Approximate Computing for Big Data Analysis

Dr. Joshua Zhexue Huang

Distinguished Professor
Director of Big Data Institute
Shenzhen University, China
zx.huang@szu.edu.cn

December 8 (Wednesday), 17.00 (Novosibirsk) / 18.00 (Shenzhen).
Zoom-link: https://us02web.zoom.us/j/83189630550?pwd=S0dmTTNkNnFGcjdlV0trZWprQk9mQT09
Meeting ID: 831 8963 0550. Passcode: 984871.

Abstract

In the era of big data, datasets with millions of objects and thousands of features have become a phenomenon in many organizations. Such datasets, often in the size of hundred gigabytes or even terabytes, can easily exceed the size of the memory of the cluster systems, creating computing problems in big data analysis. Therefore, how to effectively processing and analyzing terabyte big data with limited resources is both a theoretical and technical challenge in current big data research.

In this tutorial, we will discuss the issues of distributed data computing with a particular focus on approximate computing for big data. I will start with a general introduction to big data and challenges in big data analysis, and continue with discussions of current technologies used in big data analysis and their shortcomings. Then, I will introduce approximate computing for big data and a new method that uses multiple random samples to compute approximate results of big data. Finally, I will present the new technologies and algorithms to enable approximate computing, including the random sample partition (RSP) data model, the LMGI computing framework and the algorithm to generate the RSP data models from HDFS big data files. LMGI is a non-MapReduce framework that allows execution of serial algorithms independently on local nodes or virtual machines without data communications among the nodes. The new technologies present the

following breakthroughs in big data computing: analyzing big data without memory limit, executing serial algorithms directly in distributed computing, and extending the scalability of data analysis to the scale of terabytes on small clusters.

Biography

Dr. Joshua Zhexue Huang is a distinguished professor at College of Computer Science and Software Engineering and the founding director of Big Data Institute of Shenzhen University. Prof. Huang is known for his contributions to the development of a series of k-means type clustering algorithms in data mining, such as k-modes, fuzzy k-modes, k-prototypes and w-k-means, which are widely cited and used, and some of which have been included in commercial software. He has extensive industry expertise in business intelligence, data mining and big data analysis. He has been involved in numerous consulting projects in Australia, Hong Kong, Taiwan and mainland China. Dr Huang received his PhD degree from the Royal Institute of Technology in Sweden. He has published over 200 research papers in conferences and journals with over 10000 citations. In 2006, he received the first PAKDD Most Influential Paper Award. He has served as conference and program chairs of several national and international conferences in the areas of data mining and big data. He is recognized as a scientist of Career Scientific Impact in Stanford University World's top 2% scientists list.