Saurabh Kataria

Office Address
312 IST Bldg.
Penn State University, PA 16802
skataria@ist.psu.edu
http://www.personal.psu.edu/ssk164/

Permanent Address
Apt. M-01
445 Waupelani Drive
State College, PA 16801
(814) 876-0852

Objective

To obtain a full time research position that utilizes my research and analytical skills.

Education

- **College of Information Science & Technology, Penn State University.**
  Ph.D., IST, Fall 2006 - till date, Expected Graduation, Fall 2011
  Advisor: Dr. Prasenjit Mitra
- **Institute of Technology, Banaras Hindu University**, Varanasi, India.
  B. Tech, Computer Science and Engineering, August 2001 - June 2005

Research Interest

Applied Machine Learning, Statistical Text Mining, Document Network Analysis, Information Extraction

Research Experience

Research Intern, **Yahoo! Labs, India** Summer 2010
- Supervisors: Dr. Prithviraj Sen and Dr. Rajeev Rastogi
- Project: Entity Disambiguation using Hierarchical Topic Models and Wikipedia

Research Intern, **Xerox Research Center, Europe** Summer 2009
- Supervisors: Dr. Luca Marchesotti and Dr. Florent Perronnin
- Project: Font Retrieval on a Large Scale

Research Asst., **Intelligent Information Systems Lab, PSU** Fall 2007 - Present
- Supervisors: Dr. Prasenjit Mitra and Dr. C. Lee Giles
- Project: Chem^{x}Seer/CiteSeer^x Digital Library for Academic Publications.

Refereed Publications

- **Entity Disambiguation with Hierarchical Topic Models**, Saurabh Kataria, K. Kumar, Rajeev Rastogi, Prithviraj Sen, S. Sengamedu, *17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* **(KDD-2011).**
- **Text Classification using Abstract Features**, Cornelia Caragea, Adrian Silvescu, Saurabh Kataria, Doina Caragea, Prasenjit Mitra, *Symposium on Abstraction, Reformulation, and Approximation* **(SARA-2011).**
- **Context Sensitive Topic Models for Author Influence in Document Networks**, Saurabh Kataria, Prasenjit Mitra, C. Lee Giles, *To appear in 22nd International Joint Conference on Artificial Intelligence* **(IJCAI-2011).**
- **Utilizing Context in Generative Bayesian Models for Linked Corpus**, Saurabh Kataria, Prasenjit Mitra, Sumit Bhatia, *Association for the Advancement of Artificial Intelligence* **(AAAI-2010).**
- **Font Retrieval on a Large Scale: an Experimental Study**, Saurabh Kataria, Luca Marchesotti, Florent Perronnin, *International Conference on Image Processing* **(ICIP-2010).**
- **Generative Models for Authorship Networks**, Saurabh Kataria, Prasenjit Mitra, C. Lee Giles *Workshop on Machine Learning for Social Computing, Neural Information Processing Systems* **(MLSC-NIPS-2010).**

• **Automatic Extraction of Data Points and Text Blocks from 2-Dimensional Plots in Digital Documents**, Saurabh Kataria, William Browuer, Prasenjit Mitra, C. Lee Giles *Association for the Advancement of Artificial Intelligence* **(AAAI-2008).**

• **Segregating and Extracting Overlapping Data Points in Two-dimensional Plots**, William Browuer, Saurabh Kataria, Sujatha Das, Prasenjit Mitra, C. Lee Giles *Joint Conference on Digital Libraries* **(JCDL-2008).**

• **On Utilization of Graph Images in Digital Documents for Efficient search**, Saurabh Kataria, *Graduate Symposium at Joint Conference in Digital Libraries* **(JCDL-2008)**

• **Automated Analysis of Images in Documents for Intelligent Document Search**, Xiaonan Lu, Saurabh Kataria, William Brouwer, James Z. Wang, Prasenjit Mitra, C. Lee Giles *International Journal on Document Analysis and Recognition* **(IJDAR-2008).**

Publications in Progress

• **Large Scale Inference for Topic Modeling Based Context Sensitive Citation Recommendation**, *Work in Progress.*

• **Application of Gleanson Theorem to Document Clustering**, *Work in Progress.*

Patent

• **Dynamic Font Replacement**
Inventors: Saurabh Kataria, Luca Marchesotti, Florent Perronnin, Filed August 2009.

