Xuhong Zhang

📱 +1 (770) 576 0251 • ✉ xuhongnever@gmail.com • 🌐 www.xuhongzhang.com/

Education

University of Central Florida **2013–Fall 2017**
PhD in Computer Engineering, GPA 3.94,

Georgia State University **2011–2013**
Master in Computer Science, GPA 3.84,

Harbin Institute of Technology **2007–2011**
Bachelor in Software Engineering, GPA 84/100,

Skills

- 5+ years experiences in Java and Python.
- Sampling theory, Approximate analytics.
- Common machine learning algorithms.
- 3+ years experience in developing distributed systems.
- Hadoop/MapReduce, Spark, Hama, HDFS.
- SQL, Linux, Data Structures, Algorithm.

Research projects

Enabling efficient approximations on sub-datasets in Hadoop **2015-Now**

- I developed a system call Sapprox to enable both efficient and accurate approximations on arbitrary sub-datasets of a large dataset. Sapprox does not cache offline samples. Instead, we developed a probabilistic map to estimate the occurrences of a sub-dataset at each logical partition of a dataset (storage distribution) in the distributed system, and make good use of such information to facilitate online sampling. The speedup over existing systems is up to $20\times$. <github.com/zhangxuhong/SubsetApprox>

Reversible deterministic block management for HDFS **2014-2015**

- To reduce the memory and maintenance overhead of HDFS' table based block management, we replace it with a reversible deterministic block management. Given a HDFS block, its locations can be mathematically calculated. Given a node, the blocks on it can also be reversely calculated. Our method is expected to double the capacity of current Hadoop clusters.

Minimizing communication delay in Apache Hama via vertex categorization **2014**

- To minimize the communication delay in Apache Hama, we prototyped a new system called Zebra. Zebra implements a runtime computation and communication scheduler to overlap computation in the next superstep with communication in the current superstep. Zebra can achieve average 2× speedup over Hama. <github.com/zhangxuhong/Zebra>

Vision-based web page segmentation and bids information retrieval **2012-2013**

- Developed for Online Data Services, LLC in Atlanta. A new web page segmentation algorithm is proposed. The main block of a page and the bids in it are automatically detected. <github.com/zhangxuhong/WebPageSegmentation>

Selected publications

[1] Xuhong Zhang, Jun Wang, and Jiangling Yin. Sapprox: Enabling efficient and accurate approximations on sub-datasets with distribution-aware online sampling. ***Proc. VLDB Endow.***, 10(3), 2016.

[2] Jun Wang, Jiangling Yin, Jian Zhou, Xuhong Zhang, and R. Wang. Datanet: A data distribution-aware method for sub-dataset analysis on distributed file systems. In *2016 IEEE International Parallel and Distributed Processing Symposium (**IPDPS**)*, pages 504–513, May 2016.

[3] Jun Wang, Xuhong Zhang, Jiangling Yin, Huafeng Wu, and Dezhi Han. Speed up big data analytics by unveiling the storage distribution of sub-datasets. ***IEEE Transactions on Big Data***, 2016.

[4] Jun Wang, Xuhong Zhang, Junyao Zhang, Jiangling Yin, Dezhi Han, Ruijun Wang, and Dan Huang. Deister: A light-weight autonomous block management in data-intensive file systems using deterministic declustering distribution. *Journal of Parallel and Distributed Computing (**JPDC**)*, 2016.

[5] Xuhong Zhang, Ruijun Wang, Xunchao Chen, Jun Wang, Tyler Lukasiewicz, and Dezhi Han. Achieving up to zero communication delay in bsp-based graph processing via vertex categorization. In *Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on*, pages 112–121. IEEE, 2015.