacy questions in the context of complex, not simple networks.

F. Shashi Shekhar, Bhavani Thuraisingham, and Latifur Khan - Spatial and Spatio-temporal data mining challenges

1) Talk overview: The large number of geo-spatial data sets has given rise to spatial and spatial-temporal data mining. Applications include geo-spatial intelligence for security, surveillance of crime mappings for public safety, and containing the spread of infectious disease. Classical data mining approaches that assume independent, discrete transactions are not applicable since spatial data is highly auto-correlated and exhibit high degrees of heterogeneity. Further, existing technologies have inadequate models for dealing with richer temporal semantics and heterogeneity. The talk discussed a number of challenges for spatial data mining including: the validity of the independent assumption, the difficulty in capturing pattern continuity, the need to develop approaches for spatial anomaly and outlier detection. The second part of the talk focused on semantic integration of geospatial data. The speakers described a new approach using geospatial web services that integrates the OWL-S service ontology and the geospatial domain specific ontology to facilitate semantic matching services from multiple heterogeneous, independent data sources. They illustrated this approach using data created from ASTER, a thermal emission and reflection radiometer. While the approach is promising, a number of challenges still need addressing, including pixel and semantic merging of neighborhood regions, handling irregular shapes, scalability, and developing security policies for geo-spatial data. Privacy and security have not been sufficiently explored in the context of geo-spatial data. What is an individual's right to privacy? For example, if Google maps can capture an individual's residence, how can the individual maintain privacy?

2) Recommendations:

- Develop scalable approaches for detecting multiple spatial anomalies and for handling uncertain, heterogeneous spatial data.
- Develop richer temporal semantics for geo-spatial data. Current models are inadequate.
- Develop test data sets that can be used to evaluate different methods for spatial-temporal data mining.
- Define privacy for geospatial data and consider implications of public source geo-spatial data on individual privacy.

G. Michael Berry - Automating the Detection of Anomalies and Trends from Text

1) Talk overview: Nonnegative matrix factorization (NNMF) has been widely used to approximate high dimensional data comprised of nonnegative components. This talk presented a NNMF algorithm for detecting anomalies and trends from unstructured text documents. By preserving nonnegativity, NNMF extracts concepts and topics from the document and enables a non-subtractive combination of parts to form a whole. More specifically, the algorithm parses the documents and produces a reduced-rank representation of an entire document space. The resulting feature and coefficient matrix factors are then used to cluster the documents. The speaker demonstrated the performance of the algorithm by using data from the Aviation Safety Reporting System (ASRS). The results show that anomalies of the training documents can be directly mapped to NNMF-generated feature vectors. Dominant features of test documents can then be used to generate anomaly relevance scores for effective anomalies detection. The speaker also explored the challenges of using Multiplicative Method (MM) for solving NNMF.

2) Recommendations:

- Develop scalable, robust and incremental algorithms for solving NNMF.
- Identify ways to interpret the features generated by NNMF. How sparse (or smooth) should factors (W, H) be in order to produce as many interpretable features as possible?
- Study NNMF on multimodal data. Can we build a nonnegative feature space from objects in both images and text? Are there opportunities for multimodal document similarity?

H. Joe Kielman and Ted Senator- Panel on Future Research Challenges and Needed Resources for "Data Mining in Security, Surveillance, and Privacy Protection"

1) Talk overview: The first talk was given by Dr. Kielman from Department of Homeland Security. Dr. Kielman expressed his concerns about public surveillance – a huge amount of data about individuals are constantly being collected through surveillance systems by governmental and private organizations. Using this information effectively is an important issue that has not been carefully investigated. He pointed out that it might be interesting to develop privacy protection techniques that work in real time. His talk also covered other issues such as privacy of DNA figure prints. Dr. Kielman particularly suggested to develop systems that do not use real private data but simulated synthetic data, which is a kind of research they are interested to fund.

