

# Fei Chen

Email: [fei.chen4@hp.com](mailto:fei.chen4@hp.com)

---

---

## QUALIFICATIONS

- Project management experiences in establishing and maintaining excellent customer relationship in both China and US, developing reasonable project scope with deliverable results and leading team to accomplish planned goals.
- Extensive research experiences and solid knowledge in Data Mining, Databases, Information Extraction, Information Retrieval, NLP and Machine Learning.
- Hands-on experiences in building large-scale information management systems and data mining tools.
- Published 10+ papers at top international conferences and journals in computer science, with totally 400+ citations.
- Strong collaboration relationships with both US and China universities.
- Excellent communication skills of presenting information concisely to a wide range of audience.
- Strong capability of working both independently and adaptively within a team.

## WORKING EXPERIENCE

### ***Project Leader, HP Labs China***

May 2011-now

#### ***Project: Real-time System Log Monitoring and Analysis***

- Met with the customer team to understand the background of both teams and develop a 2-year project plan which aligns with customer business plan and will result in great research values to HP Labs.
- Developed project and research agenda which is target for 2-3 researchers in 2 years.
- Presented the project agenda to the CTO and VP of the customer and received positive feedback. They decided to list the project with us as one of the top 10 projects for them in Year 2012.

### ***Project Leader, HP Labs China***

May -Aug 2011

#### ***Project: Query Log Analysis for Enterprise Portals***

- Communicated with this internal customer to understand the project background and their system architecture and develop a practical solution which exploits query logs to improve search experiences of HP.COM portal.
- Led a team of 2 to develop, implement and evaluate the solution, and build a demo based on our solution.
- Showed demo to the customer and collected feedback and deployment requirements.
- Successfully delivered the code which is being deployed at HP.COM portal.

### ***Project Leader, HP Labs China***

Aug 2010- now

#### ***Project: Context-Aware Enterprise Search***

- Led a team of 3 to meet with the internal customer to understand the current product and its limitations.
- Proposed a project which aims to systematically exploit a rich set of contexts available in enterprises to improve enterprise search.
- Developed and implemented a solution which exploits product profiles, stored as structured data, to improve keyword search over unstructured data.

- Delivered the solution to customers and helped them develop several applications based on the techniques delivered. Received very encouraging feedback from their Senior VP after showing the demo.
- Identified several research problems from this project. Working in collaboration with University of Texas, University of Delaware and Hong Kong University of Science and Technology separately on 3 research problems.

**Research Intern, IBM Almaden Research Center**

2008

**Project:** *Efficiently Mining Large-scale Enterprise Data on High Performance Clusters.*

- Designed distributed algorithms to efficiently mine a large-scale set of documents for enterprise intranet search.
- Implemented the distributed mining algorithms using Hadoop, a software framework supporting distributed computations.
- Among the first to develop solutions to association rule mining over Hadoop platform.

**Research Assistant, University of Wisconsin-Madison**

2005 – 2010

**Advisor:** *AnHai Doan and Raghuram Ramakrishnan*

**Project:** *Optimizing Information Extraction (IE) over Large-scale Evolving Text Data*

- Proposed a paradigm of reusing previous IE efforts to execute an IE program efficiently over large-scale evolving text corpora such as Wikipedia.
- Designed, implemented and evaluated Cyclex, a scalable system that correctly and efficiently executes an IE program repeatedly over evolving text corpora.
- Designed, implemented and evaluated Delex, a scalable system that extends Cyclex to complex IE workflows that consist of multiple IE programs.
- Published 2 papers at top Database Conferences on the research results.

**Project:** *DBLife*

- One of the 3 main architects for [DBLife \(http://dblfe.cs.wisc.edu/\)](http://dblfe.cs.wisc.edu/), a structured Web portal that automatically extracts, manages and analyzes community-related information and news from Web for the database research community.
- The system has helped spawning off PSOX, a sister project on community information management at Yahoo! Research.
- Designed and implemented algorithms of extracting entity relationships and events from Web.
- Designed and implemented a mass-collaboration service that allows community users to participate in maintaining the portal in a Web 2.0 manner.

