『大数据』方法论及示例

→ 大数据为什么**能**
→ 什么不**能**
→ 怎么才**能**

刘占亮

大数据的通常定义

- WIKIPEDIA
  - **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.
- 百度百科
  - 大数据（big data），是指无法在可承受的 时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。
- 国务院《促进大数据发展行动纲要》
  - 大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速 发展为对数量巨大、来源分散、格式多 样的数据进行采集、存储和关联分析，从中 发现新知识、创造新价值、提升新能力的新一代信息技 术和服务业态。

科学传统的理性主义

- 计算机科学中的实践
  - 从理性或直觉建立问题的模型
  - 通过对少量样本数据的观察归纳出模型(人工或机器学习)
  - 通过模型判别新样本
- 为什么需要模型
  - 从有限的经验中得到普遍性的规律
  - 历史上数据收集和分享的困难
  - 模型缓解了数据的不足
    - 泛化:从已知到未知
    - 模型的参数数目定义了样本的空间大小
  - 模型的内在困难
    - 是否总能够从特殊推到一般
    - 复杂模型:股市预测

经验主义的复活 ——大数据为什么**能**

- 大数据时代
  - 大量新技术使得数据的收集和分享 变得非常容易
    - 传感器
    - 互联网
- 数据越多越不需要模型
  - 覆盖度：对所有或大部分事件，有 样本覆盖
  - 精度：对高频事件，有足 够多样本来提升精度
- 传统方法 vs 大数据方法
  - 新样本-->特征表示-->利用小数据 训练的模型进行判断-->得出结论
  - 新样本-->查找已知大数据 样本中的相同或相似 样本判断-->得出结论

例子一：搜索引擎查询结果排序

- 问题描述：给定一个查询，对返回的网页结果进行相关性排序
- 传统解决问题方法：排序模型
  - 概率检索模型
  - 统计语言模型
  - 神经网络模型
- 大数据方法：用户点击数据挖掘
  - 给定一个查询，根据用户对网页的点击率排序
  - 需要大量的数据：查询数*网页数
- 效果提升明显：是现在商用搜索引擎中的最强feature

例子二：机器翻译

- 问题：将一种语言自动翻译为另一种语言
- 传统方法：语料库+翻译模型
- 大数据方法：平行语料挖掘
  - 从互联网上自动发现大量的双语语料
  - 统计词语、短语、甚至句子之间的对照关系
- 非常显著的性能提升，是目前最好的方法，如：Google翻译

大数据的通常定义

- WIKIPEDIA
  - **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.
- 百度百科
  - 大数据（big data），是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。
- 国务院《促进大数据发展行动纲要》
  - 大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速 发展为对数量巨大、来源分散、格式多 样的数据进行采集、存储和关联分析，从中 发现新知识、创造新价值、提升新能力的新一代信息技 术和服务业态。

多大算大，多像算像

- 大数据是现代社会在掌握了海量数据收集、存储和处理技术基础上所产生的一种以群体智慧进行判断和预测的能力。
- 内在含义
  - 经验主义 > 理性主义
  - 数据 > 模型
  - 相关关系 > 因果关系
  - 数据 > 平台 > 模型
- 数据多到能对**整个**样本空间进行**充分**覆盖
  - 预测投硬币：样本空间{0, 1}，1000个样本足以得到置信度很高的 预测
  - 机器翻译：样本空间，所有可能的句子？

模型真的没用吗？ ——*大数据什么不**能***

- 数据总是不够
  - 样本空间太大
    - 机器翻译例子中所有可能的句子
    - 自动求解学生考试题目
  - 样本空间实时变化
    - 查询结果排序例子中，新查询和新网页不停出现
- 模型需要和数据结合，提供适当的泛化能力
  - 如何结合？
  - 何种程度？

例子：智能和认知科学

- 数据驱动的方法论（大数据，深度学习）能够解决终极问题吗？
- 基本每隔几十年都会吹响人工智能的号角

大数据应用开发的流程 ——大数据怎么才**能**

1. 确定问题的数据类型和样本空间
2. 收集到尽可能多(或足够多)的相关数据来覆盖样本空间
    a. 不要特别在意数据质量和格式
3. 选择(或搭建)合适的大数据处理平台
4. 针对应用对数据进行预处理
    a. 格式转换、数据抽取、数据集成(多源数据融合)
    b. 数据质量控制
5. 处理数据
6. 结果解读和应用

数据在哪？

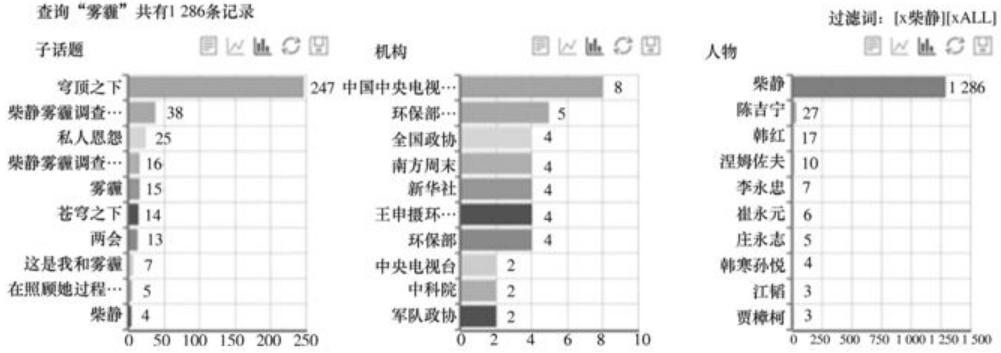
- 政府/权威机构
  - 统计局宏观数据
  - 行业数据（人口、交通、天气、卫星等）
  - 金融机构交易数据
  - 机构内部数据
- 互联网数据
  - Web数据（新闻、论坛、微博、微信等）- 主动通过Web开放的数据
  - 互联网公司私有数据 - 业务数据/用户行为数据
  - 运营商数据 - 用户行为数据

应用示例DEMO

- 互联网大数据分析引擎
  - 通用实时数据抓取引擎
  - 智能数据抽取清洗
  - 自然语言处理(非结构数据的结构化)
  - 高效的数据多维索引架构
  - 数据深度智能分析(互联网数据的OLAP)
  - 数据可视化展示
- 应用于财经新闻领域
  - 数据源:hao123财经资讯收录的34个网站
  - 日均新闻量10000篇(过去十年所有财经新闻数量1500万+)
  - http://finance.zliu.org/hot.php
- 国家安全生产监督管理局
  - 2007年~2015年的2万多条数据
  - http://misc.zliu.org/do.php?w=ALL

查询“雾霾”共有1 620 808条记录

（a）初始分析结果



（b）选定“柴静”后的分析结果

结论分享

- 数据源非常相关
  - 新闻、论坛、微博、微信，可以得出完全迥异的 结论
  - 物理世界与数字世界的映射
  - 问题的样本空间(是不是大数据)
- 挑战
  - 大数据思维
  - 计算机技术
  - 领域知识