XIN WANG

PERSONAL INFORMATION:

Name	Xin Wang
Email	Xinwang-2010@hotmail.com

EMPLOYMENT HISTORY:

DATA RESEARCH, INOME(PREVIOUSLY, INTELIUS INC)

PRINCIPAL *SCIENTIST (07/2010-NOW)*

I work on building machine learning models to solve the entity resolution problem of linking billions of public records from different sources together into hundred of millions of single people profiles. My job functions including:

1. Build new record linkage models for our data builds
2. Link professional/business profiles from sources like linkedin and zoominfo to public records
3. Relative classification
4. Data collection and management with AMTurk

Other minor stuff: wrote the pipeline to sample and transform raw data, generate sql databases, and send HITS to Turks and load their feedback into judgments that can be used by our training algorithm. I also wrote our own Turker management system--keeping records, qualification and adjustment, error reporting, training examples, bonusing Turkers and other stuff.

Participated in the KBP Cold Start Task 2012.

BING SEARCH, MICROSOFT CORP

Software Design Engineer (10/2008-06/2010)

Work on Extensible Answer Platform team for Bing Search Platform. The team is responsible for developing a framework and maintaining the platform to run all Bing services.

My work at the team involves different aspects of our platform:

1) Query understanding:
   a. Developed Query Analysis Service using our framework. The service provides classification results for queries to downstream services in real time. The service was written in C++.
2) Query triggering:
   a. Enriched the pattern-based (grammar) query triggering mechanism for our framework with dictionary, synonym and other functionalities (also in C++).
   b. Developed a prototype (neural network based) to provide the deterministic grammars used by the framework with probabilitistic matching capacities. I did all of the design, implementation, training and tweaking of the accuracy, and the evaluation work.
   c. Researched on using probabilistic grammars on the query triggering for individual services;

    d. Worked on the Bender project which developed an easy-to-use tool for our partners especially non technical ones to develop and evaluate their own services (C++ and SQL).

3) Query serving:
    a. Worked on execution plan for gallery services and categorical search (C++)
    b. Worked on the definition, parsing of query augmentation related to query refinements.
4) Wire Format—Bing's equivalent of Protocol Buffer by Google or Thrift by Facebook:
    a. Implemented unmanaged readers/writers for our managed libraries for the wire format (C#);
    b. Implemented XPath style random access readers/writers for the managed libraries (C#);
    c. Implemented versioning checking for schema validation (C++);
5) Performance Analysis:
    a. Analyzed performance bottleneck of our wire format with F1 Profiler;
    b. Analyzed the impact of garbage collections of C# on the performance of the wire format;
    c. Improved the implementation of the C# libraries for the wire format to minimize the influence of garbage collections on the performance of the libraries with FxCop and CLR Profiler.
6) Query forensics:
    a. Building statistical models for the performance counters of our distributed system (Iron Python, and Sho);
    b. Using statistics to understand the correlation of performance counters, and various run time issues of the distributed system (Perl, Iron Python, Excel, Sho to process the data; running SQL like jobs to collect query log data from virtual cloud like clusters);
7) Scope job prediction:
    a. This was a hack-day project inside Bing team. Our project was selected as the second one of the four winners out of 80 projects presented. I built the statistical model (Rep-Tree) for the prediction tool with Weka, and then wrote the libraries to call the learned model from C# code.
8) DRI responsibilities (by rotation, one week in about 2 months):
    a. Monitoring and solving live site issues;
    b. Investigating and fixing build, unit tests, and acceptance test breaks;
    c. Pushing BRS to deploy QFEs for partners;
    d. RI, FI code from and to main branches;

TABLET PC INK PARSING TEAM, MICROSOFT CORP

Software Design Engineer (8/2004 - 10/2008)

I was the team resident expert on machine learning, and my job responsibilities involved:

a. Developed ground-breaking ink parsing technologies (shipped as part of Windows Vista, Windows 7, and Office 14):
    - Shape recognition engines for diagramming;
    - Container-connector recognition engines for diagramming;
    - Ink Annotation parsing engines for diagramming;
    - Using EM algorithm to re-estimate the priors and transition probabilities for the Hidden Markov Model used for writing and drawing classification;
    - Publishing and evangelizing the technologies developed by the team;
    - Serving as reviewers and program committee members at professional workshops.
b. Research on new technologies for ink document parsing (collaborating with researchers from Microsoft Research):

