Workshop proposal for KDD 2007, San Jose, California, August 12-15, 2007

Title: From heterogeneous sensor data to knowledge discovery for tactical and strategic decision-making

1.0 Description

Remote and in-situ sensors, including wireless sensor networks and large-scale or wide area sensor infrastructures, as well as RFIDs, satellites and GPS systems, are becoming ubiquitous across multiple application domains ranging from disaster management, intelligence analysis and homeland or national security, to transportation, supply chain, ecology, climate change and the management of critical infrastructures. On the one hand, massive volumes of disparate data, typically dimensioned by space and time, are being generated in real-time or near real-time. On the other hand, the need for faster and more reliable decisions is growing rapidly in the face of emerging challenges like natural or man-made disasters, intelligence analysis for security, competitive demand-driven logistics chains, ecological management in the face of environmental change, and the rapid globalization of economies and peoples. The critical challenge for researchers and practitioners is to generate actionable insights from massive, disparate and dynamic data, in real-time or near real-time. We broadly categorize the emerging challenges for knowledge discovery and data mining (KDD) into three overlapping areas: *Data Sciences*, *Computational Sciences* and *Decision Sciences*. While the workshop encourages multiple applications from the sciences and industry, here we describe the three categories in the context of two global priorities: *Natural Disasters* and *Anthropogenic Disasters*.

1.1 Global Priorities

Natural disasters like weather and geophysical extremes, including but not limited to floods, flash floods, heat waves, earthquakes, hurricanes, and tsunamis, are caused through complex nonlinear interactions among multiple variables in space and time at multiple scales. Environmental sensors, which can range from wireless sensor networks to in-situ sensor infrastructures, as well as satellite remote sensors, can generate massive and disparate observations in space and time. Computational physically-based models, which can only hope to be imperfect realizations of the real-world, can nevertheless be useful to discover new insights and develop what-if scenarios. Together, the observed and model-simulated data can be analyzed to develop predictive insights from time scales of a few hours to several decades, with varying accuracy and resolutions. The requirements range from real-time warning systems and decision-aids for imminent natural disasters, to tools for strategic policy-making, which in turn can be invaluable to anticipate the likelihoods and mitigate the impact of disasters.

Anthropogenic disasters are broadly construed here to include accidental or purposeful leakage of chemical, biological and radiation pollution in the atmosphere or water bodies, terrorist attacks, human-induced disruptions to critical transportation infrastructures or logistics chains, as well as pandemic outbreaks and disease spread. Wireless sensor networks and wired wide area sensor infrastructures can generate large volumes of heterogeneous data from multiple sensors, ranging from video or image cameras and remote sensors to weather sensors, as well as sensors for chemical, biological or radiation plumes, traffic movement, cargo monitoring in ships, airplanes and trucks, population displacements and significant change in social behavior. Computational models, which can be physically-based as in the context of plumes, empirically-based as in population or rule-based as in social behavioral change, can generate multiple emergent and/or predictive scenarios at various space-time scales. An intelligent combination of model-simulations and sensor measurements, especially when refined to produce actionable insights, can be effective tools in the hands of analysts and decision-makers, especially when accurate decisions need to be made in real-time or near real-time.

While the above priorities have been used for illustration purposes, we welcome contributions from multiple applications that represent national or global priorities, in areas related to the broad theme of this workshop.

1.2 Data Science

A new focus may be needed within the data sciences to handle the tremendous challenges encountered in high-priority applications like the two examples provided here. While data mining algorithms often, but not always, assume independence of learning samples and envisage scenarios that require either supervised or unsupervised learning, the community has developed tools and methods for mining from dependent samples as in from temporal, spatial or spatio-temporal data, as well as semi-supervised learning paradigms where firm labels are absent but crude domain knowledge may be present. However, further developments are required in these relatively new areas, especially in areas like spatio-temporal data mining and semi-supervised learning when the domain knowledge present is known to be useful, but cannot be easily translated to simple rules or embedded within learning examples. Similarly, the area of anomaly detection from massive and disparate data, both in online and offline modes, has received considerable attention in the last couple of years by the KDD community. However, much remains to be done in terms of automated extraction and detection of anomalies as well as for developing leading indicators, which can either predict such anomalies or utilize the anomalies as predictors depending on the context, in an automated mode, especially for massive and heterogeneous

