

# Saikat Mukherjee

755 College Road East, Princeton, NJ 08540  
saikat@gmail.com, <http://www.saikatmukherjee.net>

---

- WORK EXPERIENCE**
- ◇ **Siemens Corporate Research**  
Research Scientist, 04/2005 - present
  - ◇ **IBM T.J. Watson Research**  
Intern, 06/2002 - 08/2002
  - ◇ **XSB Inc.**  
Intern, 09/2000 - 10/2001
  - ◇ **Telcordia Technologies**  
Intern, 06/2000 - 08/2000
- EDUCATION**
- ◇ **State University of New York, Stony Brook, NY**  
Ph.D. in Computer Science, 2005
  - ◇ **Indian Institute of Technology, Kharagpur**  
B.Tech. in Computer Science and Engineering, 1999
- AREAS OF EXPERTISE**
- Project management, Applying innovation to practical business problems
  - Applications of Machine Learning to Text Analysis, Computational Linguistics, Knowledge-based Reasoning, Semantic Web
- PROJECTS (SIEMENS)**
- ◇ **Power Semantics**
    - System for analyzing various unstructured text service reports in Siemens Power Generation for more efficient tracking, monitoring, and searching service related activities.
    - Implemented different technologies including statistical NLP-based vocabulary mining, semi-automated ontology creation, unsupervised clustering, automatic text summarization, and semantic search.
    - Implemented Java-based application with flexible interfaces in Prefuse package and database connectivity to Siemens data sources.
    - Supervised 2 software engineers during the project.
    - Key role in marketing the concept to Siemens operating companies, secured funding, and providing support to users of the system.
  - ◇ **Data Cleansing**
    - System for reconciling inconsistencies between various data sources using a combination of learning techniques and rules.
    - Direct the technical course of the project in consultation with end users and engineers.
    - Designed and helped implement key algorithms in the system including normalized edit distances, EM-based technique for learning edit distances, canopy clustering, SVM-based classification, and numerical value reconciliation
    - Supervised 1 software engineer and 1 intern during the project.
    - Key role in introducing data cleaning concept within Siemens, marketing the concept to different Siemens operating companies, and secured funding for project.
  - ◇ **Spend Analysis**

- System for automatic classification of material master descriptions of spend transactions to commodity code hierarchies.
- Designed and implemented algorithms for hierarchical SVM-based classification, incremental classification for training with large data sets, feature weighting using commodity code hierarchy, edit distance based feature correction technique, and supporting multiple languages.
- Implemented Java-based system with interface to the underlying algorithms.
- Supervised 2 software engineers and 2 interns during the project.
- Key role in developing new projects from this concept, secured funding, and providing support to users of the system.
- ◇ **Theseus-Medico**
  - System for semantic understanding and search of medical images.
  - Developed technologies for bringing Semantic Web concepts of ontologies, metadata and their representation in RDF, RDFS, and OWL to medical imaging for semantic image understanding.
  - Developed a system which uses Radlex and FMA as clinical ontologies and Jena as the engine for semantic annotation and retrieval of medical images.
  - Supervised 1 intern during the project.
  - Wrote a significant part of the proposal which was funded by the Theseus search initiative of the German government as a multi-million euro project over 5 years.

PROJECTS  
(GRADUATE  
WORK)

- ◇ **Semantic Annotation of Database-driven Web Content:**
  - Designed algorithms and a Java-based system for automated and scalable semantic annotation of schematic Web pages using a combination of machine learning, domain ontologies, and structural document partitioning.
  - Applied the semantic partitioning technique to help develop a browser, HearSay, for the visually challenged which uses voice XML along with semantic content structuring for easy assistive browsing.
  - Applied the learning-based semantic partitioning technique to develop semantic bookmarks for focused browsing on small screen devices such as PDAs and mobile phones.
  - Applied the learning-based semantic partitioning technique in combination with automata-based process models for automating transactions on complex Web sites.
- ◇ **Segmenting Schematic Text Sequences:**
  - Designed and implemented algorithms for training hidden markov models (HMM) from partially labeled database generated text sequences where certain segments of a training sequence could be labeled while other segments could remain unlabeled.
  - Applied the HMM training technique to learn profile hidden markov models (PHMM) for segmenting protein sequences in bioinformatics.
- ◇ **Data Mining:**
  - Designed and implemented an ontology-based system, Yellowpager, for scraping Web pages to extract and mine service directories. The system was used by HillsPet, a subsidiary of Colgate Palmolive, to mine a directory of veterinarian service providers.
  - Designed and implemented a clustering-based system, CuTeX, for automatically detecting and extracting tabular data in unstructured free text. The system was used by Argus, a leading financial data aggregator, to mine free text for table extraction.
  - Developed an ontology-based process for creating structured content from raw PDF and Web documents. The process was used by Partminer, a leading semiconductor parts supplier, to extract semiconductor part information from PDF and Web documents.

