

Saurav Sahay

470 16th St NW, #4029
Atlanta, GA 30363, USA

Phone: 404-488-6882
E-mail: ssahay@cc.gatech.edu
Homepage: www.cc.gatech.edu/~ssahay

EDUCATION

Georgia Institute of Technology, Atlanta, Georgia, USA
PhD Student, Computer Science (Minor – Bioinformatics)
Advisor – Prof. Ashwin Ram
MS Computer Science

Fall 2004 onwards
3.6/4.00

University of Delhi, Delhi, India
Bachelor of Information Technology

Fall 2000 – Spring 2004

EXPERIENCE

Research Intern, Text Mining and Linguistics, IBM T J Watson Research Center

Summer 2006

Mentor: Dr. James Cooper

Developed a semantic search engine that finds clue sentences from biomedical abstracts for finding potential drug targets for genetic abnormalities especially cancer.

Research Intern, IBM India Research Laboratory, New Delhi

Summer 2005

Mentor – Dr. Sougata Mukherjea

Worked on developing and enhancing a Relation Mining system to extract relevant information from the World Wide Web for Biomedical domain. The system performs Natural Language processing (deep parsing) and uses various ontologies like UMLS for answering the queries.

Research Intern, Bioinformatics Centre, Jawaharlal Nehru University, New Delhi India

Summer 2003

Mentor – Prof. Andrew Lynn

Worked on the project ‘Potential Nucleosome Sequence Prediction’ and made programs in hidden markov models for nucleosome prediction. Developed a generic Markov Library for DNA sequence training and prediction.

PROJECTS

Conversational Search for Health Information Access (PhD Thesis Topic)

Spring 2008 onwards

Faculty – Prof. Ashwin Ram

We are developing a system called Cobot (for Community/Conversational bot) that consists of active agents participating in an online community of health information seekers. The agents monitor user conversations with other users in the community and provide personalized as well as community based recommendations to users. The agents dynamically connect users to conversations and other users using a dynamic data structure called socio-semantic net and leverage information from past conversations.

STELLAR (SSTR Tools Enabling Lessons Learned Access and Reachback)

Summer 2009

Faculty – Prof. Ashwin Ram

This project involves developing tools and methods for efficient reachback from lessons learned (LL)/best practices documents and repositories. The goal is to facilitate commissioned forces to carry our Stability, Security, Transition and Reconstruction (SSTR) operations effectively with their limited human, experiential and knowledge based resources.

Medical Rule Learning (MERLIN)

Fall 2006 – Spring 2008

Faculty – Prof. Ashwin Ram, Prof. Eugene Agichtein (Emory), Prof. Ernest V. Garcia (Emory)

MERLIN is a joint research project between CCL, Georgia Tech and Emory University. It aims at exploring how to automatically extract, encode and reason from medical knowledge from published articles and internet resources and create expert system for a Myocardial Perfusion Imaging expert system. (<http://home.cc.gatech.edu/ccl/46>)

GeneTrek

Fall 2004 – Spring 2006

Mentor – Prof. Shamkant B. Navathe

Explored techniques of Information Extraction and machine learning for Text mining of biomedical literature. Worked on feature extraction methods for classification of PubMed documents using supervised learning techniques.

Mining association rules across multiple tables with multiple minimum support

Spring, 2004

Mentor: Dr. Naveen Kumar, Head, Department of Computer Science, University of Delhi

As part of my final semester undergraduate project, worked with a team of researchers at the Department of Computer Science, University of Delhi to implement and extend association rule mining algorithms that worked on a single table and with single minimum support value for finding association rules.

PUBLICATIONS

Journal Articles

3. **Saurav Sahay**, Baoli Li, Eugene Agichtein, Ernest V. Garcia, Ashwin Ram. *Medical Knowledge Identification from Literature Abstracts for updating an expert system*. (Under re-submission)
2. **Saurav Sahay**, Sundaresan Venkatasubramanian, Anushree Venkatesh, Priyanka Prabhu, Bharat Ravisekar, Ashwin Ram. *iReMedI - Intelligent Representation and Reasoning from Medical Information*. (Under re-submission)
1. **Saurav Sahay**, Sougata Mukherjea, Eugene Agichtein, Ernest V Garcia, Shamkant Navathe, Ashwin Ram. *Discovering Semantic Biomedical Relations utilizing the Web*. ACM Transactions on Knowledge Discovery from Data, 2(1):3, 2008

