BHAVANA DALVI

5509, Gates Hillman Center
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213.

www.cs.cmu.edu/~bbd/
bbd@cs.cmu.edu

RESEARCH INTERESTS

Information Extraction, Machine Learning, Text Mining, Knowledge Base Population, Semi-supervised Learning, Information Retrieval

EDUCATION

PhD in Computer Science (QPA: 3.84/4) Spring 2015 (expected)
Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, USA
Advisers: Prof. William W. Cohen,
and Prof. Jamie Callan

Master of Technology in Computer Science (CPI: 9.89/10) May 2007
Indian Institute of Technology (IIT),
Bombay, India
Adviser: Prof. S. Sudarshan

Bachelor of Engineering in Computer Science (Percentage: 74.5%) June 2005
Pune University, Maharashtra, India

HONORS

- Google PhD fellowship in Information Extraction 2013-2015.
- Best paper runner-up award, Automatic Knowledge Base Construction Workshop, NAACL 2012.
- Awarded Honorable mention for presentation at LTI Student Research Symposium Aug 2013 and Aug 2014.
- Received student travel award for attending SDM 2013 conference.
- Barbara Lazarus Women@IT Fellowship from CMU for 2009-2010.
- Ranked First among M.Tech. students, Computer Science IIT Bombay.
- Awarded A. K. Doshi charitable trust award for most outstanding student among MTech computer science students at IIT Bombay.
- Awarded the Nilesh Vashee fellowship for best performance in the M.Tech. batch in Fall 2005, Spring 2006 & Fall 2006.
- All India Rank 5 (percentile - 99.98), Computer Science GATE 2005.
- Ranked 4th in Pune University, B.E. Computer Science Examination (2005).
- Affiliated with Poonawala scholarship for excellent academic performance(2005-2007).

RESEARCH EXPERIENCE

Carnegie Mellon University, Pittsburgh, *Research Assistant* with Prof. William W. Cohen (Fall 2009-present)
Thesis Topic: Semi-supervised learning in the presence of unanticipated classes. Traditional semi-supervised learning (SSL) techniques consider the missing labels of unlabeled datapoints as latent/unobserved variables, and model these variables, and the parameters of the model, using techniques like Expectation Maximization (EM). We consider two extensions to traditional SSL methods which make it more suitable for many Automatic Knowledge Base Construction tasks. First, we consider jointly assigning multiple labels to each instance, with a flexible scheme for encoding constraints between assigned labels: this makes it possible, for instance, to assign labels for multiple levels from a hierarchy. Second, we account for another type of latent variable, in the form of unobserved *classes*. In open-domain web-scale information extraction problems, it is an unrealistic assumption that the class ontology or topic hierarchy we are using is complete. Our proposed framework combines structural search for the best class hierarchy with SSL, reducing the semantic drift associated with erroneously grouping unanticipated classes with expected classes. Together, these extensions allow a single framework to handle a large number of knowledge extraction tasks, including macro-reading, micro-reading, multi-view macro- or micro-reading, alignment of KBs to wikipedia or on-line glossaries, and ontology extension.

Google Research, Mountain View, *Intern* with Dr. Anish Das Sarma and Dr. Alon Halevy (Summer 2012), Team: Structured Data.
Research Topic: Human assisted table aggregation tool for consolidating information from semi-structured sources on the Web to populate knowledge-base relations.

Indian Institute of Technology, Bombay, *Master's Student*, with Prof. S. Sudarshan (Fall 2005-Spring 2007)
Research Topic: Keyword Search algorithms for external memory data graphs. The BANKS system is designed to enable keyword search on relational and semi-structured data. It constructs a graph of whole database into main memory and runs search algorithm. To remove memory constraint and generate better results, we have introduced a multi-granular graph structure and efficient use of cache layer in between search algorithm and disk based graph. We also proposed iterative and incremental graph search algorithms to do keyword search using multi-granular graph data structure.

