Dr. Kaushik Dutta
NUS Singapore
Abstract of talk

(Dr. Kaushik Dutta)

"Data Analytics and Cloud: A path to Big Data Analytics"

Data mining has been a very well researched topic from the early days of computer usages. It has been applied in many areas such as market basket, stock market and weather prediction – to mention a few. One of the popular approaches of data mining from complex data-set is machine learning. Both supervised and unsupervised machine learning based data mining techniques have been applied to solve many traditional research problems such as image processing and textual analysis.

The input of any data mining and machine learning technique is data. At present, the volume of data is growing at an unprecedented rate. Data are being produced everywhere, from Facebook, Twitter, YouTube to Google search records, and more recently, mobile apps. According to a research report from researchers in International Data Corporation (IDC), the digital data, which can be analyzed by computers, will double about every two years from now until 2020. In addition to the volume, the format of data is much more diversified now – from structured data stored in traditional databases to unstructured data such as text, images and so on. Actually, the vast majority of new data being generated is unstructured. Data mining, particularly applying machine learning techniques on such large complex data-set is not straight forward. For example, in one of the use case the volume of data growth is in the range of 500 GB per week resulting in several tens of terabytes data. Processing and mining such big data with traditional single machine approach is not feasible.

The cloud services such as Amazon AWS and Rackspace enables us to host such big data without a lot of upfront investment. The cloud services allow users to apply traditional data mining and machine learning based approaches in a cloud based environment, where the computing resources required to store and process big data in the range of terabytes can grow and shrink as and when required.

In this presentation, Dr. Dutta will first present an overview of data-mining and machine learning based approaches including Support Vector Machine, Artificial Neural Network, Genetic Algorithm, Classification and Clustering approaches. Then he will describe some specific approaches of handling large volume of data, specifically big data in the range of terabytes. He will elaborate cloud based services available from Amazon and Rackspace to handle such big data. Lastly he will demonstrate how the cloud based services and machine learning approaches can be combined together to do analytics of big data. He will explain some details of cloud based big data analytics including big data based text processing and, clustering and classification on big data.

Lastly he will present some research problems in the area of big data analytics. He will provide some high level approaches to solving these problems that may be helpful for researchers and academics for continuing further investigation and research.