Yi Yang
Apt 303c, 301 10th St NW
Atlanta, GA 30318
404-263-6322
yiyang@gatech.edu

Date: February 24, 2012

Dear Sir/Madam,

I am a PhD student in School of Computer Science from Georgia Institute of Technology. My research interests are in Information Retrieval, Data Mining, Natural Language Processing and Information Security. I am currently looking for intern positions in your company.

I was introduced to modeling in Machine Learning, Data Mining, Natural Language Processing and Information Retrieval in a NLP lab in Tsinghua University. I have worked on several projects related with Web Mining. I am interested in tweet mining and review mining. The quality of tweets posted every day varies in large range. Predicting the quality of a tweet is very challenging due to its short length, broad topics, and various intents of posting. We proposed several quality measures that have never been studied in ranking short documents, and investigate the influence of difference features. We only sampled a small number of documents in respondence of each query for manual annotation. We proposed several regularization factors to utilize the unlabeled data and to produce prediction more consistently and reliably, and discovered different observations from prior work. Nowadays, texts on the Web have become a valuable source of opinions on products such as reviews. An important issue with reviews is review spam or trustworthiness of reviews. The difficult of this issue is how to define review spam, and we tried to solve it by extricating some features like duplication, similarity and rating information. We proposed a list of elaborately-designed features which characterize the text of the reviews, as well as features that characterize the reviewers who wrote the reviews. Then we utilized the co-training framework to take advantage of the two views (review text and reviewer) of product reviews to detect review spam.

I am also interested in Text Mining. One of the research projects I have involved with patent similarity measurement. We built a claim tree for each patent, and advocated a new model which iteratively adjusted the weight of each node. In the end, approached solving the problem using multiple methods (PLSI, LDA, two-way Poisson Model), but settled on IR techniques primarily due to run-time complexity. In this case, we got a pretty encouraging result. I also addressed multi-entity query segmentation and classification problem in academic search. This work provides a better query understanding function and could make users apply advance search in a single box. The segmentation step using bigram model and classification step using unigram model are processing at the same time. We further propose a new iterative reinforcement model which combines the results attained in segmentation step and classification step respectively. Moreover, the model is not only effective but also efficient.

These days, my work focused on identifying fake followers in Twitter. Twitter users may spend money to purchase their followers, which would damage the social relationships and result in loss of efficacy and accuracy of social network analysis. Moreover, some spammers follow a lot of famous people to propagate spam, which takes a large number of low quality information to Twitter. I addressed this problem by employing graph-based semi-supervised learning methods with features belonging to four groups: profile-based features, graph-based features, behavior features and tweets content features. The project is ongoing, and by this time I am trying to conduct some experiments to prove my approach can achieve better performances than heuristic methods.

As a candidate with strong academic and research background, I am very familiar to the technologies such as Java, Hadoop and machine learning models. I want to utilize my relevant experience in Information Retrieval, Data Mining and Natural Language Processing to improve system performances in your company. I am confident that I am able to make a positive and productive contribution to your team. I am looking forward to hearing from you. Please feel free to contact me. Thank you for your time and attention.

Following are my references with who you can talk about me.

- Prof. Calton Pu
  CERCS, Georgia Tech
  Phone: (404)385-1106
  calton.pu@cc.gatech.edu

- Minlie Huang
  AI Lab, Tsinghua University
  aihuang@tsinghua.edu.cn

- Xiaoyan Zhu
  AI Lab, Tsinghua University
  zxy-dcs@tsinghua.edu.cn

- Jie Tang
  KEG, Tsinghua University
  jietang@tsinghua.eud.cn

Sincerely,
Yi Yang
PhD Student
Georgia Institute of Technology
Phone: 404-263-6322
Email: yiyang@gatech.edu