The following talk was presented by Dr. Senator from SAIC. Dr. Senator first discussed the definition of "data mining" from both technical and political perspectives. Next, he defined his notion of privacy, which is the ability to prevent linkage of identity to information. Then, he offered some recommendations for doing data mining responsibly: 1) consider privacy implications before beginning the project; 2) be completely transparent regarding the purpose of data collection, how will the data be used, who will have access to it, how it will be secured, where and for how long is the data retained, can individuals access and correct their personal information, etc. He emphasized the fact that technology is only part of the solution for privacy protection while data, processes, policies, authorities, and laws are at least as important and difficult. Finally, he suggested some research problems and needed recourses for the next generation of privacy-preserving data mining.

2) Recommendations:

- Develop techniques to protect genomic information.
- Develop scalable, real time privacy protection techniques.
- Develop synthetic data generation techniques for privacy protection.
- Develop identity-free pattern discovery techniques.
- Develop network/graph anonymization algorithms.
- Develop provably auditable data mining systems.
- Develop privacy enforcement mechanisms and formalize privacy policies.
- Develop relationship-preserving anonymization algorithms.
- To achieve the goal, we need privacy officers who understand technology as well as scientists, managers, and users who understand privacy.

I. Overall Recommendations

Pioneered by Agrawal & Srikant and Lindell & Pinkas' work from 2000, there has been an explosive number of publications in privacy-preserving data mining. Many techniques have been proposed, questioned, and improved. However, compared with the active and fruitful research in academia, applications of privacy-preserving data mining for real-life problems are quite rare. Without practice, it is feared that research in privacy-preserving data mining will stagnate. Furthermore, lack of practice may hint to serious problems with the underlying theories/concepts of privacy-preserving data mining. Identifying and rectifying these problems must be a top priority for advancing the field.

This session served as a forum for researchers, scientists, and engineers to discuss the challenges and opportunities of data mining in privacy protection, surveillance and security. We integrate the contributions from all the speakers and present a comprehensive set of recommendations for the next generation of research.

- Develop a formal theoretical framework for privacy. What is privacy? What is privacy-preserving data mining? What problems are we trying to solve and are the current state-of-the-art approaches reasonable solutions? How does privacy different from security, anonymity, cryptography, etc.?
- Develop privacy models and techniques for specific domains. It is clear that privacy is a domain dependent concept: homeland security, healthcare, business secrecy, entertainment, and ubiquitous computing are each different and pose different privacy requirements. Using the general, formally defined theory of privacy developed in recommendation 1, we should extend the models, evaluation techniques, and privacy metrics for these different domains.
- Develop approaches for complex data. Most data is not simple and independent, e.g. temporal spatial, social networks and graphs, and text. Therefore, for privacy to be useful, it must be a viable option for domains containing complex, dynamically changing data.
- Develop metrics that adequately quantify levels of privacy in different types of data. How much inferences exists in a real world data set prior to use of privacy-preserving techniques? What are 'acceptable' levels of privacy, both from a policy perspective and from a computational one?
- Understand economic and legal aspects of privacy protection. Privacy is no less a societal and economical concept than it is a technological challenge. However, the majority of work on privacy-preserving data mining does not focus on these aspects. Future research directions could include: 1) the economics of privacy; 2) modeling of privacy legislation and automated proofs of adherence; and 3) privacy and utility trade-off.
- Develop practical performance-aware PPDM techniques. Some of the technological impediments to privacy are rooted in the performance of the algorithms. Future work should pay attention to this issue. Research includes efficiency improvements to known algorithms, scalable privacy models, etc.
- Create a benchmark data set repository for data that needs to be privatized. This would let researchers compare algorithms on known data sets to more clearly understand the differences among various approaches.

Privacy-preserving data mining is a fruitful area that has been slow to gain steam. As more researchers, engineers and legal experts delve into this area, standards and theory will begin to take shape. As these are established, the next generation of PPDM will be a fertile ground for all concerned with privacy implications in society.

VII. Ubiquitous, Distributed, and High Performance Data Mining[5]

A. Introduction

We begin with some definitions. High performance data mining refers to the use of high performance computers for data mining.

[5] This section was prepared by João Gama and Robert Grossman. João Gama is with the Laboratory of Artificial Intelligence and Decision Support, University of Porto, Portugal. Robert Grossman is with the National Center for Data Mining, University of Illinois at Chicago and Open Data Group.

