主题报告 3

大数据研究面对的挑战

Challenges of Big Data in Scientific Discovery

华云生教授 | 香港中文大学常务副校长、伟伦计算机科学与工程学讲座教授

报告摘要 Abstract

大数据正是近年来最热门的学科研究领域之一。大数据研究能加强科学、工程、医学、医疗、金融、商业的转化，以至最终为社会本身。在这次报告中，我们会探讨大数据的关键特质（容量、速度、多样性和准确性），以及它们与科学和工程上一些应用之关系。要真正处理大数据，新的范例转移是必要的。成功的大数据应用，需要以原处方法从大数据中自动提取新的知识，而无需将数据集中收集和保存。传统的算法复杂性理论可能不再保留，因为数据的规模会由于太大而不能储存或提取。为了探讨大数据在科学研究中的潜力，我们需要解决在数据复杂性、计算复杂性和系统复杂性的挑战。以建立核数据为基础，我们建议一个新方法去克服大数据应用的复杂性。核数据是原始数据中密集而又易处理的代表，当中包含相似的结构、数据特质或更高层次的特性。我们会透过在科学和工程上各种应用的例子，来阐明这些挑战，及在实际应用上的战略思考。

Big Data is emerging as one of the hottest multi-disciplinary research fields in recent years. Big data innovations are transforming science, engineering, medicine, healthcare, finance, business, and ultimately society itself. In this presentation, we examine the key properties of big data (volume, velocity, variety, veracity, and value) and their relation to some applications in science and engineering. To truly handle big data, new paradigm shifts will be necessary. Successful applications in big data will require in situ methods to automatically extracting new knowledge from big data, without requiring the data to be centrally collected and maintained. Traditional theory on algorithmic complexity may no longer hold, since the scale of the data may be too large to be stored or accessed. To address the potential of big data in scientific discovery, challenges on data complexity, computational complexity, and system complexity will need to be solved. We propose a new approach based on identifying kernel data to harness the complexity of big data applications. Kernel data is a compact and manageable representation of the original data, with similar structure, data properties, or meta-properties. We illustrate these challenges and approaches by drawing on examples in various applications in science and engineering.