Baojun Su

+86(512)18625240866
freizsu@gmail.com

Do everything seriously

Educational Background

2008 – 2011 **Master**, *Zhejiang University*, Computer Application Technology.

Research Area: Large scale online learning, text classification, collaborative filtering algorithm.

2004 – 2008 **Bachelor**, *Jiangsu University of Science and Technology*, Information and Computing Sciences.

Work Background

2012.9 – **Software Develop Engineer**, *Microsoft STC*, Suzhou.
Distributed key-value datastore, campaign recommendation.

2011.4 – 2012.8 **Applied Research Engineer**, *Netease Youdao*, Beijing.
Crawler, web page parser and analyzer, ranking.

Technical Ability

Programming Language: Java, Python, C, C#, Matlab, SQL.

Database Technology: SQL Server, Mongodb, Kyoto Cabinet, self implemented distributed key-value datastore.

Profiling: Algorithm profiling, service profiling.

Large-Scale Data Processing: Experienced with big data processing technology, familiar with several stacks including CoWork/ODFS/OMap in Youdao and Cosmos in Microsoft.

Data Mining: Familiar with common models and algorithms of data mining, has in-depth study of online learning, text classification and collaborative filtering algorithms.

English: 471 points in CET-6, good English reading and speaking ability.

Projects

Distributed Key-Value Datastore

Introduction: Online serving data in TB and at the same time with low latency, with the ability to plugin user logic as SPROC.

Timeline: 2013.10 – 2014.9

Duty: System design, part of implementation, performance tuning.

Results: Compared to the old system hosted on SQL Server, the .95 has been reduced to 1/3 to 1/10 with the same request, and at the same time increased the flexibility and the readability of SPROC.

Product Recommendation Service

Introduction: Recommend inventories to advertisers in display advertising business.

Timeline 2012.9 – 2013.4

Duty Experiments and algorithms implementation including content-based and collaborative filtering algorithms, and common utility such as evaluation, profiling and instrumentation.

Results The feedbacks from users are mostly positive, the accuracy improved from 66% to 73%.

Timeliness Web Crawler

Introduction Crawl news from portal sites, forums, microblogs with low latency.

Timeline 2011.10 – 2012.4

Duty Project leader with three team members, architecture design and implementation.

Results The timeliness coverage rate has been catching up with Baidu after this system going online, compared to nearly no timely results before.

Efficiency Optimization of DNS Resolve

Introduction The DNS lookup latency has been becoming the bottleneck of main crawler's efficiency and we try to speedup it.

Timeline 2011.4 – 2011.6

Duty Performance optimization.

Results The DNS lookup latency has been sped up by above 10 times.

Terminator

Introduction A two-layer ensemble spam filter, open sourced at github.

url `https://github.com/freiz/terminator`

Timeline 2009 – 2010

Results Implemented my ensemble algorithm and ensemble 8 advanced classifiers, can achieve the best results on *All* public datasets.

Experience

Open Source

2010 NSNB spam filter won the first place in large-scale spam filtering competition of Eights Symposium of Search Engine and Web Mining, which is implemented in Python and open sourced at `http://code.google.com/p/nsnb`.

Publications

- Baojun Su, Congfu Xu. Not so naïve online Bayesian spam filter. In: Proceedings of the 21st conference on Innovative Application of Artificial Intelligence (IAAI 2009), July 14-16, 2009, Pasadena, CA, pages 147-152.
- Congfu Xu, Chunliang Hao, Baojun Su. Research on Markov logic networks. Chinese Journal of Software, 2011, 22(8): 1699-1713. (In Chinese with English abstract)