Research Statement

Jimeng Sun (jimeng@cs.cmu.edu)

My primary research interests are in the fields of data mining, machine learning, and database systems. My research focuses on the design, analysis, implementation, and experimental evaluation of *streaming data mining algorithms*. Data streams arise in many contexts: sensor measurements, machine monitoring, network traffic flows, stock quotes, phone call records. The crucial question in such applications is how to effectively and efficiently summarize the data. Unfortunately, most of the developments on data mining, machine learning and theoretical computer science in general are 1) not suitable for general streaming setting or 2) not applicable for real streaming applications. Practical mining tools for streams are still very scarce. Therefore, the central theme driving my work is to search for *"Real data, elegant algorithms, practical tools"* to bridge the gap between theoretical developments and real-world applications.

My research balances both theory and system, with an emphasis on developing elegant, scalable, theoretically sound mining tools for massive data from real applications. Working on data mining, a largely interdisciplinary area by nature, I found myself deeply inspired by all kinds of elegant algorithmic developments and also extremely excited by a variety of challenging real-world applications. I believe the key to success is through constant communication and collaboration with domain experts from different areas as well as industrial practitioners. As a graduate student, I have had the opportunity of working with researchers from diverse fields such as data mining, machine learning, databases, networking, storage system, financial accounting. Apart from the synergy created from the combination of these different perspectives, these collaborations have given me a broader viewpoint of the applicability of my work, and also shed light on new directions of research.

Thesis Research

My thesis focuses on incremental pattern discovery targeting at streaming applications where data are arriving continuously in real-time. I try to answer the following questions in general settings: *1) how to find patterns (main trends) in real-time? 2) how to efficiently update the old patterns when new data arrive? 3) how to utilize the patterns to solve other problems such as anomaly detection?*

Unlike standard stream mining where data are semi-infinite time series, I envision the streams in a more general setting and proposed a powerful data model *tensor stream* (TS), which includes time series (standard streams), time-evolving graphs, and data cubes. Under this tensor stream model, several data mining tools are developed, which is described in three categories.

Streams (1st order TS) Traditional data mining assumes data are available up front and analyzes the offline data by scanning through multiple times, whereas many real applications (sensor network, Internet traffic forensic, cluster monitoring) generate huge volume of streams continuously that need to be analyzed on the fly. To tackle this challenging problem, I introduce a simple and scalable algorithm, SPIRIT [4] that incrementally summarizes the input streams into key trends from which the original values can be reconstructed with small error. SPIRIT requires $O(n)$ time and space complexity per timestamp (n is the number of streams), which achieves tremendous improvement compared to $O(n^2)$ time and space that standard principal component analysis (PCA) requires. SPIRIT is further extended for the distributed environment [6], which avoids many problems the centralized approach has, such as communication constraint, power consumption, and privacy concern.

SPIRIT and its variants have great practical significance. I have successfully applied SPIRIT to

several application domains, including anomaly detection on wireless sensor network [5], water distribution monitoring [4], data center monitoring [1]. In particular, I have developed the InteMon system [1, 2] for monitoring a Petabyte data center in real-time. The novelties are 1) *automatic anomaly detection:* it successfully identified some nontrivial anomalies due to "broken correlation" that traditional threshold-based approaches failed to find, and 2) *adaptive historical compression:* it smartly transferred correlations into compression and achieved over 10:1 compression ratio.

InteMon has been presented and demonstrated at multiple venues where very positive feedback are given especially from the industrial practitioners. The success of SPIRIT and its variants also helped us win Pennsylvania Infrastructure Technology Alliance (PITA) grant for two consecutive years (2004-05, 2005-06) for the amount of $128K.

