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.

Research 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 Teaching Experience

Research Assistant at CS Department — Aug 2011 – present
Worked on (details in Research Projects Section):

- Label Prediction in Social Networks using NLP
- Improving Classification Accuracy by Replicating Training Data on Social Networks
- TREC - Knowledge Base Acceleration Track
- Supervised LDA for Masquerader Detection

Teaching Assistant at CS Department — Aug 2014 – May 2015
Compilers and Programming Systems (Graduate Level)- Responsible for creating and grading projects and assignments. Facilitating discussion on the forums.

Undergraduate Research Assistant, NIT Durgapur, India — May 2008 - April 2010
Worked on (details in Research Projects Section):

- Data Mining over NonBinary Data Sets
- Compression and Encryption for Secure Communication

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.

Associate IT Consultant at ITC Infotech, India — July 2010 – June 2011
Worked with Product Lifecycle Management team for*Brown Shoes*, a footwear company. Customized software called FlexPLM that runs on WindChill according to client's requirement and built web pages using JSP and JavaScript. Was the best performer amongst campus recruits after boot camp.

Research Intern at Knowledge and Data Engineering, Germany — May 2009 - July 2009
Worked on Semantic Analysis in Query Logs. 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.

Semantically characterized query term relatedness and compared it to the Bookmarks because of their similarities. The results of this work are published at ECML PKDD 2009.

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.

Research 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.
Project Guide: Dr. Dan Goldwasser
Assistant Professor, Department of Computer Science, Purdue University

Improving Classification Accuracy by Replicating Training Data on Social Networks Sep 14 – Present
Improving Classification Accuracy by replicating the training data is a well studied problem in the text domain. Techniques like Marginalized Denoising Autocoders have improved classification performance by marginalizing or taking expectation over the training data without actually replicating them. Similar approaches to solve problems like label prediction, link prediction in the Network Domain have not been explored. Initial results we obtained for label prediction after replicating data by flipping labels, dropping nodes, dropping/rewiring edges are promising.
Project Guide: Dr. Jennifer Neville
Associate Professor, Department of Computer Science, Purdue University

Quantitative Analysis of words and categories in Multiclass Regression Dec 13 – Present
Trying to apply a joint high dimensional Bayesian Variable and Covariance Selection model to the multiclass textual classification. The word features are the variables and hence, variable selection problem corresponds to finding words that are good predictors overall and for specific categories. The covariance selection gives information about dependencies between the multiple categories.
Project Guide: Dr. Anindya Bhadra
Assistant Professor, Department of Statistics, Purdue University

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
Goal was to identify documents related to entities (140 Wikipedia and 20 Twitter) that are worthy of citation in their profiles. Filtered and preprocessed, using *Python* and *GnuParallel*, around 6.5 TB of compressed 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 measure was close to 0.6.
Project Guide: Dr. Luo Si
Associate Professor, Department of Computer Science, Purdue University

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 . Also showed that by using the latent classes obtained from the model as feature, we could build classifiers that give the same performance as those that used all the feature, essentially a huge feature space reduction.
Project Guide: Dr. Seregy Kishner
Adjunct Assistant Professor, Department of Statistics, Purdue University

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.
Project Guide: Dr. Luo Si
Associate Professor, Department of Computer Science, Purdue University

Intelligent tutoring system using Alice Jan 2012 – present
Built a system that would capture code as students write programs in Alice Programming language. Built a tutor out of While module to give live feedback and decide student promotions based on the code captured. Developing a recommendation system to indicate common programming fallacies to prevent a student from the same and to improve the programming experience.
Project Guide: Dr. Luo Si, Dr. Buster Dunsmore & Dr. Steve Cooper
Prof. CS Purdue University, CS Purdue University, CS Stanford University

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.
Project Guide: Dr. Jennifer Neville and Dr. Ramana Rao Kompella
Assistant Professors, Department of Computer Science, Purdue University

Data Mining over NonBinary Data Sets July 08 – May 10
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.
Project Guide: Dr. Anirban Sarkar and Dr. Narayan C Debanath
Assistant Prof.,Dept. of MCA, NIT Durgapur and Professor, Dept. of CS, Winona State Univ.

