中国科学院大学
University of Chinese Academy of Sciences

计算机与控制学院
School of Computer and Control Engineering

移动互联网技术

第四章　互联网大数据处理技术

概述

王文杰
wangwj@ucas.ac.cn

本章要内容

1. 概述
2. 语料库介绍
3. 中文分词
4. 词性标注
5. 文本相似度分析
6. 大数据分析模型与算法

概述

- 大数据处理主要包括大数据的分析和挖掘技术
- 所谓数据分析，即对已知的数据进行分析，然后提取出一些有价值的信息，
  - 比如统计出平均数、标准差等信息，数据分析的数据量有时可能不会太大
- 所谓数据挖掘，是指对大量的数据进行分析与挖掘，得到一些未知的、有价值的信息等
  - 比如从网站的用户或用户行为数据中挖掘出用户的潜在需求信息，从而对网站进行改善等。
- 数据分析与数据挖掘密不可分，数据挖掘是数据分析的提升

概述

- 数据的分类：
  - 结构化数据：
    - 结构化数据是指被标签定义了其内容、意义和用法的数据。
    - 结构化数据一般包括：
      - 数据本身
      - 对数据的描述

      } 按照严格的规则进行组织
    - 例如，存储在数据库中，可以用二维表结构来逻辑表达实现的数据
  - 半结构化数据：
    - 自描述、数据结构和内容混杂在一起的数据，如XML、HTML等
  - 非结构化数据：
    - 除去以上两种类型的数据，如视频、音频、图片等

概述

- 数据的表示：
  - 结构化的数据模型可以用二维表（关系型）来表示，如关系型数据库中的数据等
  - 半结构化数据模型可以用树和图来表示，如HTML
  - 非结构化数据没有数据模型。
- 结构化数据的特点是先有结构、再有数据

  半结构化数据的特点是先有数据，再有结构。

概述

- **大数据处理技术主要包括：**
  - 结构化处理技术
  - 词性标注技术
  - 语法分析
  - 语义分析技术
  - 大数据分析模型与算法

Python互联网大数据处理相关模块简介

1. **Numpy**：由多维数组对象和用于处理数组的例程集合组成。很多模块都依赖他，比如pandas、 scipy、matplotlib等
2. **Pandas**：主要用于进行数据探索和数据分析。
3. **Matplotlib**：用于作图，解决可视化问题。
4. **Scipy**：主要进行数值计算，同时支持矩阵运算，并提供了很多高等数据处理功能，比如积分、傅里叶变换、微分方程求解等。
5. **statsmodels** ：主要用于统计分析
6. **NLTK**：自然语言处理模块
7. **jieba**：结巴分词
8. **Gensim** ：主要用于文本挖掘
9. **sklearn、keras**：机器学习

相关模块的安装

- 模块安装的顺序与方式建议如下：
  1. numpy、mkl（下载安装）
     https://www.lfd.uci.edu/~gohlke/pythonlibs/
  2. pandas（网络安装）
  3. matplotlib（网络安装）
  4. scipy（下载安装）
  5. statsmodels（网络安装）
  6. Gensim（网络安装）

本章要内容

1. 概述
2. 语料库介绍
3. 中文分词
4. 词性标注技术
5. 文本相似度分析
6. 大数据分析模型与算法

语料库介绍

- 语料库(corpus)
  - 语料库(corpus) 就是存放语言材料的仓库 (语言数据库)。
  - 现代语料库是指存放在计算机里的原始语料文本或经过加工后带有语言学信息标注的语料文本的汇集
- 基于语料库进行语言学研究—语料库语言学 (corpus linguistics)
  - “语料库语言学已经成为语言研究的主流。 基于语料库的研究不再是计算机专家的独有领 域，它正在对语言研究的许多领域产生愈来愈 大的影响。”（ J. Thomas 等）

语料库介绍

- 语料库语言学是以语料库中实际存储的真实语言材料作为唯一的研究对象，以语言现象的出现概率为依据。
- 因此，语料库可以如实地反映语言现象，克服语言学家观察语言现象时的主观性
- 例如：
  - **Start 或 begin**，在口语中哪个更常用？
  - 老师经常说：**Let's begin!**之类的话吗？

