DATA SCIENTIST, TECH DIRECTOR & DBA OF CITESEERX

Career Overview

Ten years work experience under Linux/Unix environment; Latex and MS Excel. Â· Five years work experience of building/maintaining production MySQL databases and Apache Solr, debugging and optimizing ETL work flows, based on scholarly big data. Â· Five years work experience of search engine architecture and infrastructure, deploying and implementing web application features Â· Five years work experience of designing, coding, and testing LAMP website powered by MySQL databases and Apache Solr, using frameworks such as Django and Spring. 1 Update on February 10, 2017 Â· Five years programming experience with Python; familiar with load balancing, virtual environment, firewall (e.g., iptables), and file systems. Â· Three years work experience of managing software projects on open source software platforms, e.g., GitHub. Â· Two years experience of analyzing logs using MapReduce; Deep Learning architectures of RNN and CNN on video data; experience with Amazon AWS, Microsoft Azure Cloud, Google Cloud, and Google Analytics; Experience with NLP tools, Bash, Java, R, Ruby on Rails, RESTful API. Â· Backgrounds in Physics, Math, and Statistics; Familiar with ML, NLP, ANN, IR, and genetic algorithms.

Work Experience

06/2013 to Current

Data Scientist, Tech Director & DBA of CiteSeerX State Of Louisiana ï¼ Ville Platte , LA

- I started with the web crawling module of CiteSeerX in 2011, then expanded to the full architecture around 2013.
- My job duties include administrating the MySQL database and Apache Solr index servers, hacking the source code (Python/Java/Perl) to fix security vulnerabilities, developing new web application features, managing 100+ terabytes production and research data, maintaining 30+ physical and virtual servers to facilitate production and research, and developing software to improve web crawling, information classification and extraction.
- By the end of 2014, I was able to run the entire search engine single handed.
- In 2015, I proposed infrastructure and software solutions to overcome scalability bottlenecks and blueprinted the next generation of CiteSeerX.
- By the end of 2016, I had scaled the data collection from 3 million to over 10 million documents.
- Currently, the system can keep running for several months without major issues.
- The 200+ page system document wrote by me significantly flattens learning curve for new admins.
- I used to assist 3+ professors to build private cloud and GPU infrastructure.
- I also have experience of working on a Hadoop cluster, and programming with MapReduce.
- Post-doctoral Scholar June 2011 - present.

06/2006 to 05/2011

Research Assistant Decatur Public Schools ï¼ Decatur , IL

- Utilize astronomical big data, compiled from archives of space- and ground-based telescopes, such as the Hub- ble Space Telescope and the Sloan Digital Sky Survey to investigate important correlations between physical parameters of Active Galactic Nuclei and quasars.
- Publish 7 peer reviewed journal articles.

08/2004 to 05/2006

Teaching Assistant Astronomy & Astrophysics, Pennsylvania State University ï¼ City , STATE

- Lecture non-science college students on astronomical fundamentals.

Education and Training

August, 2011

Ph.D : Astronomy and Astrophysics Pennsylvania State University ï¼ City , State , USA Astronomy and Astrophysics

Ph.D : Computational Science Computational Science

July 2004

B.S : Physics and Astronomy University of Science and Technology of China Hefei China Physics and Astronomy

Interests

Entity Recognition in Scientific Document Ongoing Leader Research Â· Recognize and extract semantic domain knowledge entities from scientific documents Video Compression with ANN Ongoing Co-leader Research Â· Perform near-lossless video compression using artificial neural network models Migrating CiteSeerX to a Private Cloud Published in 2014 Leader System Â· Migrate CiteSeerX production servers to a private cloud with virtualization techniques Document Classification in Digital Libraries Published in 2014 and 2016 Co-leader Research Â· Automatically and accurately classify PDF documents with ML and structural features PUBLICATIONS Â· See http://fanchyna.wixsite.com/jianwu/pubs for all publications. OTHER INFORMATION Â· PC members of 5 conferences/workshops Â· Reviewers for 14 top-tier conferences/journals/transactions, including WWW, SIGIR, and TKDE Â· Collaborated with people from UNT, Microsoft, AllenAI, and Internet Archive 2 Update on February 10, 2017

Skills

Apache, AI, big data, conferences, content, data collection, Database, features, Hub, Java, managing, MySQL, NLP, next, search engines, page, PDF, Perl, programming, proposals, publications, Python, research, scientific, servers, developing software, teaching, typing, articles

Additional Information

- HONORS AND AWARDS Best paper nomination in the 8th International Conference on Knowledge Capture 2015 Best application paper in the 26th Annual Conference on Innovative Applications of Artificial Intelligence 2014 Best paper nomination in the IEEE International Conference on Cloud Engineering 2014 Zaccheus Daniel Fund 2009 Zaccheus Daniel Fund 2007 Stephen B. Brumbach Fellowship 2006 USTC Excellent Graduate Student Award 2004 SELECTED PROJECTS Entity Recognition in Scientific Document

Ongoing Leader Research Â· Recognize and extract semantic domain knowledge entities from scientific documents Video Compression with ANN Ongoing Co-leader Research Â· Perform near-lossless video compression using artificial neural network models Migrating CiteSeerX to a Private Cloud Published in 2014 Leader System Â· Migrate CiteSeerX production servers to a private cloud with virtualization techniques Document Classification in Digital Libraries Published in 2014 and 2016 Co-leader Research Â· Automatically and accurately classify PDF documents with ML and structural features PUBLICATIONS Â· See http://fanchyna.wixsite.com/jianwu/pubs for all publications. OTHER INFORMATION Â· PC members of 5 conferences/workshops Â· Reviewers for 14 top-tier conferences/journals/transactions, including WWW, SIGIR, and TKDE Â· Collaborated with people from UNT, Microsoft, AllenAI, and Internet Archive 2 Update on February 10, 2017