Compression and Encryption for Secure Communication Jul 09 – Nov 09
The ever increasing internet traffic constantly urges the need for enhancing communication security. So, we developed an algorithm for performing encryption and lossless compression at the same time in order to increase bandwidth utilization and to secure data transmission. We essentially converted the message into a bi-tuple using mapping techniques and encoded only one elements of the tuple.
Project Guide: Mr. Prasenjit Chowdhury and Mr. Jaydeep Howlader
Sr. Lecturer, Department of MCA and Lecturer, Department of IT, NIT Durgapur

Semantic Analysis in Query Log Data May 09 – July 09
Mining for semantic information from search engine query logs bears great potential for both the optimization of search engines and bootstrapping Semantic Web applications. Further, the formalization of log data into Logsonomies retains semantics information. Therefore we analysed and semantically characterized query term relatedness by grounding it to WordNet and compared it to prior results of Folksonomies.
Project Guides: Dr. Gerd Stumme and Dr. Andreas Hotho
Professor and Senior Researcher, Department of EE/CS, University of Kassel, Germany

MULET : A Multilanguage Encryption Technique Mar 09 – Oct 09
The use of a multilingual approach in cryptography was not prevalent. Focused on encryption of plain text over a range of languages supported by Unicode. Used mapping techniques to make the algorithm fast, efficient and easier to implement. Further, the replacement strategy used ensures better security. We believe this will facilitate the localization of Cryptographic Software tools.
Project Guide: Mr. Prasenjit Chowdhury
Sr. Lecturer, Department of MCA, NIT Durgapur

Software Development and Course Projects

Exact inference for a positive Markov random field Mar 13
Implemented maximum cardinality variable elimination ordering to get maximal cliques in a graph. Constructed a clique tree using a maximum spanning tree algorithm. Performed calibration using belief propagation to compute the univariate marginal probabilities for all variables. Given evidence, estimated posterior probability for the rest of univariate posterior marginals. Used R.

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.

Unparser and Interporcedural Analysis on GCC-4.7.0 Sep 12 - Dec 12
Built an unparser that traverses the abstract syntax tree to regenerates the C source code. Performed dataflow analysis on function call graphs to identify the use of uninitialized global and local variables.

Movie Recommendation System Apr 12
Implemented in *Matlab* Memory-Based Collaborative Filtering Algorithm to make recommendations for movies. Also, developed new methods to increase the recommendation accuracy and improve the overall F1 measure.

Text Categorization system for Detecting Email Spam Mar 12
Developed a simple text categorization system for detecting email spam using the Naive Bayes text categorization algorithm and studied the empirical results with a real world email collection. Compared features from relevant (i.e., spam) and in the irrelevant models. Applied smoothing methods and feature selection to improve the categorization performance. Used C++, Lemur Toolkit.

Useful to Usable (U2U) Aug 2011 – Dec 2011
Worked as *Research Assistant* at ITaP (IT at Purdue) at *Rosen Center for Advanced Computing*. Processed climate data collected over 30 years for understanding the Transforming Climate Variability for Cereal Crop Producers. Developed a Joomla component using PHP for generating plots of the variables like temperature, rainfall etc., for a given location and integrated it with Drinet Hubzero.

Software Security Nov 11 – Dec 11
Performed a series of software vulnerability exploits (carried out Buffer Overflow, Format String and return to libc attacks), explored unsafe and insecure programming techniques (like dangers of executable stack) and evaluated the efficacy of operating system defences against them.

MiniDB Database Sep 11 – Dec 11
Starting with a few basic components, developed components of a single-user relational database management system in Java. Specifically, built a Buffer Manager and implemented functionality to evaluate Relational Operators and Queries.

Web Security Sep11 – Oct 11
Carried out a series of attacks like cross site scripting, cross site request forgery, and SQL injection attacks to exploit some vulnerabilities in an online bulletin-board. Implemented modifications to the web site to prevent such attacks. Also performed a brute-force dictionary attack to recover passwords of 100 users using most frequent password list obtained from the Web.

Online Admission System Jan 10 - April 10
Lead a team of 7 members for Software Engineering course project to automate University Admission System. It checks for the inclusion of all required documents and automatically ranks each student's application based on a number of criteria. Developed a website by incorporating necessary logic with HTML and JavaScript as the front end, JSP, Websphere and Oracle Database servers for the backend.