The terms distributed data mining and ubiquitous data mining are used in two related but distinct senses. In the first sense the data itself is distributed or ubiquitous. For example, the data for protecting civilian cyber-infrastructure is naturally distributed, and the data associated with mobile cell phones is naturally ubiquitous.

In the second sense, the processing platform used to compute statistical and data mining models may be distributed or ubiquitous. There are a variety of distributed architectures that are used today for data mining including grid systems, peer-to-peer platforms, and cloud-based infrastructures. An example of a ubiquitous computing platform is a distributed sensor network to monitor an ecological system in which processing is done using the sensor device itself, locally using geographical neighborhoods of sensors, and in larger regional or centralized processing centers.

To summarize, distributed data mining refers to data mining in which the data, the processing platform, or both are geographically distributed. Ubiquitous data mining refers to data mining in which data, the processing platform, or both involve mobile, wireless devices.

The questions discussed by the speakers in the high performance, distributed and ubiquitous section of the workshop can be divided into several categories:

Algorithms. What are appropriate algorithms for distributed, high performance and ubiquitous data mining?

Middleware. What is the appropriate middleware and computing platforms for distributed, high performance and ubiquitous data mining?

Community infrastructure. What community infrastructure is required to make progress in distributed, high performance and ubiquitous data mining? Are the appropriate data sets available? Are the necessary computing platforms and test beds available to researchers? Is there a critical mass of interested computer and application scientists to work with data mining researchers?

Privacy and policy issues. Distributed and ubiquitous computing can face difficult privacy and policy issues. Important questions discussed at the workshop include: How can data be shared between organizations for distributed data mining, such as distributed intrusion detection systems? What type of privacy policies does broadcasting geo-location from mobile devices require?

Although these four categories are relevant to the entire workshop, the latter three assume an usual importance for high performance, distributed and ubiquitous data mining. While platforms and Middleware are widely available for data mining in general, there are relatively few, if any, mature, robust high performance, distributed or ubiquitous data mining platforms available to researchers. This creates a barrier to doing work in this area.

Another barrier is the relative lack of community data sets for high performance, distributed and ubiquitous data mining that can be used in some of the same ways that the UCI KDD Archive has been used by the data mining community.

Finally, distributed and ubiquitous data usually involves data from different organizations or multiple individuals which increase the complexity of the privacy and policy issues.

B. Infrastructures for High Performance Data Mining

Two talks refer Data Mining using High Performance platforms. Both refer the need of algorithms with massive levels of parallelism, and adaptive state management. When efficiency is a issue, data miners should develop code tunable algorithms exploiting the key characteristics of the new processors.

1) Marc Snir - High Performance Computing and Data Mining: Discuss research directions in High Performance Computing Systems, discuss the implication of these technologies for data mining and the use of data mining in dynamic load balancing problems.

A key idea behind the talk: *Data Miners need to think big when doing Data Mining on Petascale platforms*. The main obstacle to petascale data mining is dreaming of grand challenges that need it.

Main requirement: *Need algorithms with massive levels of parallelism.*

- Memory bandwidth is limited (costly) but computer power is essentially free.
  Solutions: Need even higher levels of parallelism! Caching and locality: Need algorithms with good locality
- Petascale data mining requires tuneable code generators to adapt to platform and data. (locality + scalability).
- Load Balancing Problem: Amount of computation in DM kernels heavily data dependent – work partitioning results in load imbalances. Hard solution: develop good work predictors and do explicit dynamic task migration.

2) Srinivasan Parthasarathy - Architecture Conscious Data Mining : Efficient and scalable algorithms: code optimization and tuning.

Many state-of-the-art data mining algorithms under-utilize processor resources:

- Data intensive algorithms use lots of memory accesses, that implies high latency penalty.
- Mining algorithms are extremely irregular in nature: the behavior for different data and parameters is hard to predict.
- Use of pointer-based data structures - poor ILP
- Do not leverage important features of modern architectures automated compiler/runtime systems are handicapped because of 1, 2 and 3.

The challenge here is the use of adaptable algorithms design - moldable task partitioning (based on current load balance of the system) with an adaptive state management.

C. GRID and Distributed Data Mining

1) Ian Foster - Grid and Data Mining: More Related Than You Might Think: Second generation grids emphasize the delivery of data and software as services, the federation of services to meet community needs, and the construction of infrastructures designed to host services.

