

# Saurabh Kataria

Ph.D. Candidate,  
College of Information Science & Technology,  
Pennsylvania State University,  
State College, PA-16801.

Email: [skataria@ist.psu.edu](mailto:skataria@ist.psu.edu)  
Phone #: **(814)876-0852**  
Homepage: <http://www.personal.psu.edu/ssk164>

---

## Education

- Aug 2006 – Present  
Ph.D. Candidate, College of Information Science & Technology, Penn State University.
- Aug 2001 – May 2005  
B. Tech. (with honors), Computer Science and Engineering, Institute of Technology, IT-BHU, India.

## Research Interests

Applied Machine Learning, Information Extraction, Social Network Analysis

## Work Experience

- **Jun '07 – Present**  
Working as a graduate assistant for Chem<sup>x</sup>seer (<http://chemxseer.ist.psu.edu>) project which is a digital library for chemistry related literature and provides advanced search capabilities such as table search, chemical entity search and figure search. My responsibilities include:
  - Metadata extraction pertaining to charts (e.g. 2-D line, bar and pie charts) from research papers to enable figure search
  - Figure data extraction for post-processing
  - Image Classification to different chart categories
- **May '05 - July '06**  
Worked as a software developer at GlobalLogic Software Solutions, Noida, India.

## Related Projects

- **Figure Search**  
Findings in scientific studies are usually reported as charts such as 2-D line, bar and pie charts. Digital library search engines such as citeseerx and chem<sup>x</sup>seer tend to underutilize this source of information while indexing its content. The primary focus of this project is to provide the digital libraries with the capability to search for the charts which requires identification and classification of the charts, information extraction and ranking of the charts relevant to user queries.
- **Semi-Supervised Image Clustering**  
Traditionally, image clustering is performed by utilizing surrounding text features or by the image features e.g. color histograms and texture features. Given the inherent noise in the real-world data, agreement in clustering induced by these two different features becomes an issue while utilizing both of the features simultaneously. The aim of this project is to perform co-clustering of images in the two incompatible feature spaces with a few known image labels. The problem reduces to semi-supervised tri-partite graph partitioning.
- **Joint Probabilistic Models for Combining Link and Content in a Social Network**  
Generative models e.g. Latent Dirichlet Allocation explains the generation of content without taking link information into the account. Although the link information is certainly informative about any network, modeling generation of links simultaneously with content, however, can be very computationally expensive tasks. This project aims at developing techniques by combining these two separate information sources present in various social networks such as citation network in a document collection, collaboration network in scientific community etc. to reveal certain characteristics e.g. community memberships of individual authors, document similarity etc.
- **Blocking storage Model for Relation Databases**  
Developed the workload sensitive database storage layout on disk on top of the traditional N-ary and Decomposition storage layouts. The later layouts are static with respect to dynamically changing query workloads that can cause inefficient data fetching operations from disk. The aim of the project was to analyze the query workload and suggest an efficient layout to the storage engine of the relational

database. For the implementation purposes, Shore database system, developed by CS at Wisconsin, was used. [Done as an independent project.]

- **Payment Processing Solutions**

Worked upon coding the software targeting the merchant services for U.S. Health Industry titled "Payment Processing Solutions". My responsibilities included design of the database, implementation of Data Access Layer on .NET platform, User Management using Active Directory Services and optimization of operations over database for the software.

## **Awards/Honors**

- Student travel grant awards for AAAI'08 (Chicago, US) and JCDL'08 (Pittsburgh, US).
- Teaching Assistantship since Aug'06 at College of Information Science & Technology at Penn State University.
- Outstanding summer intern at 'GlobalLogic Software solutions' during May '05-Aug'05
- State Merit award for standing among top 5 students in Senior Secondary Examination.

## **Technical Proficiency**

- Languages and scripting environments - C/C++, Java, Perl and Python.
- Database related - MySQL, SQL-Server and XML.
- Operating systems - Unix/Linux, Windows.

## **Publications**

- **Saurabh Kataria**, William Brouwer, P. Mitra, C. Lee Giles, "Automatic Extraction of Data Points and Text Blocks from 2-Dimensional Plots in Digital Documents", Association for the Advancement of Artificial Intelligence (AAAI)-2008.
- William Brouwer, **Saurabh Kataria**, Sujatha Das, P. Mitra, C. Lee Giles, "Segregating and Extracting Overlapping Data Points in Two-dimensional Plots", Joint Conference in Digital Libraries (JCDL)-2008.
- Xiaonan Lu, **Saurabh Kataria**, William Brouwer, James Wang, P. Mitra, C. Lee Giles, "Automated Analysis of Images in Documents for Intelligent Document Search" , under final review for publication in International Journal of Document Analysis and Recognition (IJ DAR).
- **Saurabh Kataria**, "On Utilization of Graph images in Digital Documents for Efficient search" appeared in transactions of Joint Conference in Digital Libraries (JCDL)-2008
- William Brouwer, **Saurabh Kataria**, Prasenjit Mitra, Karl Mueller, C. Lee Giles, "Knowledge discovery using data mined from Nuclear Magnetic Resonance spectral images", appeared in Microsoft eScience Workshop-2008.

## **References**

- **Prasenjit Mitra**,  
Asst. Professor,  
College of Information Science & Technology,  
Penn State University, University Park.
- **C. Lee Giles**,  
David Reese Professor,  
College of Information Science & Technology,  
Penn State University, University Park.