Shandian Zhe

Experienced in machine learning/data mining research and development

Personal Information

Email:szhe@purdue.edu **Home Page**: http://www.cs.purdue.edu/homes/szhe/
Phone: 219-629-1630 **Address**: 223 Arnold Drive, Apt. 10, West Lafayette, IN 47906

Education

Ph.D. candidate in Computer Science, Purdue University 01/2012 - present
M. Sc. in Computer Science, Chinese Academy of Sciences (CAS) 09/2007-03/2011
B. Eng. in Computer Science in Beijing Univ. of Aeronautics and Astronautics (BUAA) 09/2003-07/2007

Internship Objective

Research and software development in web mining,, social network and machine learning.

Current GPA: 3.87

Related Courses

Statistical Machine Learning: A+ Data Mining: A Algorithm: A Numerical Analysis: A-
Optimization: A Advanced Bayesian Learning: A Randomized Algorithm: A

Internship Experiences

NEC Laboratories America, INC, Research Intern 05/2014-08/2014

- Dynamic adaptive lasso for time series data analysis: The feature weights are learned (instead of being fixed) during the model learning.

Microsoft Research Asia (MSRA), Research Intern 09/2010-03/2011

- Improved focus crawling in Microsoft Academic Search by a weight propagation algorithm for crawling sites ranking.

Work Experience

Assistant Researcher Sogou Search Engine Company 05/2011-12/2011

Research Experiences

Scalable nonparametric multiway data analysis 09/2014 – 11/2014

- Proposed a scalable, nonparametric Bayesian model for large scale multiway data (i.e., tensor) analysis. The model integrates Dirichlet process and local Gaussian process to capture an undetermined number of latent clusters and potential nonlinear relationships between data entries.
- Proposed an efficient online VB-EM algorithm for model learning.
- Evaluated the proposed model on large datasets with billions of elements (DBLP, a scholarship dataset and ACC, a source code management system log) and it shows a significant better predictive performance than the state-of-the-art multiway data decomposition method.

Distributed Bayesian nonparametric tensor decomposition 08/2013 – 12/2013

- Proposed a hierarchical Bayesian model based on Gaussian Process for tensor factorization.
- Devised a distributed online learning algorithm for model estimation (inference).
- Implemented and tested the algorithm with Python under Hadoop platform. The model obtained significantly better prediction accuracy than alternatives like TUCKER and PARAFAC. In large datasets (with billions of tensor entries), the proposed algorithm outperforms a distributed version of PARAFAC--GigaTensor in terms of both running time and prediction accuracy.
- Cooperated with IBM Thomas J. Watson Research Center to use the proposed model to analyze a large log from source code management system..

Multiview learning for association discovery 09/2012 – 05/2013

- Proposed a Bayesian model to combine multiple data views for prediction task and at the same time to extract associations between different data views.
- Applied the proposed model on Alzheimer's disease data and discovered meaningful associations between brain regions and gene fragments (SNPs). The model also obtained better prediction accuracy for Alzheimer's disease status than alternative methods, e.g., lasso, elastic net and SVM.

Network and node selection 01/2012 – 08/2012

- Propose a hybrid Bayesian model to select useful networks and important nodes inside the selected networks. The model uses the structure of networks and selects useful networks (by conditional component) and nodes (by generative component), jointly.
- Applied the proposed model on biological microarray datasets and found critical regulatory pathways and genes related to some cancers. Those findings are supported by published biological literatures.

Parallel conditional random field software 03/2009 – 07/2009

- Developed a CRF software in C++, supporting all graphical structures, all types of features, maximum likelihood and pseudo maximum likelihood training, junction tree inference (i.e., exact inference), loopy belief propagation inference (approximate inference) and multi-thread learning.

Relational network discovery 09/2009 – 03/2010

- Developed a system to construct a relational network for any input entity (e.g., a person or an organization). For example, an input entity "Obama" may result in a network including "White house", "Michelle" and "Obamacare".

Event discovery from search query logs 03/2008 – 07/2008

- Developed a distributed system for event extraction from search query log, based on MPI.
- Extracted news from a one-month search query log and found many interesting news events.

Publications

1. **Shandian Zhe**, Zenglin Xu, Xinqi Chu, Yuan Qi and Youngja Park , Scalable Nonparametric Multiway Data Analysis. Artificial Intelligence and Statistics (AISTATS), 2015.
2. Changying Du, **Shandian Zhe**, Fuzhen Zhuang, Yuan Qi, Qing He, Zhongzhi Shi. Bayesian maximum margin PCA. The 29th AAAI Conference on Artificial Intelligence (AAAI-15), Austin, Texas, USA, January 25-29, 2015.
3. **Shandian Zhe,** Zenglin Xu, Yuan Qi and Peng Yu, Sparse Bayesian multiview learning for simultaneous association discovery and diagnosis of Alzheimer's Disease, AAAI 2015 (Outstanding student paper honorable mention).
4. **Shandian Zhe**, Yuan Qi, Youngja Park, Ian Molloy and Suresh Chari, DinTucker: Scaling up Gaussian process models on multidimensional arrays with billions of elements, S. Zhe, Y. Qi, Y. Park, I.M. Molloy, and S. N Chari, arXiv 2013. A collaborative project between Purdue and IBM T.J. Watson Research Center.
5. **Shandian Zhe**, Zenglin Xu, Yuan Qi and Peng Yu, Joint association discovery and diagnosis for Alzheimer's Disease by supervised heterogeneous multiview learning, in Proceedings of the Pacific Symposium on Biocomputing (PSB), The Big Island of Hawaii, 2014.
6. **Shandian Zhe,** Syed A.Z. Naqvi, Yifan Yang, and Yuan Qi**,** Joint network and node selection for pathway-based genomic data analysis, Bioinformatics, Oxford, 2013.

Skills

1. Excel in C/C++, Python, Matlab, C#; proficient in Web development, Hadoop and Java.
2. Familiar with development in both Linux and Windows environment.
3. Mandarin Chinese (native), English (fluent)