**Consultant, California Institute of Technology, Economy Department**

2003 – 2004

**Project:** *Computer Testbed for Studying the Scalability of Learning Algorithms in Large Games.*

- Constructed test bed for studying the performance of learning algorithms (e.g. reinforcement learning) in large games for studying economy phenomena such as action and bargains.
- Guided team and its leader in selection of appropriate experimental comparison strategies by machine learning expertise.

**Research Assistant, Simon Fraser University**

2002 – 2004

**Advisor:** *Martin Ester*

**Project:** *Prediction of Outer Membrane Proteins (OMP)*

- Designed and implemented a highly accurate predictor of OMPs. The predictor can achieve a rate of 98% precision and 81% recall (*much higher than 64% precision and 71% recall achieved by the state-of-the-art method in the biological domain*).
- Proposed and implemented the method of Support Vector Machine based prediction, which incorporates biology domain knowledge.
- Participated in building an online-server, [PSORTb \(http://www.psorth.org/psorth/\)](http://www.psorth.org/psorth/), the most precise bacterial localization prediction tool available (by April 1, 2005).

**Project: Interpreting Support Vector Machine (SVM) Models.**

- Pioneer the first work in interpreting SVMs in high dimensional data space.
- Designed a rule-based learning algorithm for interpreting trained SVMs; in particular SVMs trained in high dimensional data space.
- Implemented the newly designed algorithm and conducted experiments in text and biology sequence data.

## **EDUCATION**

**PH.D. in Computer Science**, University of Wisconsin-Madison, USA, 2010

Advisor: AnHai Doan and Raghu Ramakrishnan

Minor: Statistics

Dissertation: Optimizing Information Extraction over Large-scale Evolving Text Data.

**M.S. in Computer Science**, Simon Fraser University, Canada, 2004

Advisor: Martin Ester

Master's Thesis: Learning Accurate and Understandable Rules from Support Vector Machine Classifiers.

## **PROFESSIONAL ACTIVITIES**

- **External Reviewer** for International Conference on Very Large Databases (VLDB) 2006, 2009; ACM's Special Interest Group on Management Of Data Conference (SIGMOD) 2006; International Conference on Extending Database Technology (EDBT) 2010.
- **Secretary** of ACM's Committee on Women UW student chapter, 2006-2010
- **Member** of ACM, IEEE, SIGMOD

## **STUDENT ADVISED**

- Yan Zhang, Shanghai Jiao Tong University
- Aaron Feng, University of Wisconsin-Madison
- Xitong Liu, University of Delaware
- Xingtian Shi, University of Tokyo
- Ruihua Sun, Shanghai Jiao Tong University

## **Recent Talks ( 2010-2011)**

- Towards Collaborative and Automatic Enterprise Knowledge Management, *invited talk, May 2011, Sun Yat-sen University*
- A Structured-Centric Approach to Manage Unstructured Data, *invited talk, Dec 2010, Renmin University*
- Optimizing Information Extraction over Large-Scale of Text, 2010, *University of Illinois-Urbana and Champaign, Johns Hopkins University, Washington University-St Louis*

## **PUBLICATIONS**

### **PAPERS IN REFEREED CONFERENCES**

1. **F. Chen**, A. Feng, C. Re, M. Wang, "Optimizing Statistical Information Extraction Programs over Evolving Text", To appear in *Proceedings of the 28<sup>th</sup> IEEE Int. Conf. on Data Engineering (ICDE), Washington DC, USA, 2012.*
2. **F. Chen**, B. J. Gao, A. Doan, J. Yang and R. Ramakrishnan, "Optimizing Complex Extraction Programs over Evolving Text data", In *Proceedings of the 35<sup>th</sup> SIGMOD Int. Conf. on Management of Data (SIGMOD)*, Pages 321-334, Providence, Rhode Island, USA, 2009. (63/397=15.9% Accepted)