The main challenge here is the *service oriented* architectures and the issues they point out on automation, standards, and trust.

Future directions for service oriented architecture:

- Enable a separation of concerns and responsibilities between those who operate the physical resources that host services, those who construct services, and those who access services
- Provenance is an important issue to address in a substantial and principled manner. Progress in science depends on one researcher's ability to build on the results of another. the results of these activities are only useful when published if other researchers can determine how much credence to put in the results on which they build, and in turn convince their peers that their results are credible. Ultimately we will need to automate these processes.
- Methods used to specify and execute large computations involving many loosely coupled activities. Such computations arise, for example, in large-scale data analyses and parameter studies.
- Research occurs within *communities*, and the formation and operation of communities can be facilitated by appropriate technology. However, many challenges remain: scalability in the number of participants, trust, shared vocabulary, and other implicit knowledge break down as communities extend beyond personal connections.

2) Robert Grossman - Distributed Discovery in E-Science: Lessons from the Angle Project: Lessons Learned in the Angle project:

- The current grid-based distributed computing infrastructure is well-suited for sharing cycles, but less suited for making discoveries using distributed data.
- Necessity of developing algorithms that scale as the number of records, the number of dimensions and the number of analytic models increase.
- Developing policies and procedures so that anonymized data could be shared between organizations. This process is an important lesson to keep in mind when designing projects that require sharing data.

3) Domenico Talia and Paolo Trunfio - How Distributed Data Mining Tasks can Thrive as Services on Grids: Long Term Vision: pervasive collections of data mining services and applications will be accessed anytime and anywhere, and used as public utilities.

- Grid services features as a way to develop data mining services accessible every time and everywhere by providing a sort of knowledge discovery ecosystem formed of a large numbers of decentralized data analysis services.
- Mobile Grid Services allow remote users to execute data mining tasks on a Grid from a mobile devices.
- The Grid as an effective cyber infrastructure for implementing and deploying geographically distributed data mining and knowledge discovery services and applications.
- Future uses of the Grid are mainly related to the ability to utilize it as a knowledge-oriented platform able to run worldwide complex distributed applications. Among those, knowledge discovery applications are a major goal. To reach this goal, the Grid needs to evolve towards an open decentralized platform based on interoperable high-level services that make use of knowledge both in providing resources and in giving results to end users

4) Hillol Kargupta - Thoughts on Human Emotions, Breakthroughs in Communication, and the Next Generation of Data Mining: Key Idea: Client-server model of computing may not be appropriate for the new generation of applications on the Internet that relies upon user's content (e.g. social networking, media sharing).

Alternative: P2P - Global communication through local control and interaction.

- Less reliance upon single-entity owned centralized client-server model of computation with more emphasis decentralized emergence on global behavior through local interactions;
- privacy-sensitive content analysis and match-making between the source and the interested parties in a distributed decentralized environment.
- We need distributed data mining algorithms for large asynchronous networks.
- Local distributed algorithms are particularly interesting because of their scalability.
- P2P web mining for social networking and P2P scientific data mining are some of immediate applications.

D. Ubiquitous Data Mining

Ubiquitous Data Mining is the topic that requires higher levels of distributed computation.

1) Michael May - Research Challenges in Ubiquitous Knowledge Discovery: The ubiquity of data, and ubiquity of computing. Challenges:

- Across a large sector of challenging application domains, further progress depends on advances in the fields of machine learning and data mining; increasing the ubiquity sets the directions for further research and improved applications.

- Ubiquitous Knowledge Discovery requires research beyond independent and identically distributed data.
- Ubiquitous knowledge discovery requires new approaches in spatio-temporal data mining, privacy-preserving data mining, sensor data integration, collaborative data generation, distributed data mining,

2) Xindong Wu, Jeffrey E. Stone, and Marc Greenblatt - User Centered Biological Information Location by Combining User Profiles and Domain Knowledge: The challenges from *Intelligent Digital Libraries* where data resides on the web. The main goal is improve search using user profiles and emerging semantic web technologies to guide a *semantic* search.