Awards Honors

- Student travel grant awards for KDD'11 (San Diego).
- First runner up for best internship award at Xerox Research Center Europe.
- Student travel grant awards for AAAI'08 (Chicago) and JCDL'08 (Pittsburgh).
- Outstanding summer intern at GlobalLogic Software Ltd. during summer'05
- Ranked among top 1% of 0.2 million students in IIT Joint Entrance Exam., 2001
- State merit award for standing among top 5 students in Senior Secondary Exam.

Research Projects

• **Citation Recommendation (RefSeer):** RefSeer supplements digital library project CiteSeer^x with capability of recommending citations based upon a textual description of the author's scientific interest. RefSeer is content based recommendation system that is built upon extension of Latent Dirichlet Allocation (LDA), i.e., topic models. The inference algorithm learns word-topic and citation-topic association simultaneously and recommends citations that are most probable with input description of scientific interest. The context of citations is used exclusively while learning the associations off-line. Currently, a map reduce based inference algorithm is being implemented to support CiteSeerX scale inference on full document content. **[AAAI'10, NIPS-MLSC'10, IJCAI'11]**

• **Entity Disambiguation using Crowd-Sourced Catalog:** Disambiguating entity references by annotating them with unique ids from a catalog is a critical step in the enrichment of unstructured content. In this project, I focussed upon application of statistical machine learning based models (hierarchical variants of topic models) for the annotation task. The developed approach not only give a coherent way of bridging entity content association to entity id association but also improves upon the existing surrounding context window based entity disambiguation approaches. **[KDD'11]**

• **Relevance Clustering:** Gleason theorem is a fundamental theorem in quantum mechanics that provides a probabilistic connection between a trace class linear operator over a vector space, which is a fundamental quantum mechanics construct known as density matrix, and the probability of any subspace being relevant to the whole vector space. From an information retrieval perspective, Gleason theorem provides a basis for a document being relevant to any construct of feature space. In this project, we consider

clusters as the aforementioned feature space construct, and explore how the documents are clustered into coherent group of similar relevances. **(ongoing)**

• **Font Retrieval on Large Scale:** In this project, I focussed on font retrieval using a query-by-example paradigm: given a font, retrieve the the most visually similar fonts. A font is described by (a) rendering a set of reference characters, (b) extracting a feature vector for each reference character and (c) concatenating character level descriptors. The similarity between two fonts is simply the similarity between the vectorial representations. The main challenge in this project was to extract features that are scale invariant and most descriptive of an underlying font image. The descriptors that were chosen to evaluate were drawn from the literature on typed and handwritten text analysis. [**ICIP'10**]

• **Information Extraction from Scientific Charts:** Findings in scientific studies are usually reported as charts such as 2-D line, bar and pie charts. Digital library search engines such as CiteSeer^x and Chem^{x}Seer tend to under-utilize 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. [**AAAI'08, JCDL'08, IJDAR'08**]

Services

- Reviewer for WWW(08,09,10), KDD(08,09,10), ICDM(08,09),AAAI(08,09,10), CIKM(09,10)

Conference Talks

- JCDL, June 2008
- AAAI, July 2008
- NIPS (MLSC workshop), December 2010.

Skill Set

- Languages: C/C++, JAVA, Perl, Matlab, Shell scripting
- Distributed Computing: MPI, Hadoop (Map/Reduce).
- Machine Learning Packages: Mallet, WEKA, Mahout, Natural Language Toolkit (NLTK).
- Database related: MySQL, SQL-Server and XML.

References

- Available upon Request.