闲聊大数据

张瑞
墨尔本大学计算与信息系统系
HTTP://WWW.RUIZHANG.INFO

内容侧重

- 上次Lucy更多从Business Intelligent角度讲
- 我更多从概念，技术，应用角度
- 我的简介
  - 本科清华，博士新加坡，现在墨尔本大学副教授

什么是大数据：民间版

- 你要买比萨饼，问什么时候能送到。回答说40分钟：因为手机泄露了你的位置。他进一步解释说，如果你自己过来，可以快一点：因为知道你骑的是助动车。你要 买海鲜比萨，他建议你换一个：因为他知道你最近的体检报告。你要用信用卡付账，他不同意：因为他知道你的信用卡恶意透支，被封了。

什么是大数据

- Buzz word
- Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. [Wikipedia]
- Key features
  - 3V/4V: Volume, Velocity, Variety, Veracity

大数据实例

- Sloan Digital Sky Survey: 200GB /day
- Facebook: 2.5 Billion Pieces Of Content And 500+ Terabytes / day
- Google: 20,000 TB/day

大数据技术

- Data management -- Databases
  - Database design
  - Indexing
  - Query optimization
- Data mining
  - Classification
  - Clustering
- Machine learning
  - Pattern recognition
  - Computational learning theory

大数据算法实例

- Google Page Rank
- Spatial index
- 我现在做的项目示例
  - Guess ethnic group by name
  - Destination prediction
- 我的expertise：大规模高效处理算法，10多人博士生/博士后/老师团队
  - Hadoop, Mapreduce, GPU cluster
  - Data crawling/information extraction/data mining

联系方式

欢迎各种交流，特别是合作项目。我负责墨尔本大学计算机系里大数据组对外合作联系。

网页: http://www.ruizhang.info

或者Google “Rui Zhang”

问题和其他话题