Praveen Kumar Gurumurthy

Contact Information

Computer Science, Purdue University
142 Halsey Dr Apt 7
West Lafayette, IN 47906, USA

Cell: 765-637-1367
HomePage: http://gpraveenkumar.com
E-mail: gpraveenkumar5@ymail.com

Objective

Seeking full-time position for starting in December 2015.

Interests

Machine Learning, Data Science, Social Network Analysis, Data Mining, Information Retrieval, Big Data Analytics, Natural Language Processing, Semantic Web, Text mining.

Education

PhD in Computer Science, Purdue University — Aug 2011 – May 2017 (expected)
Advisor: Dr. Jennifer Neville and Dr. Luo Si — GPA : 3.79/4

Masters in Statistics and Computer Science — Aug 2011 – May 2014
Purdue University — GPA : 3.71/4

Bachelor of Technology in Computer Science and Engineering — Jul 2006 – May 2010
National Institute of Technology, Durgapur (West Bengal, India) — GPA : 9.17/10

Research and Work Experience

Data Scientist Intern at Apple — May 2015 – Aug 2015
Worked at the *iAd Data* team on *User Segmentation and Behavioural Targeting*. Ideated and prototyped a new product - *Lookalike Segments*. Used Latent Semantic Analysis (SVD) to find latent feature/traits among users. Segmented user based on their click behaviour pattern, their relation to Apps and latent features. Used Hive to preprocess data, Python and R for data processing. Built a visualization tool using D3 to qualitatively analyze and compare User and Lookalike Segment. Also implemented the more superior Probabilistic Latent Semantic Analysis technique.

Data Scientist Intern at Linkedin — May 2014 – Aug 2014
Worked with the *Data Sciences* team on *Clustering Fields of Study (Majors)*. Constructed networks of Fields of Study (FoS) using features like member skills, inferred classmates. Detected clusters of FoS using Louvain's Modularity (hierarchical community detection algorithm for graphs/networks) to improve and modify Linkedin's existing FoS taxonomy. Clusters obtained were significantly better than that of traditional hierarchical agglomerative clustering. Used visualization tools like Gephi and D3 to analyze the graph, clusters and taxonomy. Used Apache Pig and Hadoop to preprocess data, Python to do all the data processing and HDFS to store the results.

Research Assistant at CS Department — Aug 2011 – present
Working on Improving Classification Accuracy by Replicating Training Data on Social Networks. Techniques like Marginalized Denoising Autocoders have improved classification performance in the text domain without actually replicating them but by marginalizing over them. Exploring similar approaches to solve problems like label prediction, link prediction in Social Network. Initial results we obtained for label prediction after replicating data by flipping labels, dropping nodes, dropping/rewiring edges are promising.

Teaching Assistant at CS Department — Aug 2014 – May 2015
Spring 2015 - Data Mining (Graduate)
Fall 2014 - Compilers and Programming Systems (Graduate)

Research Intern at Knowledge and Data Engineering, Germany — May 2009 - July 2009
Semantically characterized query term relatedness by analyzing Query Logs and and compared it to query terms in *Bibsonomy*. Wrote Perl Scripts to compute tag-tag, tag-user and user-user similarities. Built a Framework in Java to integrate Perl scripts for easier experimentation.

Undergraduate Research Assistant, NIT Durgapur, India — May 2008 - April 2010
Worked on Data Mining over NonBinary Data Sets. Binary dataset representation gives information about the presence or absence of an item in the search space, but does not provide information about the strength of its presence, which could be more effective in generating association rules close to real life situations. Developed algorithms for mining frequent itemsets and association rules, generating weighted association rules and clustering in non-binary search space.

Selected Projects

Label Prediction in Social Networks using NLP — Mar 15 – Present
Trying to predict the gender, political preferences and religious views of Users(Nodes) on Social

Networks like Facebook. Initially used only network features and techniques like *Gibbs Sampling* for prediction. Started looking at textual features to improve the prediction accuracy. For instance, by using Facebook wall posts of users and their connections, their gender can be predicted with 92% accuracy.

Empirical Analysis of Personal Email Network Nov 13
Constructed and analyzed three different types of ego networks obtained from Gmail consisting of about *seven and half years* of emails. Applied clustering and community detection algorithms to detect communities based of my email communications and compared them with communities detected from my facebook friendship network. Interestingly, I could recover a good number of them.

TREC - Knowledge Base Acceleration Track May 13 – Aug 13
Identified documents related to 140 Wikipedia and 20 Twitter entities that are worthy of citation in their profiles. Filtered and preprocessed, using *Python* and *GnuParallel*, around 6.5 TB of data consisting of social data, news articles etc. Main challenge was that the entities had very few training examples, in the order of 10. Built a model similar to one-vs-all classifier and F1 was close to 0.6.

Supervised LDA for Masquerader Detection Feb 13 – Apr 13
Extended a work of the PhD Thesis of Malek Ben Salem, that builds user-profiles based on search behaviour with a predefined taxonomy of applications and processes to detect masquerader attacks and intrusion detection. Built a novel method by using a variation of LDA to build the taxonomy automatically . Showed that by using the latent classes obtained as feature, we could build classifiers that give the same performance as those that used all the feature, a huge feature space reduction.

Indiana Social Search Feb 12 – Aug 12
Crawled download Google News articles and Tweets and stored them in MySql Databases. Built *Multiclass SVM* classifiers to classify them into predefined categories. Developed a PHP front end to display the results. Set up *Cron Jobs* to repeat the process several times a day. Also, built models to compare the effectiveness of using news articles to classify tweets and vice-versa.

Categorization of Yelp Reviews Sep 12 - Oct 12
Built from scratch in Python, *Naive Bayes* classifiers for Yelp's Academic dataset to classify reviews as Useful, Funny, Cool, Positive. Generated learning curves for the classifiers. Applied smoothing methods and feature selection (using unigram vs. bigram features) to boost classifier performance. Used *K-Means* to cluster data with latitude, longitude, review count and stars as features.

Sampling and Analysis of Social Network Activity Graphs Sep 11 – Dec 11
Constructed email network activity graphs of senders and receivers from the Purdue email data. Sampled data over two day window spans and computed various graph properties like the average degree, density etc. for these windows and the aggregate graph. Compared and contrasted email user activity with those of friendship networks like facebook.

Technical Skills

- Programming languages: C/C++, Python, Java, Perl, PL/SQL, Visual Basic.
- Data analysis and visualization: R, Matlab, Gephi, Lemur and Indri toolkits, RapidMiner, D3.
- Big Data technologies: Hive, Apache Pig, Hadoop, GNUParallel.
- Databases: HDFS, Titan, Oracle, MySQL, IBM DB2, MSSQL.
- Web Development: HTML, JavaScript, PHP, JSP, Ajax, Joomla.
- Tools: GitHub, SVN, Eclipse, NetBeans, Star UML, Adobe Dreamweaver and Flash.

Achievements

- Honoured as the best Graduate Teaching Assistant in 2014-15.
- Received **two scholarships** from *Director, NIT Durgapur* and *NITDAA (NITD Alumni Association)* for my internship at KDE group, University of Kassel, Germany.
- Positioned **1st** in *Open Project*, the Project cum Paper presentation contest in Mukti 10, the Annual Technical Symposium on GNU/Linux and Free Software, 5th - 7th February, 2010.
- **1st** in *Concepts* by IEEE Student Branch, NIT Durgapur for the best project abstract proposed.

Activities

- Treasurer of Computer Science Graduate Student Board, Purdue University. Sep 11 – Aug 13
- Executive co-ordinator Maths N Tech Club, NIT Durgapur Sep 07 – Apr 10