WARREN SHEN

Department of Computer Science, Office 4354
1210 West Dayton Street
Madison, WI 53706

Telephone: 217-377-3564
Email: whshen@cs.wisc.edu
Homepage: http://pages.cs.wisc.edu/~whshen

RESEARCH INTERESTS

Applying database, Web, and AI techniques to data management problems with a focus on information extraction and integration for managing unstructured data.

EDUCATION

Ph.D. Computer Science, **University of Illinois at Urbana-Champaign** (expected) *Summer 2009*
Advisor: Professor AnHai Doan
Dissertation Title: "Toward an Integrated Information Extraction Management System"

M.S. Computer Science, **University of Illinois at Urbana-Champaign** *2005*
Advisor: Professor AnHai Doan

B.S. Computer Science, **Stanford University**, Stanford, CA *2002*

RESEARCH AND WORK EXPERIENCE

Research Intern, University of Wisconsin, Madison *Sept. 2008 – present*
Continuing dissertation research on building information extraction systems. Studied the benefits and limitations of leveraging relational database system technology for information extraction (submitted for publication).

Research Assistant, University of Illinois, Urbana-Champaign *Aug. 2003 – Aug. 2008*
Conducted dissertation research on information extraction and integration, key challenges in managing unstructured data.

- Developed Xlog, a declarative language for information extraction, and extended relational query optimization techniques to automatically optimize Xlog programs. Extended Xlog to support best-effort information extraction.
- Developed solutions to leverage data source properties and integrity constraints to maximize entity matching accuracy, a fundamental problem in information integration.

Intern, Yahoo! Research, Santa Clara, CA *April 2007 – Aug. 2007*
Supervisors: Raghu Ramakrishnan and Philip Bohannon
Worked on the design and development of PSOX, a platform to support declarative and extensible information extraction.

Intern, Stanford Biology Department, Stanford, CA *June 2001 – Sept. 2001*
Supervisor: Professor Virginia Walbot
Designed and built a Web application to design DNA segments suitable for microarray experiments in the Maize Gene Discovery Project.

Intern, Framfab, Sunnyvale, CA *June 2000 – Feb. 2001*
Designed and built a Java software simulation of a network of wireless devices.

SOFTWARE

One of two main developers of DBLife (http://dblife.cs.wisc.edu), a Web portal for the database research community that automatically extracts and integrates community-related information and news from the Web. Significantly contributed to every core component of the system, with a special focus on extraction and integration components.

TEACHING EXPERIENCE

Teaching Assistant, University of Illinois at Urbana-Champaign *Fall 2003*
CS 105: Introduction to Computing with Application to Business
Taught a weekly lab section, designed and graded exams and homework, and conducted weekly office hours.

Teaching Assistant, University of Illinois at Urbana-Champaign *Fall 2002, Spring 2003*
CS 173: Discrete Math
Designed and graded exams and homework, conducted review sessions, and conducted weekly office hours.

Section Leader, Stanford University *Spring 2000 – Spring 2001*
CS 106: Programming Methodology and Abstractions
Taught a weekly section, helped students in weekly one-on-one meetings, and graded homework and exams.

PROFESSIONAL ACTIVITIES

External Reviewer for Fnt 2008, VLDB 2008, JAIR 2007, IJCAI 2007, NGITS 2006, ICDE 2006, WebDB 2005, VLDB 2005, SIGMOD 2005, ASIAN 2005
Volunteer at KDD 2007
Member of SIGMOD, ACM, IEEE