Sphynx Bank Online Aug 08 – Dec 08
Built a website as an online comprehensive solution to Internet Banking for The Great Mind Challenge 2008, a competition conducted by IBM. Used IBM technologies like DB2 and WebSphere for the back end and JSP was used for designing the front end pages.
Project Guide: Mr. Jaydeep Howlader
Lecturer, Department of IT, NIT Durgapur

Publications

1. G. Praveen Kumar, Anirban Sarkar, Ilhyun Lee, Haesun Lee and Narayan C. Debnath, *"A Novel Approach for Hierarchical Clustering in Non - Binary Search Space"*, In the Proceedings of 8th International Conference on Industrial Informatics (INDIN 10), pp. 693 - 697, Osaka, Japan, 13-16th July, 2010.
2. G. Praveen Kumar and Anirban Sarkar, *"Weighted Association Rule Mining and Clustering in Non-Binary Search Space"*, In the Proceedings of the 7th International Conference on Information Technology: New Generations (ITNG 10), pp. 238 - 243, Las Vegas, Nevada, USA, 1214 April, 2010.

3. G. Praveen Kumar, Arjun Kumar Murmu, Biswas Parajuli, and Prasenjit Choudhury, *"MULET : A Multilanguage Encryption Technique"*, In the Proceedings of the 7th International Conference on Information Technology: New Generations (ITNG 10), pp. 779 - 782, Las Vegas, Nevada, USA, 1214 April, 2010.
4. G. Praveen Kumar, Biswas Parajuli, Arjun Kumar Murmu, Prasenjit Choudhury and Jaydeep Howlader, *"A Lossless MOD-ENCODER Towards a Secure Communication"*, In the Proceedings of the International Conference on Recent Trends in Information, Telecommunication and Computing (ITC 10), pp. 330 - 332, Cochin, Kerela, India, 1213 March, 2010.
5. Dominik Benz, Beate Krause, G. Praveen Kumar, Andreas Hotho, Gerd Stumme, *"Characterizing Semantic Relatedness of Search Query Terms"*, In A. Nurnberger, M. Berthold (eds.): Proc. Workshop on Explorative analytics of Information Networks at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2009), pp. 119 - 135, Bled, Slovenia, 11th September, 2009.
6. G. Praveen Kumar, Anirban Sarkar and Narayan C. Debnath, *"A New Algorithm for Frequent Itemset Generation in Non-Binary Search Space"*, In the Proceedings of 6th International Conference on Information Technology: New Generations (ITNG 09), pp. 149 - 153, Las Vegas, Nevada, USA, 2729 April, 2009 ***(Nominated for Best Paper Award)***.

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.
- Adjudged **1st** in *Concepts* by IEEE Student Branch, NIT Durgapur for the best project abstract proposed amongst 40 abstracts.
- Stood **1st** in *The Brand Game* for designing and marketing a Mutual Fund firm in aarohan2k9, a National Level Techno-Management Festival held during 26th February - 1st March, 2009.
- Awarded **2nd** prize in *Konfigure*, the System Administration contest in Mukti 09, the Annual Technical Symposium on GNU/Linux and Free Software, 2nd - 8th February, 2009.
- Judged **3rd** best undergraduate performer by Sun MicroSystems for the project The Ultimate Exam Simulator in Share 2008.
- Secured a place in the **Top 10** among 138 participants in the Network Management Training Program, organized by Goa Institute of Management, Goa and NETTECH INC.

Activities

- Member Computer Science Graduate Student Board, Purdue University.
  - Treasurer — Sep 12 – Aug 13
  - CS Senator to Purdue Graduate Student Government — Mar 12 – Sep 12
  - PhD Representative to the Department — Sep 11 – Mar 12
- Sponsorship Head of Maths N Tech Club, NIT Durgapur — Apr 09 – Apr 10
- Executive co-ordinator Maths N Tech Club, NIT Durgapur — Sep 07 – Apr 09

Relevant Coursework

- Machine Learning
- Social Network Analysis
- Natural Language Processing
- Data Mining
- Information Retrieval
- Social and Economic Networks: Models and Analysis
- Probabilistic Graphical Model
- Mining Massive Datasets
- Introduction to Probability
- Introduction to Mathematical Statistics
- Computational Statistics
- Bayesian Applied Decision Theory
- Big Data in Education
- Computing for Data Analysis
- Statistics One
- Information Security
- Database Systems
- Compiler and Programming Systems
- Algorithm Design, Analysis and Implementation