Graphs (2nd order TS) Another important part of my thesis is on graph mining, where the data are modelled as graphs such as who-calls-whom, who-emails-whom social networks, source-destination Internet traffic networks, who-bought-what market basket graphs. The goal is to find communities and anomaly nodes in the graphs. On this territory I developed an efficient algorithm using random walk with restarts and graph partitioning technique that simultaneously constructs communities and finds anomalies on a bipartite graph [8]. Through further collaboration on collaborative filtering with one of the co-author, Deepayan Chakrabarti from Yahoo Research, I helped my advisor Christos Faloutsos obtain $75K Yahoo! Research Alliance gift.

Real graphs are sparse (i.e., most nodes have small number of neighbors), however, most existing methods are oblivious to this property. As a consequence, the mining result are often lack of sparsity, which leads to expensive computation and storage. Based on this observation, I developed *Compact Matrix Decomposition* (CMD) [10]. Compared to state of the art, CMD method achieves *orders of magnitude improvements* on both speed and space. In practice, CMD has been applied to network monitoring in the CMU intranet traffic (500GB) which exhibits early success on its anomaly detection capability.

Tensors (high order TS) Many applications require a more complex data model than time series and graphs. For example, Internet traffic packets consist of source, destination, port number and time; cluster monitoring data have machine id, sensor modality and time; or more generally, data cubes in data warehouses. Powerful as they may be, matrix (including time series or graph) based tools cannot handle such data directly. The crux is that matrices have only two "dimensions" (e.g., measurements vs. time, customers vs. products), while we may often need more, like source, destination, port, timestamps. This is exactly what a tensor is, and of course, a tensor is a generalization of a matrix (and of a vector, and of a scalar). My contribution is to envision all such problems as tensor problems, to use the vast literature of tensors to our benefit, and to introduce new tensor analysis tools, tailored for streaming applications.

In particular, I developed a suite of techniques for analyzing and compressing tensor streams DTA/STA [9] and WTA [7], which successfully applied to many different applications including network forensics computing, financial transaction monitoring, social networks (Enron, DBLP, Cellphone communication). Based on the tensor work, I filed for a patent with Spiros Papadimitriou and Philip Yu at IBM Watson. Also along this line, I will present a tutorial on matrix and tensor tools for large scale data mining at SIAM data mining conference 2007 with Christos Faloutsos at CMU and Tammy Kolda from Sandia national lab.

Other Research

Outside my thesis, I have diverse research experiences that spanned many different fields.

Privacy preservation Data mining and privacy preservation are important areas with a seemingly conflicting goal. One is to reveal patterns in data and the other is to hide sensitive information. How to achieve both? Is it all possible? What are the tradeoffs? The problem becomes even more challenging for the streaming environment. I developed techniques for the privacy preservation in streams [3], which is the first such study on streaming data. Both a mathematical analysis and experimental evaluation on real data are provided to validate the correctness, efficiency, and effectiveness of proposed algorithms. Based on these novel techniques, I am filing for a patent with Feifei Li, Spiros Papadimitriou, George Mihaila, Ioana Stanoi and Philip Yu at IBM Watson.

Fraud Detection I spent two summers (2003, 2004) on the Sherlock project at Center for advanced research at PriceWaterhouse Coopers, the number one accounting firm, working on developing automatic fraud detection techniques for financial data. After my first summer there, I have become an external collaborator for the center. Throughout the collaboration, I have successfully applied/extended several of my research developments on fraud detection application, such as SPIRIT, DTA/STA. A patent is also filed with Dr. David Steier and Krishna Kumaraswamy on systems and methods for investigation of financial reporting information. After my successful summer there, the center for advanced research gave a $50K research gift to Carnegie Mellon university. Currently, I am still collaborating with PWC on applying and developing state of the art data mining techniques on financial applications.

Mobile computing Prior to my PhD study at Carnegie Mellon, I obtained a Master degree at Hong Kong university of Science and Technology, specialized on spatio-temporal databases, spatial access method, selectivity estimation, approximate query processing. I completed the master study in one year and published several papers at top databases conferences and journals including VLDB, ICDE, TODS. Those papers have already received over 140 citations on google scholar.