3. A. Doan, J.F. Naughton, A. Baid, X. Chai, **F. Chen**, T. Chen, E. Chu, P. DeRose, B. Gao, C. Gokhale, J. Huang, W. Shen, B. Vuong, “The Case for A Structured Approach To Managing Unstructured Data”, In *Proceedings of the 4<sup>th</sup> Biennial Conference on Innovative Data Systems Research (CIDR)*, Asilomar, CA, USA, 2009.
4. **F. Chen**, A. Doan, J. Yang and R. Ramakrishnan, “Efficient Information Extraction over Evolving Text Data”, In *Proceedings of the 24<sup>th</sup> IEEE Int. Conf. on Data Engineering (ICDE)*, Pages 943-952, Cancun, Mexico, 2008. (75/617 = 12.1% Accepted)
5. P. DeRose, W. Shen, **F. Chen**, A. Doan and R. Ramakrishnan, “Building Structured Web Community Portals: the Case for an Incremental and Compositional Approach”, *Proceedings of the 33<sup>rd</sup> Int. Conf. on Very Large Data Bases (VLDB)*, Pages 399-410, Vienna, Austria, 2007. (45/276 = 16.3% Accepted)
6. P. DeRose, W. Shen, **F. Chen**, Y. Lee, D. Burdick, A. Doan and R. Ramakrishnan, “DBLife: A Community Information Management Platform for the Database Research Community”, *Proceedings of the 3<sup>rd</sup> Biennial Conference on Innovative Data Systems Research (CIDR)*, Pages 169-172, Asilomar, CA, USA, 2007.
7. R. She, **F. Chen**, K. Wang, M. Ester, J. L. Gardy, F. S.L. Brinkman, “Frequent-Subsequences-Based Prediction of Outer Membrane Proteins” ,*Proceedings of the 9<sup>th</sup> ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, Pages: 436 – 445, Washington, D.C., USA, 2003. (46/298 = 15% Accepted)
8. R. She, **F. Chen**, K. Wang, M. Ester, J. L. Gardy, F. S.L. Brinkman, “Identifying Bacterial Outer Membrane Proteins using Frequent Subsequences – A Data Mining Approach”, *Proceedings of the 14<sup>th</sup> Int. Conf. on Intelligent Systems for Molecular Biology (ISMB-2003)*.

#### **PAPERS IN REFEREED JOURNALS**

9. A. Doan, J.F. Naughton, A. Baid, X. Chai, **F. Chen**, T. Chen, E. Chu, P. DeRose, B. Gao, C. Gokhale, J. Huang, W. Shen, B. Vuong, “Information Extraction Challenges in Managing Unstructured Data”, *SIGMOD Record*, Winter 2008.
10. A. Doan, R. Ramakrishnan, **F. Chen**, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen, “Community Information Management”, *IEEE Data Engineering Bulletin, Special Issue on Probabilistic Databases*, 29(1), 2006. Invited Paper.
11. J.L. Gardy, M.R. Laird, **F. Chen**, S. Rey, C.J. Walsh, M. Ester, and F.S.L Brinkman, “PSORTb v.2.0: Expanded Prediction of Bacterial Protein Subcellular Localization and Insights Gained from Comparative Proteome Analysis”, *Bioinformatics*, 21(5), 2004.

#### **OTHER PUBLICATIONS**

12. **F. Chen**, “Learning Accurate and Understandable Rules from Support Vector Machine Classifiers”, *Master Thesis, School of Computing Science, Simon Fraser University, 2004.*

## **COMPUTER SKILLS**

- **Programming language:** Java, C++, Perl, Matlab, Mathematica and R
- **Environment:** Linux and Windows

## **REFERENCES**

Available upon requests.