Conference Publications

6. **Saurav Sahay**, Anushree Venkatesh, Ashwin Ram. *Collaborative Information Access: A Conversational Search Approach*. 'Reasoning from Experiences on the Web' Workshop at 8th International Conference on Case based Reasoning. ICCBR 09
5. **Saurav Sahay**, Sundaresan Venkatasubramanian, Anushree Venkatesh, Priyanka Prabhu, Bharat Ravisekar, Ashwin Ram. *iReMedI – Intelligent Retrieval from Medical Information*. ECCBR 08.
4. **Saurav Sahay**, Eugene Agichtein, Baoli Li, Ernest V Garcia, Ashwin Ram.(2007) *Semantic Annotation and Inference for Medical Knowledge Discovery*. 2007 NSF Next Generation Data Mining (NGDM) Symposium, October 10th – 12th, 2007 Baltimore.
3. **Saurav Sahay**, Baoli Li, Ernest V Garcia, Eugene Agichtein, Ashwin Ram.(2007) *Domain Ontology construction from Biomedical Text*. ICAI 2007
2. Sougata Mukherjea, **Saurav Sahay** (2006) *Discovering Biomedical relations utilizing the world-wide web*. Pacific Symposium on Biocomputing 2006
- 1.N. Polavarapu, S. B. Navathe, R. Ramnarayanan, A. Haque, **S. Sahay**, Ying Liu. (2005) *Investigation into Biomedical Literature Classification using Support Vector Machines*. Proceedings of 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005), Stanford University, August 8-11, 2005.

Book Chapters

3. **Saurav Sahay**, Ashwin Ram. *Conversational Search for Health*. “Recent Advances in Healthcare using Case-based Reasoning” Publisher: Springer-Verlag, Germany. (In Preparation, Release: March 2010)
2. S.B. Navathe, **Saurav Sahay**, Hari Prasad. *Information Retrieval*, in Fundamentals of Database Systems, Elmasri, R. and Navathe, S.B., Addison Wesley, 2010 (In preparation)
1. S.B. Navathe, **Saurav Sahay**, . *Emerging Database Technologies and Applications: Genome Data Management* in Fundamentals of Database Systems, Elmasri, R. and Navathe, S.B., Addison Wesley, 2006

Posters

2. Baoli Li, **Saurav Sahay**, Shreekanth Karvaje, Bharat Ravisekar, Joseph Irwin, Neha Sugandh, Cesar Santana, Eugene Agichtein, Ashwin Ram, Ernest V Garcia. *Locating Applicable Knowledge Sentences in Medical Literature*. The 6th Georgia Tech-ORNL International Conference on Bioinformatics poster presentation, 2007.
1. Shreekanth Karvaje, Bharat Ravisekar, **Saurav Sahay**, Baoli Li, Ernest Garcia, Ashwin Ram.(2007) *Discovering Causal Sentences with Automatically Learned Patterns*, ISBRA '07 Poster

GRADUATE COURSEWORK

- CS 7001 Introduction to Graduate Studies
- CS 8803 Data and Information Management
- CS 7641 Machine Learning
- CS 8803 Natural Language Processing
- CS 8803 Probabilistic Graphical Models
- CS 7620 Case based Reasoning
- CS 6400 Database Systems Concepts and Design
- CS 6505 Computability and Algorithms
- CS 8803 Health Informatics
- Biol 4755 Mathematical Biology
- Biol 7023 Bioinformatics
- Biol 4803 Biologically Inspired Design
- Biol 8803 Introduction to Bioinformatics and Genomics
- Biol 6608 Prokaryotic Molecular Genetics
- Chem 6572 Macromolecular Structure
- CS 8803 Foundation of Machine Learning and Data Mining (Audit)
- CS 8803 Discrete Algorithms – CS & E (Audit)
- CS 8803 Cognitive Foundations HCC/LST (Audit)
- CS 8803 Knowledge-Based AI (Audit)
- CS 4803 Computer Science Ventures (Audit)

TEACHING ASSISTANT

- CS 7650 Natural Language
- CS 4440 Emerging Database Technologies and Applications
- CS 6400 Database Systems Concepts and Design