PUBLICATIONS

Publications in Rigorously Reviewed Conferences

1. **Warren Shen**, Robert McCann, Pedro DeRose, Raghu Ramakrishnan, AnHai Doan. “Toward Best-effort Information Extraction”. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pp. 1031-1042, Vancouver, Canada, June 2008. (78/435=17.9% accepted)
2. Robert McCann, **Warren Shen**, AnHai Doan. “Matching Schemas in Online Communities: A Web 2.0 Approach”. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pp. 110-119, Cancun, Mexico, April 2008. (119/617=19.2% accepted)
3. Pedro DeRose, Xiaoyong Chai, Byron Gao, **Warren Shen**, AnHai Doan, Philip Bohannon, Jerry Zhu. “Building Community Wikipedias: A Human-Machine Approach”. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pp. 646-655, Cancun, Mexico, April 2008. (75/617=12.1% accepted)
4. **Warren Shen**, AnHai Doan, Jeffrey Naughton, Raghu Ramakrishnan. “Declarative Information Extraction Using Datalog with Embedded Extraction Predicates”. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pp. 1033-1044, Vienna, Austria, September 2007. (45/276=16.3% accepted)
5. Pedro DeRose, **Warren Shen**, Fei Chen, AnHai Doan, Raghu Ramakrishnan. “Building Structured Web Community Portals: A Top-Down, Compositional, and Incremental Approach”. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pp. 399-410, Vienna, Austria, September 2007. (45/276=16.3% accepted)
6. **Warren Shen**, Pedro DeRose, Long Vu, AnHai Doan, Raghu Ramakrishnan. “Source-aware Entity Matching: A Compositional Approach”. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pp. 196-205, Istanbul, Turkey, April 2007. (122/659=18% accepted)
7. **Warren Shen**, Xin Li, AnHai Doan. “Constraint-Based Entity Matching”. In *Proceedings of the National Conference in Artificial Intelligence (AAAI)*, pp. 862-867, Pittsburgh, Pennsylvania, July 2005. (148/803=18% accepted)

Publications in Journals

8. AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, **Warren Shen**, Ba-Quy Vuong. “Information

Extraction Challenges in Managing Unstructured Data". In *ACM SIGMOD Record, Special Issue on Managing Information Extraction, December 2008*.

9. AnHai Doan, Philip Bohannon, Raghu Ramakrishnan, Xiaoyong Chai, Pedro DeRose, Byron J. Gao, **Warren Shen**. "User-Centric Research Challenges in Community Information Management Systems". In *IEEE Data Engineering Bulletin, Special Issue on Data Management Issues in Social Sciences,* 30(2):32-40, 2007.
10. AnHai Doan, Raghu Ramakrishnan, Fei Chen, Pedro DeRose, Yoonkyong Lee, Robert McCann, Mayssam Sayyadian, and **Warren Shen**. **"**Community Information Management". *In IEEE Data Engineering Bulletin, Special Issue on Probabilistic Data Management, 29(1)*:64-72, 2006.

Other Publications

11. AnHai Doan, Jeffrey Naughton, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, **Warren Shen**, Ba-Quy Vuong. "The Case for a Structured Approach to Managing Unstructured Data". To appear in *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2009.
12. Pedro DeRose, **Warren Shen**, Fei Chen, Yoonkyong Lee, Douglas Burdick, AnHai Doan, Raghu Ramakrishnan. "DBLife: A Community Information Management Platform for the Database Research Community" (Demo). In *Conference on Innovative Data Systems Research (CIDR)*, pp. 169-172, Asilomar, California, January 2007.
13. Robert McCann, Alexander Kramnik, **Warren Shen**, Vanitha Varadarajan, Olu Sobulo, AnHai Doan. "Integrating Data from Disparate Sources: A Mass Collaboration Approach" (Poster). In *Proceedings of the International Conference on Data Engineering (ICDE)*, pp. 487-488, Tokyo, Japan, April 2005. (100/521=19% accepted)
14. Robert McCann, **Warren Shen**, and AnHai Doan. "Collective Integration of Information for Virtual Organizations". In *Proceedings of the SIGMOD Workshop on Databases In Virtual Organizations (DIVO)*, pp. 9-16, Paris, France, June 2004.

ADDITIONAL INFORMATION

Citizenship: U.S.A.

REFERENCES

Available upon request.