Future Work

I am extremely thrilled by the potential of my research area. I plan to extend the scope of incremental pattern discovery framework in various directions.

In the near term, I plan to study the *general tensor analysis* which enriches the summarization schemes by allowing not only maximum variance orthogonal projection (currently used in DTA/STA), but other transformations as well such as Fourier, Wavelet or Matching Pursuit. This generalization enables data specific customization. For example, the Fourier transform is more suitable when the data are periodic or concentrate on a few dominating frequencies; while the Wavelet and Matching Pursuit often work better when data are non-stationary or even bursty.

Another interesting topic here is *probabilistic tensor stream*. A lot of interesting questions arise here: how to incorporate the data specific characteristics into the algorithm? How to put in prior knowledge in the summarization? Is there a probabilistic way of determining the size of core tensors? What are the computational cost for various steps (learning, inference)? Here, many interesting connections can be made between numeric linear algebra, information theory and probabilistic graphical model. The ultimate goal is to develop a unified view for all and to build tools that combine the best characteristics of all.

Looking further out, I plan to study two important topics on mining dynamic data: *anomaly detection* and *visualization*. In many cases, it is the abnormal patterns (anomalies) that are of interests, such as fraud detection in financial auditing and intrusion detection in network forensics. *Real-time anomaly detection* becomes even more challenging due to dynamic environments. I believe my research on stream mining can give me an edge on solving this problem.

Another interesting topic is to study the *visualization of dynamic data*. In the constantly changing environment, data grows exponentially over time. It has become increasingly difficult to understand and visualize the data due to the volume and changes. A good visualization of dynamic data can provide intuition and understanding for users which in my mind is at least as significant as the mining result itself. The data mining tools I am working on have very subtle connection to visualization. More specifically, the low-dimensional summary for the streaming data can provide an efficient and simplified view of the original data. I plan to explore this aspect to improve state of the art visualization tools for dynamic data.

From the application aspect, I plan to work with practitioners and continue building mining systems for various applications. In addition to the cluster monitoring and financial application, some other possibilities include (1) bio-surveillance application and (2) biomedical image analysis, such as fMRI brain images. As always, the developed system will be released to other researchers and practitioners. This way, my research hopefully can be quickly transferred to technologies.

Finally, I will continue following the promise "Real data, elegant algorithm, practical tools" to advance the data mining field in a constructive way.

References

[1] Evan Hoke, Jimeng Sun, and Christos Faloutsos. Intemon: Intelligent system monitoring on large clusters. In *Proceedings of the Very Large Data Bases Conference (VLDB)*, 2006.

[2] Evan Hoke, Jimeng Sun, John D. Strunk, Gregory R. Ganger, and Christos Faloutsos. Intemon: Continuous mining of sensor data in large-scale self-* infrastructures. *ACM SIGOPS Operating Systems Review*, 40(3), 2003.

[3] Feifei Li, Jimeng Sun, Spiros Papadimitrou, George Mihaila, and Ioana Stanoi. Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2007.

[4] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos. Streaming pattern discovery in multiple time-series. In *Proceedings of the Very Large Data Bases Conference (VLDB)*, pages 697–708, 2005.

[5] Jimeng Sun, Spiros Papadimitriou, and Christos Faloutsos. Online latent variable detection in sensor networks. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2005.

[6] Jimeng Sun, Spiros Papadimitriou, and Christos Faloutsos. Distributed pattern discovery in multiple streams. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[7] Jimeng Sun, Spiros Papadimitriou, and Philip Yu. Window-based tensor analysis on high-dimensional and multi-aspect streams. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2006.

[8] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Relevance search and anomaly detection in bipartite graphs. *SIGKDD Explorations Special Issue on Link Mining*, 7:48–55, 2005.

[9] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2006.

[10] Jimeng Sun, Yinglian Xie, Hui Zhang, and Christos Faloutsos. Less is more: Compact matrix decomposition. In *Proceedings of the SIAM Data Mining (SDM)*, 2007.