HONORS AND ACTIVITIES

- IBM Worldwide PhD Fellowship Finalist, 2006, 2007
- Best Graduate Machine Learning class project on 'Classification of G-Protein Coupled Receptors, based on their Specific Ligand Coupling Patterns', Prof. Charles L. Isbell, Spring 2005, with Burcu Bakir.
- Wrote proposal and sought GRA Venturelab Phase 1 funding on the Cobot Project.
- Have given guest lectures in graduate and undergraduate classes on Natural Language Processing, Databases and Emerging Database Technologies and Applications.
- Reviewed several conferences and journal papers in AI, Data Mining, Natural Language and Case based Reasoning.
- Worked on development of Research Grant Proposals in Biomedical Knowledge Acquisition, Representation and Reasoning submitted to NIH, NSF, CDC and HSI.
- Travel Grants (NIH and ECCBR) for presenting at Pacific Symposium on Biocomputing, Hawaii 2006 and ECCBR, 08.
- Mentored many undergraduate and graduate students in various research projects.
- Executive Committee Member, Aarohi - Indian Classical Music Society, Georgia Tech 2006-07
- Top 1.5% in Entrance Exam for undergraduate admission conducted by University of Delhi taken by over 40,000 students.
- Represented High School as youngest participant at State Level Programming in BASIC competition, 1996.

Brief Statement:

My primary research interest is in building intelligent systems for informatics applications. My recent work focuses on socio-informatics research, more specifically combining data and information models with social information networks to leverage information access on the web. As part of my thesis, I am developing a real time conversational information access health community that includes an intelligent information agent to scaffold the community activities by pushing relevant recommendations to users of the community. The information agents (cobot, for community/cognitive/collaborative bot) live in the community, are socially aware of users and their preferences, understand the health and medical domain and actively participate in the collaborative information access activity.

I will briefly summarize my major projects and the work involved in my research:

1) Cobot for Healthcare: Community, Collaborative and Cognitive Information Agent

In this project, we are developing a system called *Cobot* (for Community/Conversational bot) that consists of active agents participating in an online community of health information seekers. These agents monitor user conversations with other users in the community and provide personalized as well as community based recommendations to users. One of the goals of this project is to develop an innovative approach to delivering relevant healthcare information using a combination of Web 2.0 social networking and Artificial Intelligence information aggregation techniques. The long-term objective of this research is to facilitate end users efficiently find personalized health-care information using social and intelligent computing techniques.

Keywords: Agent Learning, Recommendation Systems, Case based Reasoning, Natural Language Processing, User Modeling

2) Knowledge Representation and Reasoning

The rapidly increasing volume of unstructured biomedical information poses the challenge of knowledge integration so as to build autonomic computing systems that can acquire, represent and learn such knowledge, and efficiently reason from it to aid in knowledge discovery and re-use. The construction of these automated systems to assist biomedical decision making is impeded by difficulties in formalizing knowledge and in encoding that knowledge for use by computer systems. My research focuses on developing efficient methods of information retrieval and extraction and build a semantic intelligence infrastructure using techniques of language processing, learning and reasoning. This requires automatic construction of knowledge models and ontologies for representing biological objects and processes, as well as methods for expressing hypotheses and 'biological inference rules' that will facilitate their evaluation against what is already known. My work in this area is a joint effort with the Department of Nuclear Cardiology, Emory University as part of *Merlin* (Medical Rule Learning) project.

Keywords: Ontology Learning, Information Retrieval and Extraction, Applied Machine Learning, Natural Language Processing

3) Textual Case based Reasoning

Effective encoding of information is one of the keys to qualitative problem solving. My aim is to explore Knowledge representation techniques that capture meaningful word associations occurring in documents. We have developed *iReMedI*, a TCBR based information access system. For representation we have used a combination of NLP and graph based techniques which we call as Shallow Syntactic Triples, Dependency Parses and Semantic Word Chains. To test their effectiveness we have developed retrieval techniques based on PageRank, Shortest Distance and Spreading Activation methods. The various algorithms developed and the comparative analysis of their results provides us with useful insight for creating an effective problem solving and reasoning system.

Keywords: Knowledge Representation, Language Parsing, Knowledge Re-use and adaptation, Spreading Activation