Rahul Gupta

Contact: K R Building, IIT Bombay, Powai, Mumbai, India 400076
grahul@cse.iitb.ac.in, Ph: +91-9930470980, Fax: +91-22-25720022

EDUCATION

Doctoral candidate, Computer Science and Engineering.
Indian Institute of Technology (IIT), Mumbai. Expected March 2010.
Advisor: Sunita Sarawagi, Department of Computer Science and Engineering.
Title: Training, Inference and Imprecise Storage Techniques for Collections of Conditional Random Fields.

Summary: Conditional Random Fields (CRFs) are the state of the art tools for various structured tasks. My thesis aims at enhancing the power of CRFs by tackling three key technical problems, specifically in the context of information extraction: (a) Joint training of multiple CRFs by exploiting content redundancy among unlabeled sources, and a small seed set of structured records. (b) Joint inference for exploiting rich forms of labeling consistency across data records. This includes designing a framework that supports rich consistency measures, and efficient algorithms that optimize them. (c) Efficient storage of a CRF's output in a query-friendly imprecise data model. This consists of algorithms for efficiently and accurately approximating the CRF's distribution with a mixture of multinomial distributions.

The motivation behind this choice of problems comes from various web tasks that exhibhit sparse user input, yet data sources that enjoy content redundancy.

Topics of interest: Large scale text mining systems involving graphical models, inference algorithms, learning in structured output spaces, domain adaptation, and information extraction.

Bachelor of Technology, Computer Science and Engineering.
Indian Institute of Technology (IIT), Delhi, India, August 2001.

EXPERIENCE

Research Intern Summer 2008
Community Systems Group, Yahoo! Research, Santa Clara, California, USA.
Supervisors: Sathiya Keerthi, Philip Bohannon.

Technical Staff Member June 2001 - Dec 2007
IBM Research Lab, New Delhi, India.
Groups: Knowledge Management, Media Mining.
Managers: Raghuram Krishnapuram, Sugata Ghosal.

Technical Intern Summer 2000
Embedded Systems Group, University of Dortmund, Germany
Supervisors: Rainer Leupers, Peter Marwedel.

PROJECTS

Collective training of multiple CRFs by exploiting content overlap in labeled as well as unlabeled data. Ongoing

The World Wide Tables (WWT) system for answering structured queries using HTML lists and tables on the web. Key issues include robust extraction of relevant records from unstructured web sources, followed by deduplication and ranking. 2008-2009

Lightly-supervised transductive techniques for category-specific attribute extraction from web documents. Summer 2008

More accurate training methods for Conditional Random Fields (CRFs). This included investigation of new loss functions, and efficient approximation of the accurate but inefficient slack scaling method. Fall 2007

A collective CRF-based labeling framework that exploits rich forms of consistency across records during inference. 2007-2008

Maximum a-posteriori (MAP) labeling computation algorithms for clique graphical models with cardinality-based clique potentials. 2006-2007

Efficient creation of imprecise data storage models from CRF-based information extraction models. 2005-2006

Automatic anomaly detection on e-commerce sites, like unusual price dips and impossible product attributes, using one-time limited user input. 2002-2004

Similarity search on multimedia databases incorporating negative relevance feedback from the user. 2001-2002

Efficient online algorithms for maintaining 2-edge connectivity in graphs with dynamic edge insertions and deletions. 2000-2001

PUBLICATIONS **Refereed papers**

Answering table augmentation queries from unstructured lists on the web. With Sunita Sarawagi. In International Conference on Very Large Databases (VLDB), 2009.

Accurate max-margin training for structured output spaces. With Sunita Sarawagi. In International Conference on Machine Learning (ICML), 2008.

Efficient inference with cardinality-based clique potentials. With Ajit A. Diwan and Sunita Sarawagi. In International Conference on Machine Learning (ICML), 2007.

Creating probabilistic databases from information extraction models. With Sunita Sarawagi. In International Conference on Very Large Databases (VLDB), 2006.

Map estimation in mrfs via rank aggregation. With Sunita Sarawagi. In Workshop on Open Problems in Statistical Relational Learning (co-presented with Workshop on Learning in Structured Output Spaces), ICML 2006.

LIPTUS: associating structured and unstructured information in a banking environment. With Manish Bhide, Ajay Gupta, Prasan Roy, Mukesh Mohania, and Zenita Ichhaporia. In ACM Conference on Management of Data (SIGMOD), industrial track, 2007.

Optimal Bitwise Register Allocation using Integer Linear Programming. With Christian Grothoff, Rajkishore Barik, Vinayaka Pandit, and Raghavendra Udupa. In International Workshop on Languages and Compilers for Parallel Computing (LCPC), 2006.

Adaptable Similarity Search using Non-Relevant Information. With T. V. Ashwin and Sugata Ghosal. In International Conference on Very Large Databases (VLDB), 2002. A shorter version : *Leveraging non-relevant images to enhance image retrieval performance*. In ACM International Conference on Multimedia, 2002.

EShopMonitor: A Web Content Monitoring Tool. With Neeraj Agarwal, Rema Ananthanarayanan, Sachindra Joshi, Sumit Negi, and Raghuram Krishnapuram. In International Conference on Data Engineering (ICDE), industrial track, 2004.

Invited papers, working papers, and technical reports

Joint structured models for extraction from overlapping sources. With Sunita Sarawagi. Under review.

Domain adaptation of information extraction models. With Sunita Sarawagi. In Sigmod Record, 2008.

The eShopmonitor: A comprehensive data extraction tool for monitoring Web sites. With Neeraj Agarwal, Rema Ananthanarayanan, Sachindra Joshi, Sumit Negi, and Raghuram Krishnapuram. In IBM Journal of Research and Development, Vol 48, Number 5/6, 2004.

Generalized Collective Inference with Symmetric Clique Potentials. With Ajit A. Diwan and Sunita Sarawagi. arXiv imprint. http://arxiv.org/abs/0907.0589.

Survey of Conditional Random Fields. http://www.cse.iitb.ac.in/~grahul/main.pdf.

PATENTS

US Patent 7487174. Method for storing text annotations with associated type information in a structured data store.
US Patent 7254577. Methods, apparatus and computer programs for evaluating and using a resilient data representation.
Under filing. A transductive approach to category-specific attribute extraction.

REFERENCES

Prof. Sunita Sarawagi, IIT Bombay (`sunita@iitb.ac.in`).
Prof. Soumen Chakrabarti, IIT Bombay (`soumen@cse.iitb.ac.in`).
Dr. Raghuram Krishnapuram, IBM Research, India (`kraghura@in.ibm.com`).