data dimensioned by space and time. There are a few other significant methodological areas which are critical for the high-priority applications but have received limited attention from the KDD community till date, to the detriment of both the methodological areas and the KDD community. We pick two examples for illustration, specifically, nonlinear statistics and extreme value theory. Nonlinear statistics has been traditionally explored by a specialized group within the physics community while extreme value theory is a focus area within statistics. However, pressing societal concerns ranging from natural disasters, climate change, terrorism and pandemics, have, over the last few years, transformed these relatively arcane disciplines with focused applications, to the forefront of science and technology. The KDD community can contribute significantly to make these relatively non-traditional or new data sciences, with wide applications in our day and age, into tools that can be leveraged across the sciences and business in the context of processing massive data sets in online, offline or event-based modes for real-time or strategic decisions.

1.3 Computational Science

The requirements for sensor information processing vary from archived analysis or "*offline knowledge discovery*" within a central computing, perhaps supercomputing, environments, to real-time or near real-time analysis or "*online knowledge discovery*" and event-based processing, within an in-situ or remote unit, often in a distributed processing environment. A distinction among the two modes of operation may arise as a result of the processing and data integration requirements or the speed and scalability of the data science algorithms. The applicability of non-traditional and novel algorithms to massive and disparate data dimensioned by space and time remain of interest for both situations. The computational science aspects that can be of significant interest and significance to the KDD community are what may be called *Computational Efficiency* and *Algorithmic Efficiency*. We use the former phrase in this context to include innovations in computational speeds, data structures, processing and scalability, primarily for offline processing. The latter phrase is used to describe developing the mathematical algorithms themselves in a manner such that they are amenable to fast, storage- and memory-efficient, adaptive, and real-time or near real-time, implementations for robust online implementations. In addition, the relevant challenges in the computational science include data integration and fusion as well as coordination between offline and online discovery as well as with high-performance modeling and simulations.

1.4 Decision Science

The decision scientific requirements range from strategic policy-making, for example in the context of advance disaster planning and readiness efforts, to real-time decision aids, for example to detect, thwart and/or manage disasters. The decision phase, and hence the decision scientific requirements, is typically strongly coupled to the specific application domain. In addition, this is the phase that brings in the "human in the loop" in the form of analysts, decision-makers and policy strategists. The primary requirements for which KDD tools need to be developed or refined can be categorized into the following: *Hypothesis Generation* and *Decision Support*. The hypothesis generation phase include KDD methods for advanced visualizations and analysis induced through human-computer interactions. In addition to human-driven hypothesis generation, emerging areas like automated generation of hypothesis or causal models, as well as new advances in artificial intelligence, may need to be explored. The decision support aspect comprises traditional tools like OLAP and GIS, with embedded KDD methodologies, coordination with tools in areas like risk analysis and operations research, as well as the development of efficient metrics and tools that can aid tactical, operational or strategic decision-making.

1.5 Contributions

The workshop seeks to explore traditional and novel ideas from machine learning, data mining, statistics, signal processing, nonlinear dynamics, system modeling, operational research and decision theory that can be utilized for heterogeneous data from sensors, which are in turn broadly construed as described earlier. New algorithms, or novel adaptations of traditional and non-traditional methodologies, are desired which are geared towards sensor data and sensor-based applications, with a view to high-priority application domains. Of particular interest are algorithms and approaches for real-time data analysis as well as the coordination of offline KDD tasks with online analysis, followed by decision-making in real-time or near real-time, as well as strategic policy-making. In addition, the workshop welcomes contributions from disciplines like real-time database infrastructures and computational frameworks, spatial and temporal KDD, online data fusion, as well as high performance modeling and simulation, which directly contribute to the primary focus of this workshop on data mining and decision support from sensor data. The workshop strongly encourages research papers as well as case studies and success stories. Finally, focused discussion papers on open research challenges and working papers particularly relevant to the focus of the workshop are welcome.