Indian Institute of Technology, Bombay, *Independent Study Project*, with Prof. Sunita Sarawagi (Fall 2006). Project: Workbench for Analysis and Comparison of Graphical Models for Information Extraction

OTHER EXPERIENCE

- Program committee member for AKBC 2014, EMNLP 2014, NAACL HLT 2013, AKBC 2013, EMNLP 2012.
- Reviewer for Journal of Internet and Information Systems 2011, WWW Journal 2011, ECML/PKDD 2013 Journal track, WSDM 2014(secondary reviewer), VLDB 2014(secondary reviewer).
- Teaching assistant for "Analysis of Social Media" (Fall 2012, Spring 2011), CMU.
- Teaching assistant at IIT Bombay for graduate course "Database Systems" (Spring 2007), undergraduate course "Language Processors" (Spring 2006, Fall 2006), and "undergraduate lab for C programming" (Fall 2005), IIT Bombay.

INDUSTRY EXPERIENCE

Google R&D, India, *Software Engineer*, August 2007-August 2009
I developed a query suggest feature for geographic queries to Mapmaker. I also worked on spammer detection and friend recommendation tools for Orkut.

JOURNAL PUBLICATIONS

1. Bhavana Dalvi, Meghana Kshirsagar and S. Sudarshan. “Keyword Search on External Memory Data Graphs”, *Proceedings of the VLDB Endowment (PVLDB) 2008.*

CONFERENCE PUBLICATIONS

2. Bhavana Dalvi, Einat Minkov, Partha Pratim Talukdar, and William W. Cohen, “Automatic Gloss Finding for a Knowledge Base using Ontological Constraints”, *WSDM 2015.*
3. Bhavana Dalvi, and William W. Cohen, “Hierarchical Semi-supervised Classification with Incomplete Class Hierarchies”, *Under submission.*
4. Bhavana Dalvi, and William W. Cohen, “Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances”, *Under submission.*
5. Bhavana Dalvi, William W. Cohen and Jamie Callan, “Exploratory Learning”, *ECML/PKDD 2013.*
6. Ramnath Balasubramanyan, Bhavana Dalvi and William W. Cohen, “From Topic Models to Semi-Supervised Learning: Biasing Mixed-membership Models to Exploit Topic-Indicative Features in Entity Clustering”, *ECML/PKDD 2013*
7. Bhavana Dalvi and William W. Cohen, “Very Fast Similarity Queries on Semi-Structured Data from the Web”, *SDM 2013.*
8. Bhavana Dalvi, William W. Cohen and Jamie Callan, “WebSets: Extracting Sets of Entities from the Web Using Unsupervised Information Extraction”, *WSDM 2012.*
9. Bhavana Dalvi, Jamie Callan and William W. Cohen, “Entity List Completion Using Set Expansion Techniques”, *Text REtrieval Conference, TREC 2010.*

WORKSHOP PUBLICATIONS

11. Bhavana Dalvi, Chenyan Xiong, and Jamie Callan, “A Language Modeling Approach to Entity Recognition and Disambiguation for Search Queries”, *ERD 2014, Entity Recognition and Disambiguation Challenge at SIGIR 2014.*
12. Bhavana Dalvi, William W. Cohen and Jamie Callan, “Classifying Entities into an Incomplete Ontology”, *Automatic Knowledge Base Construction workshop, CIKM’13.*
13. Bhavana Dalvi, William W. Cohen and Jamie Callan, “Collectively Representing Semi-structured Data from the Web”, *Automatic Knowledge Base Construction workshop, NAACL HLT 2012.* ***(Best paper runner-up)***
14. Leman Akoglu and Bhavana Dalvi, “Structure, Tie Persistence and Event Detection in Large Phone and SMS Networks”, *MLG workshop, KDD 2010.*

TALKS

1. “Unsupervised Learning: k-Means and Mixtures”, *Guest lecture in CS-601 Machine Learning*, CMU, October 2014
2. “Automatic Gloss Finding for a Knowledge Base using Ontological Constraints”, *LTI Student Research Symposium*, CMU, August 2014.
3. “Exploratory Learning, semi-supervised learning in the presence of unanticipated classes”, *Google Research*, Mountain View, October 2013.
4. “Exploratory Learning, semi-supervised learning in the presence of unanticipated classes”, *LTI Student Research Symposium*, CMU, August 2013.

Graduate Level Courses

Machine learning, Information retrieval, Analysis of social media, Optimization, Probabilistic graphical models, Language and statistics, Data mining, Web search and mining, Advanced database systems, Algorithms and complexity, Grammars and lexicon, Advanced semantics seminar.