Bhavana Bharat Dalvi
Third year PhD student at Language Technology Institute, CMU.
mail: bbd@cs.cmu.edu
webpage : http://www.cs.cmu.edu/~bbd/

Publications

1. **Collectively Representing Semi-structured Data from the Web,** Bhavana Dalvi, William Cohen and Jamie Callan, to appear in Proceedings of the **NAACL HLT 2012 Workshop** on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction **AKBC-WEKEX 2012**.
2. **WebSets: Extracting Sets of Entities from the Web Using Unsupervised Information Extraction,** Bhavana Dalvi, William W. Cohen and Jamie Callan, Proceedings of the The Fifth ACM International Conference on Web Search and Data Mining, **WSDM 2012.**
3. **Entity List Completion Using Set Expansion Techniques**, Bhavana Dalvi, Jamie Callan and William Cohen, Proceedings of the The Nineteenth Text REtrieval Conference, **TREC 2010.**
4. **Structure, Tie Persistence and Event Detection in Large Phone and SMS Networks**, Leman Akoglu and Bhavana Dalvi, Proceedings of the Eighth Workshop on Mining and Learning with Graphs, **KDD 2010.**
5. **Keyword Search on External Memory Data Graphs**, Bhavana Dalvi, Meghana Kshirsagar and S. Sudarshan, Proceedings of **VLDB 2008**.

Educational Record

PhD	PhD (Language Technology Insti.), CMU, Pittsburgh	2009-present	3.7/4
Post Graduation	M.Tech. (Computer Sci. & Eng.), IIT Bombay	2005-2007	9.89/10.0
Graduation	B.E. (Computer Sci. and Eng.), Pune Institute Of Computer Technology, Pune	2005	74.5 %
		2004	71.0 %
		2003	73.0 %
		2002	78.0 %

Academic Achievements

- 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 IITB.
- 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.
- Affiliated with Poonawala scholarship for excellent academic performance.

Standardized Scores

• GRE – 1410/1600
• TOEFL – 108/120

Research Experience

At Carnegie Mellon University, Pittsburgh (August 2009 - present)

- **Unsupervised Information Extraction to find Sets of Entities from the Web Using semi-structured information :** I work on open-domain information extraction methods for extracting concept-instance pairs from an HTML corpus. Our aim is to extract useful typed information from the corpus corpus, and insert it in an incomplete knowledge-base in totally unsupervised way. Future goal is to use this extracted information to improve quality of keyword search over the corpus.
- **Using entity extraction tools to improve domain specific search**
  I worked on the techniques to combine information extracted using open domain tools to improve information retrieval.
- **Topic modeling techniques applied to research paper corpora**
  This was done as course project for graduate Machine Learning course. We took a subset of citeseer papers in the area of machine Learning. We applied various topic modeling techniques like Latent Dirichlet Allocation (LDA), Correlated Topic Models (CTM) and Link LDA. We could see meaningful topics as a result of this. We could generate a topic graph using the covariance matrix outputted by CTM.

At Google R&D, India (August 2007 – August 2009)

- **Spammer Detection**
  I worked in spam detection team to recognize users with anomalous behavior and taking actions on such profiles.
- **Query Suggest for Maps specific geographic queries**
  It involved suggesting geographic places when a person is typing maps related query. Suggestions were not only based on part of query already typed but also on the viewport in which the person is typing the query and geographic rank of the candidate place.
- **Friend Recommendations**
  I evaluated potential signals for recommending friends to an Orkut user based on common friends and profile similarity.

At IIT Bombay (2005-2007)

- **M.Tech. Project :** *Keyword Search algorithms for external memory data graphs*
  Guide - Prof. S. Sudarshan (May 2006 - June 2007)
  BANKS is a system 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 multigranular 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 multigranular graph data structure.
- **R&D Project :** *Workbench for Analysis and Comparison of Graphical Models for Information Extraction*
  Guide - Prof. Sunita Sarawagi (Autumn 2006)
  In this project, our aim was to come up with metrics for comparing two graphical models. We ranked model's features based on their contribution to the prediction capability. The ranking

schemes were based on expected values of the features and feature weights. This helped in visualizing the robustness of one model over another for different datasets. To demonstrate the idea, we compared simple CRFs and Semi-markov CRFs.

- **M.Tech. Seminar :** *Keyword Search on Relational and XML Data*
  Guide - Prof. S. Sudarshan (Autumn 2005)
  We surveyed existing systems used for keyword search like BANKS, DBXplorer, DISCOVER. We also studied popular ranking techniques like PageRank, XRank, ObjectRank.

Selected Course Projects

- **Implementation of Inference Algorithms on Graphical Models**

I first implemented message passing algorithm for graphical models, assuming edge potentials are given as input. I also implemented gradient based algorithm to make it to learn edge potentials automatically from data.

- **Extension of Data-mining Tool- Weka**

We extended the Weka tool to support new classifier- "DataBoostIM" and filter- "Small-Disjuncts Filter" proposed in recent research papers and they gave better accuracy than the existing algorithms. We also implemented correlation clustering in Weka.

- **Order Optimization for Volcano Optimizer**

We modified Volcano optimizer so that functional dependencies will be inferred and propagated while logical DAG is generated and will be used for order optimization (avoiding unnecessary sort orders), to generate more efficient execution plans.

Other Academic Activities

- Worked as a reviewer for WWW Journal 2011, EMNLP 2012.
- Teaching assistant for “Analysis of Social Media” course under prof. William Cohen
- Teaching assistant(TA) for ‘Database systems’ course under Prof. S. Sudarshan
- TA for undergraduate course “Language Processors” under Prof. Uday Khedkar
- TA for undergraduate lab for C programming
- TA for advance C++ (CEP) course at IIT Bombay
- Tutor for Advanced DBMS workshop 2006, IIT Bombay.

Courses

Social Media Analysis, Graduate Machine Learning, Information Extraction, Optimization, Information Retrieval, Data Mining, Advanced Data Mining (Graphical Models), Web Search and Mining, Web Search and Mining, Algorithms and Complexity, Approximation Algorithms, Databases, Performance Evaluation of Computer Systems and Networks, Advanced Databases

Extra-Curricular Activities

- Participated in a singing at IGSA events 2010.
- Participated in Grad-Cohort 2010 and Grace Hopper conference 2010
- Participated in Google Beat 2007 singing competition.
- Computer Secretary Hostel-11, IIT Bombay.
- Interested in music, singing, painting.