## Tutorial 1

# Uncertain Clustering: Models, Methods and Applications

***Presenter:*** Zhenjie Zhang and Dr. Anthony K. H. Tung
*National University of Singapore*

### ABSTRACT :

Clustering analysis is a well studied topic in computer science with many different applications in data mining, information retrieval and electronic commerce. However, traditional clustering method can only be applied on data set with exact information. With the emergence of web-based applications in last decade, such as distributed relational database, traffic monitoring system and sensor network, more and more uncertain data are becoming ubiquitous in many real applications. No trivial solutions over such uncertain data is available on clustering problem, by extending conventional methods. In this tutorial, we discuss some new studies on uncertain clustering from theories to applications. Several different basic computational models on uncertain clustering will be presented. The models satisfy the requirements of different applications, and are all independent to the clustering criterion and underlying calculation algorithm. Based on the models, we will show how they can be incorporated with some popular clustering algorithms, such as k-means algorithm. Given the methods above, some concrete example will be presented on how to monitor the k-means clustering over moving objects with less communication cost. Finally, we will discuss some extension from k-means algorithm to some more complicated clustering method, such as EM algorithm.

### BIOGRAPHY:

- **Dr Anthony K. H. Tung** is currently an Associate Professor in the Department of Computer Science, National University of Singapore (NUS). He received both his B.Sc.(2nd Class Honour) and M.Sc. in computer sciences from the National University of Singapore in 1997 and 1998 respectively. In 2001, he receive the Ph.D. in computer sciences from Simon Fraser University (SFU). His research interests involve various aspects of databases and data mining. Anthony has published widely in various database and data mining conferences. More details of Anthony's research can be obtained at www.comp.nus.edu.sg/~atung.

- **Zhenjie Zhang** is a PhD candidate in the Database Group at the National University of Singapore. He received his B.Sc. in Computer Science from Fudan University, China. His research interests include general skyline query, unsupervised learning, and game theoretical analysis over large databases. Zhenjie presently has over 10 research papers to his name including papers in major venues such as SIGMOD, ICML and TKDE. He was a recipient of the prestigious NUS President Fellowship in 2007. More about Zhenjie's research can be found at www.comp.nus.edu.sg/~zhangzh2.

*14th International Conference on Management of Data*
*COMAD 2008, Indian Institute of Technology, Bombay, India.*
*December 17-19, 2008*

# Tutorial 2

## Graph Mining Techniques and Their Applications

***Presenter:*** Prof. Sharma Chakravarty (*Univ. of Texas at Arlington, USA)*

### ABSTRACT:

In this tutorial, we present graph mining techniques and their relevance to a number of applications. Most of the currently used mining approaches assume transactional and other forms of data. However, there are a large number of applications for which relationships among data objects are extremely important. For these applications, use of conventional approaches results in loss of information that will critically affect the knowledge discovered. Mining techniques that preserve and exploit the domain characteristics are extremely important and graph mining is one such general purpose technique that uses a graph representation facilitating representation of complex relationships.

Graph mining, as opposed to transaction mining (association rules, decision trees and others), is suitable for mining structural data. Complex relationships that exist between entities can be faithfully represented using graphs. Associations between objects in a complex structure are easy to understand when represented graphically. Most importantly, the representation in graph format preserves structural information.

In this tutorial, we overview transactional mining techniques, contrast them with the requirements of applications, and introduce graph mining as an alternative approach for a large class of applications. In the first half of the tutorial, we present details of several graph mining approaches, such as Subdue, FSG, AGM, and gSpan. In the second half of the tutorial, we present scalability issues of graph mining and how SQL-based approaches can handle graph mining on very large data sizes. Finally, we present a novel application of graph mining for classifying documents (email, web, etc.).

### BIOGRAPHY:

Sharma Chakravarthy is Professor of Computer Science and Engineering Department at The University of Texas at Arlington, Texas. He established the Information Technology Laboratory at UT Arlington in Jan 2000 and currently heads it. Sharma Chakravarthy has also established the NSF-funded, Distributed and Parallel Computing Cluster (DPCC@UTA) at UT Arlington in 2003. He is the recipient of the university-level "Creative Outstanding Researcher" award for 2003 and the department level senior outstanding researcher award in 2002.

He is well known for his work on semantic query optimization, multiple query optimization, active databases (HiPAC project at CCA and Sentinel project at the University of Florida, Gainesville), and more recently scalability issues in graph mining and its applications. His group at UTA has developed DBSubdue and DB-FSG – scalable versions of corresponding approaches for graph mining, and InfoSift – a classification system for text, email, and web that uses graph mining techniques.

His current research includes web technologies, stream data processing, complex event processing, ***mining and knowledge discovery – association, graph and text***, push/pull technologies, web content monitoring, and information integration. He has published over 140 papers in refereed international journals and conference proceedings. He has given tutorial on a number of database topics, such as graph mining, database mining, active, real-time, distributed, object-oriented, and heterogeneous databases in North America, Europe, and Asia. He is listed in Who's Who Among South Asian Americans and Who's Who Among America's Teachers.

Prior to joining UTA, he was with the University of Florida, Gainesville. Prior to that, he worked as a Computer Scientist at the Computer Corporation of America (CCA) and as a Member, Technical Staff at Xerox Advanced Information Technology, Cambridge, MA.

Sharma Chakrvarthy received the B.E. degree in Electrical Engineering from the Indian Institute of Science, Bangalore and M.Tech from IIT Bombay, India. He worked at TIFR (Tata Institute of Fundamental Research), Bombay, India for a few years. He received M.S. and Ph.D degrees from the University of Maryland in College park in 1981 and 1985, respectively.