2.0 Motivation

The expected ubiquity of sensors in the near future, combined with the critical roles they are expected to play in high-priority application solutions, point to an era of unprecedented growth and opportunities. The online knowledge discovery requirements described earlier imply immediate opportunities as well as intriguing short- and long-term challenges for practitioners and researchers in knowledge discovery. The immediate opportunities would be to develop

new data mining approaches and adapt traditional and emerging knowledge discovery methodologies to the requirements of the emerging problems. These would include development and adaptation of algorithms and innovative computational and database approaches. In addition, the knowledge discovery and data mining (KDD) community will be called upon, again and again, as partners with domain experts to solve critical application solutions in business and government, as well as in the domain sciences and engineering. The new challenges would be to broaden the definition of knowledge discovery to focus on online pattern and change detection from continuously evolving and dynamic data in real-time, develop and strengthen ties with other areas of computational and data sciences, and possibly develop innovations in areas like automated hypotheses generation by combining data and semantic information.

Besides the opportunities and challenges mentioned above, perhaps the major motivation is that the KDD community cannot afford to lose the window of opportunity. This is the time when nearly every world body and government agency, for example US federal agencies, including DHS, DOD, NASA, and the Intelligence Communities, as well as large and mid-sized business enterprises, are encountering new problems and challenges that can be solved through online knowledge discovery from sensor data, as defined here. The SensorNet® program led by the Computational Sciences and Engineering division of the Oak Ridge National Laboratory, as well as numerous other government programs and efforts in the US and across the globe, have identified the focus area of the proposed workshop as the key research challenge of the near future. The experience of Oracle's Data Mining, as well as multiple business leaders, suggest rapid growth in the business sector. Researchers and academicians, cutting across traditional disciplinary silos, recognize the value and the potential. We believe this is the time to bring the KDD community together in this important area to establish a much needed leadership position in research and practice in the near term, as well as in the longer-term.

3.0 Target Group

The workshop organizers are in a unique position to organize the target group for this workshop. The organizers from the Oak Ridge National Laboratory, who are all involved with the SensorNet® program in various ways, would provide the connection with government and national laboratories. In addition, ORNL collaborates with multiple academic and business partners in this area and some of them can provide appropriate depth and breadth for the workshop. The organizer from the industry would provide a strong link with a large software vendor and their partners and customers. The organizers from academia, as well as from ORNL, have strong links with researchers and academicians working in the area of KD, sensors and critical application domains. The primary target group will include researchers and practitioners working in the intersection of wide area sensor networks and knowledge discovery, as well as in areas like spatial statistics and data mining which are pivotal to sensor-based data analysis. A major outcome is expected to be to provide a meeting ground for experts in sensors and sensor systems, statistics, KDD, and high-priority domain sciences.

4.0 Duration

A full day workshop is proposed to accommodate the researchers and practitioners from multiple disciplines.

5.0 Workshop Organizers

Auroop R. Ganguly[1]	Olufemi A. Omitaomu[1]	Arindam Banerjee[2]	Marcos Campos[3]
Mallikarjun Shankar[1]	Vladimir A. Protopopescu[1]	George Ostrouchov[4]	Ranga Raju Vatsavai[1]
Brian A. Worley[1]	Shashi Shekhar[2]	Tailen Hsing[5]	Budhendra Bhaduri[1]

1. Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN
2. Department of Computer Science, University of Minnesota, Twin Cities, MN
3. Data Mining, Server Technologies, Oracle Corporation, Burlington, MA
4. Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN
5. Department of Statistics, The Ohio State University, Columbus, OH

6.0 Reviewers, Authors, Attendees

Preliminary lists will be provided upon request. The organizers listed above will also comprise the program committee.

7.0 Point of Contact

Auroop R. Ganguly
Geographic Information Science and Technology Group
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
1 Bethel Valley Road, MS 6085, Oak Ridge, TN 37831
Phone: +1-865-241-1305; Fax: +1-865-241-6261
Email: gangulyar@ornl.gov