Zhaohui Wu

SUNNYVALE-1020/5321, Microsoft
Phone: +1-814-880-7807 Email: zhaowu@microsoft.com

Summary

- 8+ years' research and engineering experiences on large scale machine learning applications in web search and mining, natural language processing
- Expertise in scholarly/educational information extraction and data mining
    - □ Contribute to Semantic CiteSeerX (one of the world's largest academic search engine)
    - □ Contribute to BBookx (one of the world's pioneering AI powered online learning design system)
    - □ Contribute to Yotta (One of the largest Chinese eLearning platform)

Education

The Pennsylvania State University 08/2011 - 05/2016

-- Ph.D., Department of Computer Science and Engineering
-- Advisor: Prof. C. Lee Giles

Xi'an Jiaotong University 09/2004 - 06/2011

-- Master, Department of Computer Science and Technology, June 2011
-- Advisor: Prof. Qinghua Zheng and Prof. Jun Liu
-- Bachelor, Department of Computer Science and Technology, June 2008

Professional Experiences

Microsoft 05/2016 – present
Applied Scientist Sunnyvale, CA, USA
-- Research and engineering for Bing relevance and intent with specialty on question answering and information extraction

The Pennsylvania State University 08/2011 - 05/2016
Research Assistant State College, PA, USA
-- Research on better understanding natural language using both machine learning and web knowledge
-- Research on information extraction and knowledge mining of scholarly and educational big data

Microsoft Research 05/2014 - 08/2014
Research Intern Redmond, WA, USA
-- Research on novel/emerging entity detection from daily updated news streams and online knowledge repositories

Microsoft Research 05/2013 - 08/2013
Research Intern Mountain View, CA, USA
-- Research on locally defined entity understanding aiming at finding informative passages from long documents

Xi'an Jiaotong University 09/2008 - 06/2011
Research Assistant Xi'an, Shaanxi, China
-- Research on deep web crawling, ill Internet content detection, and information extraction in educational resources

Selected Publications

Zhaohui Wu, Yang Song, C. Lee Giles. Exploring Multiple Feature Spaces for Novel Entity Discovery. **AAAI 2016**, Phoenix AZ, USA.

Chen Liang, Shuting Wang, **Zhaohui Wu**, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, C. Lee Giles. BBookX: Building Online Open Books for Personalized Learning. **AAAI 2016** [Demo], Phoenix AZ, USA.

Shuting Wang, Alexander Ororbia, **Zhaohui Wu**, Kyle Williams, Chen Liang, C. Lee Giles. Using Prerequisites to Extract Concept Maps from Textbooks. **CIKM 2016**, Indianapolis, Indiana, USA.

Madian Khabsa, **Zhaohui Wu**, C. Lee Giles. Towards Better Understanding of Academic Search. **JCDL 2016**, Newark, New Jersey, USA.

Kyle Williams, Jian Wu, **Zhaohui Wu**, C. Lee Giles. Information Extraction for Scholarly Digital Libraries. **JCDL 2016** [Tutorial], Newark, New Jersey, USA.

Zhaohui Wu, Chen Liang, C. Lee Giles. Storybase: Towards Building a Knowledge Base for News Events. **ACL 2015** [System Demo], Beijing, China.

Chen Liang, **Zhaohui Wu**, Wenyi Huang, and C. Lee Giles. Measuring Prerequisite Relations Among Concepts. **EMNLP 2015**, Lisbon, Portugal.

Zhaohui Wu, C. Lee Giles. Sense-aware Semantic Analysis: A Multi-prototype Word Representation Model using Wikipedia. **AAAI 2015**, Austin, TX, USA.

Wenyi Huang, **Zhaohui Wu**, Chen Liang, Prasenjit Mitra and C. Lee Giles. A Neural Probabilistic Model for Context Based Citation Recommendation. **AAAI 2015**, Austin, TX, USA.

Zhaohui Wu, Dayu Yuan, Pucktada Treeratpituk, C. Lee Giles. Science and Ethnicity: How Ethnicities Shape the Evolution of Computer Science Research Community. arXiv:1411.1129 (2014).

Zhaohui Wu, Yuanhua Lv, Ariel Fuxman. Searching Locally-Defined Entities. **CIKM 2014**, Shanghai, China.

Zhaohui Wu, Wenyi Huang, Chen Liang, C. Lee Giles. Crowd-sourcing Web Knowledge for Metadata Extraction. **JCDL 2014**, London, UK.

Zhaohui Wu, Jian Wu, Madian Khabsa, Kyle Williams, Hung-Hsuan Chen, Wenyi Huang, Suppawong Tuarob, Sagnik Ray Choudhury, Alexander Ororbia, Prasenjit Mitra and C. Lee Giles. Towards Building a Scholarly Big Data Platform: Challenges, Lessons and Opportunities. **JCDL 2014**, London, UK.

Wenyi Huang, **Zhaohui Wu**, Prasenjit Mitra and C. Lee Giles. RefSeer: A Citation Recommendation System. **JCDL 2014**, London, UK.

Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernandez-Ramirez, Hung-Hsuan Chen, **Zhaohui Wu** and C. Lee Giles. CiteSeerX: A Scholarly Big Dataset. **ECIR 2014**, Amsterdam, Netherlands.

Zhaohui Wu, Zhenhui Li, Prasenjit Mitra, C. Lee Giles. Can Back-of-the-book Indexes be Automatically Created? **CIKM 2013**, San Francisco, CA, USA.

Zhaohui Wu, Sujatha Das, Zhenhui Li, Prasenjit Mitra, C. Lee Giles. Searching Online Book Documents and Analyzing Book Citations. **DocEng 2013**, Florence, Italy.