## Invited Industrial Talk 1

**Title:**

BI on Data and Content Together: What will you do with the derived insights? .

**Speaker:**

Mukesh Mohania, Senior manager (Information Management), IBM India Research Lab

**Abstract:**

Faced with growing knowledge management needs, enterprises are increasingly realizing the importance of seamlessly integrating critical business information distributed across both structured and unstructured data sources. This is especially true for financial institutions, where they can potentially use this integration to gain critical insights into customer trends, fraud and market trends. In this talk, we describe a technology and tool, Linkage Discovery, that associates the customer interactions (emails and transcribed phone calls) with customer and account profiles stored in an existing data warehouse. The associations discovered by Linkage Discovery enable analytics spanning the customer and account profiles on one hand and the meta-data associated or derived from the interaction (using text mining techniques) on the other. We show that actionable insights derived using this tool can be fed back into the system to achieve measurable gains in the business. We also show that using either data or content in isolation will not provide the kind of deep analysis possible when the two are combined.

**Biography:**

Mukesh Mohania received his Ph.D. in Computer Science & Engineering from Indian Institute of Technology, Bombay, India in 1995. Currently, he is a senior manager in IBM India Research Lab, and leading Information Management research group. He has worked in the areas of distributed databases, data warehousing, data integration, and autonomic computing. He received the best paper award for his XML and context-oriented data integration work in CIKM 2004 and CIKM 2005, respectively. He received an award from IBM Tivoli Software in 2004 for his research contribution to Policy Management for Autonomic Computing product. He was also a recipient of the “Excellence in People Management” award in IBM India in 2007. He received the Outstanding Innovation Award from IBM Corporation in 2008 for his Context-Oriented Information Integration work. He is an IEEE and ACM Distinguished Speaker.

## Invited Industrial Talk 2

**Title:**

Internet Research: What’s hot in Search, Advertizing, and Cloud Computing?
Speaker: Rajeev Rastogi, Vice President, Yahoo! Labs Bangalore

**Abstract:**

Web search is one of the most widely used Internet applications, online advertizing is key for companies to make money on the Internet, and cloud computing allows Internet services to be delivered to hundreds of millions of users. In this talk, we discuss the current landscape and future trends in each of these 3 critical areas. Specifically, we highlight the role of information extraction, multimedia search, and Web classification technologies in powering Web search evolution. We also examine the key research challenges in matching ads to page views in the various advertizing models prevalent on the Internet today. And finally, we present some of the main technical challenges in realizing massive clouds with efficient utilization computing resources.

**Biography:**

Rajeev Rastogi is the Vice President of Yahoo! Labs Bangalore where he directs basic and applied research in the areas of Web search, advertizing, and cloud computing. Previously Rajeev was a Bell Labs Fellow and the founding Director of the Bell Labs Research Center in Bangalore, India. Rajeev worked at Bell Labs from 1993 until 2008. During the period, he led a number of research projects that were incorporated into Lucent products and services. These include the Data blitz main-memory database system, the Fellini multimedia storage server, and the Net Inventory auto-discovery engine. His research interests include database systems, data mining, and network management. His most recent research has focused on the areas of network monitoring, network graph compression and analysis, and information extraction.

Rajeev is active in the fields of databases, data mining, and networking, and has served on the program committees of several conferences in these areas. He currently serves on the editorial board of the CACM, and has been an Associate editor for IEEE Transactions on Knowledge and Data Engineering in the past. He has published over 125 papers, and filed over 70 patents of which 40 have been issued. Rajeev received his B. Tech degree from IIT Bombay, and a PhD degree in Computer Science from the University of Texas, Austin.

# Invited Industrial Talk 3

**Title**:
Sybase Appliance for Extreme Analytics

**Speaker:**
Shailesh Mungikar, Senior Engineer and Architect, Sybase Software (India) Pvt. Ltd.

**Abstract:**
Enterprise Data Warehouses (EDWs) are stretched beyond their performance capacity because of mixed workloads, increased number of users, and increased data volumes which, in some cases, can grow greater than 60% a year. Customers are looking to off-load their analytics applications to specialized servers. Furthermore, IT is getting increased pressure from upper management and their Line-of-Business sponsors to fix the performance problems in weeks rather than months. These business requirements of EDWs are referred to as “extreme analytics”. The Sybase Analytic Appliance enables EDWs to support extreme analytics. The presentation will cover few interesting ideas related to the Sybase Analytic Appliance (which comprises the following components):

- A column-based analytics server that requires no special tuning or indexing to deliver query results faster than traditional row-oriented relational databases
- Fully integrated ETL that supports Data-Loading for immediate analysis
- A Data Modeling Tool that reads the source data warehouse schemas and automatically generates the target appliance schema
- A high-availability Server and Storage Technology with redundant hot-swap components and Level 5vRAID
- A Business Intelligence Tool for Reporting, Analysis and Monitoring

**Short bio:**
Sybase is acknowledged as one of the world leaders in Business Intelligence and Datawarehousing products space. Sybase provides BI solutions to many leading organizations in the Telecommunications and Financial industries. Shailesh is working as a Senior Developer and Architect in the Business Intelligence space for Sybase R&D in Pune, India. Shailesh’s software research and development career spans over 15 years. His research interests include Enterprise Middleware, Application Server, Infrastructure programming, and Open Source. Previously, he has worked in Research Labs for leading

US based organizations such as BEA Systems. Shailesh holds a B.E in Computers from the Pune Institute of Computer Technology, Pune, India and a Masters Degree in Software Systems from BITS, Pilani.

# Author Index