- Designed and developed, in C and Prolog, a database interface for XSB – a logic programming environment. Available for download from <http://xsb.sourceforge.net>.
- PUBLICATIONS:
- BOOK [1] *Automated Semantic Analysis of Schematic Data: Learning-based Techniques for Scalable and Automated Semantic Understanding of Template Generated Schematic Web Content* - Saikat Mukherjee, Publisher VDM Verlag' 2008, ISBN 978-3639026740
- BOOK [2] *Logic based Approaches to Workflow Modeling and Verification* - Saikat Mukherjee, Hasan Davulcu, Michael Kifer, Pinar Senkul, and Guizhen Yang, in *Logics for Emerging Applications of Databases (LEAD)*, Eds. Chomicki, van der Meyden, and Saake, Pub. Springer Verlag, 2003.
- CHAPTER
- JOURNALS [3] *Automated Semantic Analysis of Schematic Data* - Saikat Mukherjee and I.V. Ramakrishnan, in *World Wide Web Journal (WWW)*, 11(4), 2008
- [4] *Model-directed Web Transactions under Constrained Modalities* - Zan Sun, Jalal Mahmud, Saikat Mukherjee, and I.V. Ramakrishnan, in *Journal of the ACM Transactions on the Web, (TWEB)*, 1(3), 2007
- CONFERENCE [5] *Context-Driven Ontological Annotations in DICOM Images: Towards Semantics PACS* - Manuel Moeller and Saikat Mukherjee in *International Conference on Health Informatics (HEALTHINF)*, 2009.
- [6] *Classifying Spend Transactions with Off-the-Shelf Learning Components* - Saikat Mukherjee, Dmitriy Fradkin, and Michael Roth, in *IEEE International Conference on Tools in Artificial Intelligence (ICTAI)*, 2008.
- [7] *Medical Image Image Understanding through the Integration of Cross-Modal Object Recognition with Formal Domain Knowledge* - Manuel Moeller, Michael Sintek, Paul Buiteelaar, Saikat Mukherjee, Xiang Sean Zhou, and Joerg Freund, in *International Conference on Health Informatics (HEALTHINF)*, 2008.
- [8] *Automated Fault-Tree Generation: Bridging Reliability with Text Mining* - Saikat Mukherjee and Amit Chakraborty, in *Annual Reliability and Maintainability Symposium (RAMS)*, 2007.
- [9] *Model-directed Web Transactions under Constrained Modalities* - Zan Sun, Jalal Mahmud, Saikat Mukherjee and I.V. Ramakrishnan, in *International Conference on World Wide Web (WWW)*, 2006. [**Nominated for Best Paper Award.**]
- [10] *Profiling Protein Families from Partially Aligned Sequences* - Saikat Mukherjee, Chang Zhao and I.V. Ramakrishnan, in *SIAM International Conference on Data Mining (SDM)*, 2006 (short paper).
- [11] *Browsing Fatigue in Handhelds: Semantic Bookmarking Spells Relief* - Saikat Mukherjee and I.V. Ramakrishnan, in *International Conference on World Wide Web (WWW)*, 2005.
- [12] *Bootstrapping Semantic Annotation for Content-Rich HTML Documents* - Saikat Mukherjee, I.V. Ramakrishnan and Amarjeet Singh, in *International Conference on Data Engineering (ICDE)*, 2005.
- [13] *Taming the Unstructured: Creating Structured Content from Partially Labeled Schematic Text Sequences* - Saikat Mukherjee and I.V. Ramakrishnan, in *International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE)*, 2004.
- [14] *Semantic Bookmarking for Non-Visual Web Access* - Saikat Mukherjee, I.V. Ramakrishnan, and Michael Kifer, in *ACM Conference on Computers and Accessibility (ASSETS)*, 2004.
- [15] *On Precision and Recall of Multi-Attribute Data Extraction from Semistructured Sources* - Guizhen Yang, Saikat Mukherjee, and I.V. Ramakrishnan, in *IEEE International Conference on Data Mining (ICDM)*, 2003.