The key challenge is improving the semantic search, by means of:

- Web services can provide tools for populating the library from trust sites.
- Exploiting emerging semantic web technologies by means of the annotated web resources.

3) Christos Faloutsos - Large Graph Mining: Graphs is a general structure to represent relations. Graph mining are one of the most challenging data mining problems. Key issues:

- Which laws are behind graphs?
- How to visualize graphs?
- How graphs evolve over time?

E. What are the Challenges?

We can identify two types of challenges: those that arise from the platforms used by communities, and those that arise from the limitations of the current state-of-the-art machine learning and data mining algorithms.

Here are some of the challenges raised by the speakers at the workshops related to high performance, distributed or ubiquitous data mining:

- Develop appropriate middleware to mine very large data, geographically distributed data, and ubiquitous data.
- Develop algorithms and analytic libraries that exploit the high level of parallelism that will be available in emerging petscale systems.
- Develop middleware for integrating data.
- Develop data mining algorithms and associated software that adapt automatically to the specific data of the problem.
- Develop service oriented data mining architectures.
- Develop data mining platforms and middleware that take advantage of emerging wide area, high performance networks.
- Develop tools, systems, and platforms that support privacy preserving data mining.
- Develop platforms and systems that allow local control in distributed, peer-to-peer, and ad hoc systems.
- Develop end-to-end systems that embed data mining in a transparent way.

Algorithmic challenges for high performance, distributed and ubiquitous data mining include:

- Go beyond the *i.i.d.* assumption and stationary distributions.
- Focus in the dynamic and distributed properties of data.
- Develop algorithms that exploit the locality inherent in high performance, distributed and ubiquitous data mining.
- Develop algorithms that work with the unlabeled and partially labeled data that is common in distributed and ubiquitous data mining.
- Use semantic information and background knowledge.
- Develop algorithms that scale not only with the number of records and number of dimensions, but also with the number of different heterogeneous components in a large data set and the number of statistical models.

F. Overall Recommendations

Progress in high performance, distributed and ubiquitous data mining require data sets, software, and hardware platforms. Developing these requires teams. We recommend:

- Support efforts that make large, distributed and ubiquitous data sets available to the community.
- Support efforts that make data mining middleware for high performance, distributed and ubiquitous data mining available to the community.
- Support efforts that make high performance, distributed and ubiquitous testbeds available to the community.

Progress in high performance, distributed and ubiquitous data mining also requires an interdisciplinary approach at both the application and infrastructure level.

- *Application Level:* Between data miners and domain experts. Two issues in data mining that require domain experts are: the necessity to include domain knowledge into the learning process, and the interpretation and evaluation of decision models in the context of the specific application. Both issues recommend interdisciplinary research teams: data miners and domain experts.
- *Platform Level:* Data mining using specific platforms, such as grids, data clouds, petascale platforms or microscale platforms (sensors networks), require teams that include platform experts and application experts, as well as data mining experts.

G. Conclusion

The lessons for data miners might be: *think big*, *think distributed*, *think over time*, *think over space*. Data Mining is moving from static and centralized models to evolving and distributed models. High performance and grid computing are two platforms for the realization of data mining dreams.

VIII. Poster Presentations

This session had twenty posters that covered various topics:

1) e-Science and Engineering: [1], [2], [3], [4]
2) The Web and its Semantics: [5], [6], [7]
3) Social Science, Finance, Medicine, and Other Disciplines: [8], [9], [10], [11]
4) Security, Surveillance, and Privacy Protection: [12], [13], [14], [15], [16]
5) Ubiquitous, Distributed, and High Performance Data Mining: [17], [18], [19], [20]

IX. Concluding Remarks

In summary, the NGDM'07 symposium addressed a variety of challenges that confront the data-mining community, but also explored many promising approaches to them. In coming years, we expect the field of data mining will rise to these challenges and opportunities, building on previous methods but also exploring innovative techniques that move far beyond them.

X. Workshop Website

The official workshop website is http://www.cs.umbc.edu/˜hillol/NGDM07/ .

