Big Data Mining in the Cloud

Zhongzhi Shi

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
shizz@ics.ict.ac.cn

Abstract. Big Data is the growing challenge that organizations face as they deal with large and fast-growing sources of data or information that also present a complex range of analysis and use problems. Digital data production in many fields of human activity from science to enterprise is characterized by an exponential growth. Big data technologies will become a new generation of technologies and architectures which is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.

Massive data sets are hard to understand, and models and patterns hidden within them cannot be identified by humans directly, but must be analyzed by computers using data mining techniques. The world of big data present rich cross-media contents, such as text, image, video, audio, graphics and so on. For cross-media applications and services over the Internet and mobile wireless networks, there are strong demands for cross-media mining because of the significant amount of computation required for serving millions of Internet or mobile users at the same time. On the other hand, with cloud computing booming, new cloud-based cross-media computing paradigm emerged, in which users store and process their cross-media application data in the cloud in a distributed manner. Cross-media is the outstanding characteristics of the age of big data with large scale and complicated processing task. Cloud-based Big Data platforms will make it practical to access massive compute resources for short time periods without having to build their own big data farms. We propose a framework for cross-media semantic understanding which contains discriminative modeling, generative modeling and cognitive modeling. In cognitive modeling, a new model entitled CAM is proposed which is suitable for cross-media semantic understanding. A Cross-Media Intelligent Retrieval System (CMIRS), which is managed by ontology-based knowledge system KMSphere, will be illustrated.

This talk also concerns Cloud systems which can be effectively employed to handle parallel mining since they provide scalable storage and processing services, as well as software platforms for developing and running data analysis environments. We exploit Cloud computing platforms for running big data mining processes designed as a combination of several data analysis steps to be run in parallel on Cloud computing elements. Finally, the directions for further researches on big data mining technology will be pointed out and discussed.

Z. Shi, D. Leake, and S. Vadera (Eds.): IIP 2012, IFIP AICT 385, pp. 13–14, 2012.

Acknowledgement. This work is supported by Key projects of National Natural Science Foundation of China (No. 61035003, 60933004), National Natural Science Foundation of China (No. 61072085, 60970088, 60903141), National Basic Research Program (2007CB311004).