[16] *Automatic Annotation of Content-Rich Web Documents: Structural and Semantic Analysis* - Saikat Mukherjee, Guizhen Yang, and I.V. Ramakrishnan, in International Semantic Web Conference (**ISWC**), 2003.

[17] *Automatic Discovery of Semantic Structures in HTML Documents* - Saikat Mukherjee, Guizhen Yang, Wenfang Tan, and I.V. Ramakrishnan, in International Conference on Document Analysis and Recognition (**ICDAR**), 2003.

[18] *Extraction Techniques for Mining Services from Web Sources* - Hasan Davulcu, Saikat Mukherjee, and I.V. Ramakrishnan, in IEEE International Conference on Data Mining (**ICDM**), 2002 (short paper).

[19] *A Clustering Technique for Mining Data from Text Tables* - Hasan Davulcu, Saikat Mukherjee, and I.V. Ramakrishnan, in SIAM International Conference on Data Mining (**SDM**), 2002.

WORKSHOP

[20] *Web Transactions on Handhelds with Less Tears* - Jalal Mahmud, Zan Sun, Saikat Mukherjee and I.V. Ramakrishnan, in WWW Workshop on Empowering the Mobile Web (**MobEA IV**), 2006.

[21] *Learning Semantic Bookmarks for Mobile Handheld Devices* - Saikat Mukherjee and I.V. Ramakrishnan, in ISWC Workshop on Semantic Web technology for Mobile and Ubiquitous Applications (**SWMU**), 2004.

[22] *On the Power of Semantic Partitioning of Web Documents* - Guizhen Yang, Saikat Mukherjee, Wenfang Tan, I.V. Ramakrishnan and Hasan Davulcu, in IJCAI Workshop on Information Integration on the Web (**IIWeb**), 2003.

PROFESSIONAL  $\diamond$  **Program Committee:**

ACTIVITIES

European Semantic Web Conf. (ESWC) 2009, 2008, 2007.

Intl. Symposium on Data, Information, and Knowledge Spectrum (ISDIKS) 2007.

IEEE Intl. Conf. on Data Mining (ICDM) 2006.

Intl. Conf. on Ontologies, Databases, and Applications of Semantics (ODBASE) 2005.

VLDB Workshop on Information Integration on the Web 2004.

$\diamond$  **Journal Reviewer:** Journal of Web Semantics, Data Mining and Knowledge Discovery (DMKD), Multimedia Tools and Applications Journal (MTAP), IEEE Pattern Analysis and Machine Intelligence (PAMI), Information Fusion Journal.

$\diamond$  **Conference Reviewer:** ACM Symposium on Principles of Database Systems (PODS) 2006, International Conference on Logic Programming (ICLP) 2005, Principles and Practices of Declarative Programming (PPDP) 2003.

PENDING

PATENT AP-  
PLICATIONS

[1] *Scalable Semantic Image Search* - Saikat Mukherjee, S. Kevin Zhou, Xiang Zhou, Martin Huber, Jorg Freund, Volker Tresp, Sonja Zillner, Alok Gupta, and Dorin Comaniciu.

[2] *Managing Service Requirements for Airports* - Amit Chakraborty, Saikat Mukherjee, Paul Camuti, and Ramesh Viswanathan.

[3] *System and Method for Integrating Heterogeneous Biomedical Information* - Xiang Zhou, Dorin Comaniciu, Alok Gupta, Zhuowen Tu, Daniel Fasulo, Lu-yong Wang, Saikat Mukherjee, and Amit Chakraborty.

[4] *Method and System for Generating and Validating Clinical Reports with Built-in Automated Measurement and Decision Support* - Peiya Liu, Sridharan Palanivelu, Amit Chakraborty, Dorin Comaniciu, Christoph Dickmann, Sultan Haider, Saikat Mukherjee, Stefan Scholl, Jurgen Vaupel, and Volker Wetekam.

[5] *System and Method for Integration of Medical Information* - Saikat Mukherjee and Amit Chakraborty.