- 2-d conditional random fields for recognition of container and connector;
- Hierarchical conditional random fields for the analysis of lists in ink documents;
- Probabilistic grammars for analysis of lists in ink documents;

c. Functioned as a consultant to teammates on how to use machine learning or pattern recognition techniques to solving problems in their work:
   - Line grouping engine project;
   - Block grouping engine project;
   - Using recognition engines for correction user interface;

d. Owner of the team's computational geometry libraries:
   - Improved the libraries by implementing new algorithms, e.g. the rotating calipers based algorithms I implemented improved some of the basic geometric operations used by the team from O(nlogn) to O(n);
   - Fixed various bugs related to the libraries including bugs related to floating point computation, and degeneration problems in geometry;

e. Owner of the team's machine learning libraries, responsibilities include redesigning the libraries, maintaining them during day to day work, porting them from C# to C++
   - AdaBoost trees
   - CRF
   - Neural networks
   - Hidden Markov models of the team's Computational Geometry libraries

f. Worked on internal labeling tool for ink documents:
   - Worked with program managers and labelers on labeling scenarios and guidelines;
   - Implemented features for the labeling tool used by the labelers and the team;
   - Implemented features for analyzing and debugging geometric algorithms visually;

OREGON STATE UNIVERSITY

Graduate Research Assistant (7/1997 - 9/2004)

Work on developing new algorithms in value function approximation and policy gradient reinforcement learning for solving large scale combinatorial optimization problems such as job-shop scheduling problem. (C++, Unix, Splus, Matlab, and ILog CPLEX, Linux Clusters);

IBM T.J. WATSON

Intern (6/1999 - 9/1999)

Work with the operation research team on using constraint programming methods to solve combinatorial optimization problems encountered by the steel manufacturing industry (Prolog, and Eclipse for constraint programming).

EDUCATION HISTORY:

OREGON STATE UNIVERSITY-UNITED STATES-OR

Computer Science, PH.D-Doctorate, 5/2006

Thesis Advisor: Thomas G. Dietterich (Founding President of International Machine Learning Society)

Thesis Title: Model Based Approximation Methods for Reinforcement Learning

GPA: 3.8

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Computer Engineering

MS-Masters of Science 7/1996

NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY

Major: Computer science and engineering

BS-Bachelor of Science

PATENTS:

Patents (mostly still in disclosure and applications phase):

1. Parsing Ink Annotation (2006)

2. Classifying Digital Ink into Writing and Drawing (2007)

3. Shape Recognition using Partial Shapes (2008)

4. Engine Support For Parsing Correction User Interfaces (2007)

PUBLICATION LIST:

AT INOME:

Wang X., Kardes H., Sun A., Chen L., Aggrawal S., Borthwick. A. Probabilistic Estimates for Attribute Statistics and Match Likelihood for People Entity Resolution (Submitted to VLDB 2014)

Kardes H., Agrawal S., Wang X., and Sun A. (2014) CCF: Fast and Scalable Connected Computation in MapReduce, IEEE International Conference on Computing, Networking and Communications, 2014

Sun A., Wang X., Xu S., Kiran Y., Shakthi P., Borthwick A., and Grishman, R. (2012) Intelius_NYU Cold Start System, KBP 2012

AT MS:

Wang, X. and Sashi, R. (2007) Ink Annotations and their Anchoring in Heterogeneous Digital Documents Paper, ICDAR 2007

Wang, X., Biswas, M. and Sashi, R. (2007) Addressing Class Distribution Issues of the Drawing/Writing Classification in an Ink Stroke Sequence, Sketch-based Interface and Modelling Workshop 2007, (Paper selected for reprise at SIGGRAPH 2007, and invited for publication as a full-length journal paper in the Journal of Computer Graphics.

Wang, X. Shilman, M and Sashi, R (2006). Parsing Ink Annotations On Heterogeneous Documents, Sketch-based Interface and Modelling Workshop 2006

IN GRADUATE SCHOOL:

Wang, X. Model Based Approximation Methods for Reinforcement Learning (2006) Ph.D thesis. (To be published by Verlag Dr Muller Publishing)

Wang, X. and Dietterich, T. G. (2003). Model-based Policy Gradient Reinforcement Learning. International Conference on Machine Learning, ICML-2003, Washington, DC, 776-783.

Dietterich, T. G. and Wang, X. (2002). Batch value function approximation via support vectors. In Dietterich, T. G., Becker, S., Ghahramani, Z. (Eds.) Advances in Neural Information Processing Systems 14. (pp. 1491-1498). Cambridge, MA: MIT Press.

Wang, X. and Dietterich, T. G. (2002). Stabilizing value function approximation with the BFBP algorithm. In T. G., Becker, S., Ghahramani, Z. (Eds.) Advances in Neural Information Processing Systems 14. (pp. 1587-1594). Cambridge, MA: MIT Press.

Wang, X., Dietterich, T. G. (2000). Efficient value function approximation using regression trees. Pages 51-54 of collective article: J. Boyan, W. Buntine, and A. Jagota (Eds.), Statistical Machine Learning for Large Scale Optimization. Neural Computing Surveys, 3, 1-58.

Wang, X., Dietterich, T. G. (1999). Efficient Value Function Approximation Using Regression Trees. In Proceedings of the IJCAI Workshop on Statistical Machine Learning for Large-Scale Optimization, Stockholm, Sweden.

Wang, X. From Simplex Methods to Interior-Point Methods: A Brief Survey on Linear Programming algorithms (1999) (http://citeseer.ist.psu.edu/old/352670.html)