语料库语言学研究的内容

- 语料库的建设与编纂
  - 出发点是：如何使得在其基础上开展的语言调查是合理的和可靠的
- 语料库的加工和管理技术
  - 主要指用于语料分析、标注、维护和检索软件工具。
- 语料库的使用
  - 如对于下面问题的研究：
    - 汉语文本中交集性切分歧义的研究
    - 汉语基本名词短语识别研究
    - 基于结构语义空间的汉语语义排歧模型

语料库的分类

- 按应用取向分为：通用型和专用型语料库
- 按信道分为：笔语和口语语料库
- 按语言属性分为：单语、双语、多语语料库
- 按语言变体分为：本族语、译语、学习者语料库
- 按时间分为：共时、历时语料库
- 按语料状态分为：静态和监控语料库

几个典型语料库

- 布朗语料库 (Brown Corpus)
  - 20世纪60s，Francis 和 Kucera 在布朗(Brown)大学建立，是世界上第一个根据系统性原则采集样本的标准语料库， 100万词规模
- LLC口语语料库(London-Lund Corpus of Spoken English)
  - 1960s 伦敦大学著名语言学家Quirk 组织，瑞典隆德(Lund) 大学教授 Svartvik 主持录入计算机 ，最终规模 50万词
  - 5大类：面对面交谈；电话交谈；讨论；采访；辩论，未 经准备的当众评论、论证、演讲，经准备的当众演讲
  - 标注：语调、节律、关键词(语段)，词类、出现次数、搭 配关系等

几个典型语料库

- 朗文语料库 (Longman Corpus)
  - January 1981- November 1990， 2800 万词
  - 10个分布广泛的领域：自然和纯科学、应用科学、社会 科学、世界事务等
- 宾夕法尼亚大学(UPenn)树库(Tree Bank) (http://www.ldc.upenn.edu/)
  - 美国宾夕法尼亚大学计算机系 M. Marcus 教授主持
  - 1993年完成约300万词次英语句子的语法结构标注
  - 2000年完成第一版汉语树库，约10万词次，4185个 句子
  - Chinese Tree Bank (CTB) 中汉语词性(part-of-speech) 被划分为33类，23类句法标记(Syntactic tags)

几个典型语料库

- 北京大学开发的CLKB
  - 现代汉语语法信息词典：8万词、360万项语法属 性描述
  - 汉语短语结构规则库：600多条语法规则
  - 现代汉语多级加工语料库：实现词语切分并标注 词类的基本标注语料库1.5亿字，其中精加工的有 5200万字，标注义项的有2800万字
  - 多语言概念词典：10万个以同义词集表示的概念
  - 平行语料库：含对译的英汉句对100万
  - 多领域术语库：有35万汉英对照术语 4.6 典型语料库介绍

几个典型语料库

- **WordNet (http://wordnet.princeton.edu/) –词汇知识库**
  - **普林斯顿大学(Princeton University) 认知科学实验室 George A. Miller 教授领导开发。**
  - **开发目的：解决词典中同义信息的组织问题**
  - **规模：95600 英语词条，其中，51500个简单 词，44100 个搭配词。70100个词义(同义词集合)。**
  - **五大类词汇：名词、动词、形容词、副词、虚词。 (实际上 WordNet 中仅包含前4类)**

几个典型语料库

- **WordNet (http://wordnet.princeton.edu/) –词汇知识库**
  - 特色：根据词义（而不是词形）组织词汇信息，从 某种意义上讲，它是一部语义词典。
  - **WordNet 按语义关系组织：语义关系看作是同义词 集合之间的一些指针，语义关系是双向的。如果词 义{x1 , x2 , …} 和{y1 , y2 , …}之间有一种语义关系R， 则在{y1 , y2 , …}和{x1 , x2 , …} 之间也有语义关系R。 属于这两个同义词集合的单词之间的关系也是R。**

几个典型语料库

- **WordNet (http://wordnet.princeton.edu/) –词汇知识库**
  - **– 4 种语义关系：**
    - **同义关系(synonymy)**
    - **反义关系(antonymy)**
    - **上下位关系(hypernymy)或称从属/上属关系： 如：{枫树}是{树}的下位，{树}是{植物}的下 位。**
    - **部分关系(meronymy)或称部分/整体关系**

技术不会停下脚步，学习永无止境。