XI. Workshop Organization

A. General Chair

Hillol Kargupta, University of Maryland Baltimore County & Agnik LLC

B. Steering Committee

Rakesh Agrawal, Microsoft Research
Christos Faloutsos, Carnegie Mellon University
Jiawei Han, University of Illinois at Urbana-Champaign
Hillol Kargupta, University of Maryland Baltimore County
Vipin Kumar, University of Minnesota
Rajeev Motwani, Stanford University
Philip Yu, IBM T. J. Watson Research Center

C. Report Committee

Tim Finin, University of Maryland Baltimore County
João Gama, University of Porto
Robert Grossman, University of Illinois at Chicago
Diane Lambert, Google Research
Huan Liu, Arizona State University
Kun Liu, IBM Almaden Research Center
Olfa Nasraoui, University of Louisville
Lisa Singh, Georgetown University
Jaideep Srivastava, University of Minnesota at Twin Cities
Wei Wang, University of North Carolina at Chapel Hill

D. Local Support Team

Kanishka Bhaduri, University of Maryland Baltimore County
Kamalika Das, University of Maryland Baltimore County
Souptik Datta, University of Maryland Baltimore County
Haimonti Dutta, Columbia University

XII. Acknowledgments

The NGDM'07 symposium is supported by the U.S. National Science Foundation under Grant No. IIS-xxxxxxx. We also acknowledge the sponsorships from Agnik, SAIC and University of Maryland Baltimore County. We are grateful to all the participants in the symposium for their contributions. We thank the report committee and the steering committee for preparing this report. We thank Kun Liu for coordinating the effort, and Pat Langley, Olfa Nasraoui and Chris Clifton for their constructive comments to make the report even better.

References

[1] J. Campbell, K. Molines, and C. W. Swarth, "Data mining for ecological field research: Lessons learned from amphibian and reptile activity analysis."
[2] H. Jasso, W. Hodgkiss, C. Baru, T. Fountain, D. Reich, and K. Warner, "Spatio-temporal characteristics of 911 emergency call hotspots."
[3] J. Tang, R. Yang, D. Barbara, and M. Kafatos, "Mining conditions in rapid intensifications of tropical cyclones."
[4] D. Lo and S.-C. Khoo, "Software specification discovery: A new data mining approach."
[5] L. Raschid, P. Srinivasan, and W.-J. Lee, "A framework for discovering associations from the annotated biological web."
[6] N. I. Yasui, S. Saruwatari, X. Llora, and D. E. Goldberg, "Message feature map toward effective facilitation on on-line discussions."
[7] H. Lauw and E.-P. Lim, "A multitude of opinions: Mining online rating data."
[8] S. Tsumoto and S. Hirano, "Data mining for risk management in hospital information systems."
[9] C. L. Giles, P. Mitra, K. Mueller, J. Z. Wang, B. Sun, L. Bolelli, X. Lu, Y. Liu, I. Councill, W. Brower, Q. Tan, A. Jaiswal, J. Kubicki, B. Garrison, and J. Bandstra, "Chemxseer: An echemistry web search engine and repository."
[10] S. Sahay, E. Agichtein, B. Li, E. V. Garcia, and A. Ram, "Semantic annotation and inference for medical knowledge discovery."
[11] R. Martin, "Investigation of optimal alarm system performance for anomaly detection."
[12] M. Ahluwalia, Z. Chen, A. Gangopadhyay, and Z. Guo, "Preserving privacy in supply chain management: A challenge for next generation data mining."
[13] L. Xiong, P. Jurczyk, and L. Liu, "Mining distributed private databases using random response protocols."
[14] M. Meiss, F. Menczer, and A. Vespignani, "A framework for analysis of anonymized network flow data."
[15] K. Das, K. Liu, and H. Kargupta, "A game theoretic perspective toward practical privacy preserving data mining."
[16] M. Kantarcioglu, B. Xi, and C. Clifton, "A game theoretical framework for adversarial learning."
[17] N. Balac, Olschanowsky, H. Karimabadi, T. Sipes, S. Ferenci, R. Chandran, R. Fujimoto, and A. Roberts, "Distributed data mining system with gateway for virtual observatories."
[18] M. Boley, "Intelligent pattern mining via quick parameter evaluation."
[19] L. Yang and M. Sanver, "Multiresolution data aggregation and analytical exploration of large data sets."
[20] J. Gama, "Issues and challenges in learning from data streams."