Zhaohui Wu, Prasenjit Mitra, C. Lee Giles. Table of Contents Recognition and Extraction for Heterogeneous Book Documents. **ICDAR 2013**, Washington, DC, USA.

Zhaohui Wu, C. Lee Giles. Measuring Term Informativeness in Context. **NAACL-HLT 2013**, Atlanta, GA, USA.

Qinghua Zheng, **Zhaohui Wu**, Xiaocheng Cheng, Lu Jiang, Jun Liu. Learning to Crawl Deep Web. ***Information Systems***, 38(6):801-819, 2013.

Jun Liu, Lu Jiang, **Zhaohui Wu**, Qinghua Zheng, Yanan Qian. Mining Learning-Dependency between Knowledge Units from Text. **The *VLDB Journal***, 20(3): 335-345, 2011.

Jun Liu, Lu Jiang, **Zhaohui Wu**, Qinghua Zheng. Deep Web Adaptive Crawling based on Minimum Executable Pattern. ***Journal of Intelligent Information Systems***, 36(2): 197-215, 2010.

Zhaohui Wu, Lu Jiang, Qinghua Zheng, Jun Liu. Learning to Surface Hidden Web content. **AAAI 2010**. Atlanta, GA, USA.

Zhaohui Wu, Lu Jiang, Qinghua Zheng, Jun Liu, Zhenhua Tian, Junzhou Zhao. A Peep at Pornography Web in China. **WebSci 2010**. Raleigh, NC, USA.

Lu Jiang, **Zhaohui Wu**, Qian Feng, Jun Liu, Qinghua Zheng. Efficient Deep Web Crawling Using Reinforcement Learning. **PAKDD 2010**. Hyderabad, India.

Patents

A Feature Dictionary Generating Method for Text Classification based on LZW Compression Algorithm. Publication Number: ZL200810232557, Jun. 2010.

Searching Locally Defined Entities. Publication Number: WO2015168344 A1, Nov. 2015.

Systems and Methods for News Event Organization. Publication Number: US20160188590 A1, Jun. 2016.

Projects

Semantic CiteseerX

08/2015~ 05/2016 *Research Assistant (with Prof. Lee Giles)*
To build a scholarly knowledge base for next generation CiteseerX; key work includes: crawl open scholarly knowledge resource; define semantic schemas and extract scholarly knowledge triples from various resources.

Bbookx: bring open educational resource for teaching and learning

09/2014~05/2016 *Technical Mentor, System/Algorithm Designer*
To build a system that allows users to dynamically create course materials, such as textbooks, using open educational resource based on advanced web search algorithms. URL: https://book.tlt.psu.edu/; http://bbookx.psu.edu/
Media coverage: Tech Times, TechRepulic, Center for Digital Education, Education World, Penn State News, The Daily Collegian Online, T.H.E. Journal, EdSurge, etc.

News event detection

10/2013~ 05/2015 *Research Assistant (with Prof. Lee Giles)*
To study news event detection from large scale web news and social media and to build event knowledge base using 30+ years' global event data for exploring news event chains and story timelines.
URL: http://breckenridge.ist.psu.edu:8000/storybase/

Academic Book Search

08/2011~05/2012 *Research Assistant (with Prof. Lee Giles)*
An open book search engine that indexed more than 59,000 books, with title, author, ISBN, Date, Copyright, bibliography extracted from PDF. Collective metadata extraction and book citation analysis are conducted on dataset from this project. URL: http://sundance.ist.psu.edu:8080/solr1/index.html

Deep/Hidden Web Crawling and Integration

11/2008~05/2011 *Self-support (collaborate with Lu Jiang)*
MEP-based adaptive crawling method and a Reinforcement Learning Framework were proposed respectively. Released DWIM (Deep Web Intelligent Miner), an open source deep web mining platform.

Ill Internet Content Monitoring System

09/2008~10/2010 *National Key Technology Research Project (Research Assistant)*
Work on the Identification and Countercheck of Ill Internet Content; Research on Pornographic Web Page Recognition and Filter based on Text, URL and Web Page Structure (partly supported by Microsoft Research Asia).
URL: http://netculture.xjtu.edu.cn
Media coverage: DANWEI, Gbtimes, Hastac, Factsanddetails, etc.

Yotta: A Novel Cloud-based E-learning Platform

03/2009~07/2010 *National High-Tech R&D Program (Research Assistant)*
Work on knowledge mining and management of Yotta: research on mining Learning-dependency between knowledge units from educational resources (e.g. textbooks, wikipedia, etc); be responsible for the knowledge map datasets construction.

Professional/Community Services

Conference PC Member: AAAI2017, WSDM2017, ScholarlyBigData@IJCAI2016, BigScholar2015; LTI-SRS2015;

Journal Reviews: Plos One, Journal of Intelligent Information Systems, Journal of Engineering and Computer Innovations, Computer and Information Science, Computer Methods and Programs in Biomedicine

Organizer: Penn State SIG Computational Informatics Seminar; CVPR2013 workshop on Symmetry Detection

Reviewer/Sub-reviewer: WWW2017, DocEng2016, TPDL2016, NAACL2016, WWW2016, AAAI2016, WWW2015, JIIS, LAS2015, WSDM2015, DocEng2015, EMNLP2014, CIKM2014, SIGIR2014, WWW2014, CIKM2013, WWW2013, SIGIR2013, JCDL2013, TPDL2013, SIGIR2012, JCDL2012, CIKM2012, TPDL2012

Invited Talks: Microsoft Research 2013; Xi'an Jiaotong University 2015; Stanford University 2016