数据库系统概论新技术篇

社交网络大数据分析
--社交媒体初探（1）

赵鑫

中国人民大学信息学院

2017年5月

讲述提纲

1 社交媒体概述

2 常用数据处理技术

3 常见任务与解决方法

4 未来研究展望

5 小结

社交媒体概述

- 社交媒体（**Social Media**）是指

  基于社交应用服务的内容生产与交换平台。

- 很多情况下，不只局限于拥有用户关系的网络平台，

  而泛指公开的内容发布平台

社交媒体概述（续）

- 包含多种应用
  - 微博
  - 朋友圈
  - 博客
  - 电子商务平台
  - 视频平台
  - 。。。

社交媒体概述（续）

- 包含多种数据类型

图像	音频	视频	文本	关系
				

本讲主要关注基于文本与关系的数据挖掘研究

讲述提纲

1 社交媒体概述

2 常用数据处理技术

3 典型任务与解决方法

4 未来研究展望

5 小结

常用数据处理技术

- 独立于特定任务的数据处理技术
  - 文本数据处理
  - 关系数据处理

常用数据处理技术（续）

常用文本预处理技术

常用文本语义分析技术

常用网络关系建模技术

常用数据处理技术（续）

- 中文预处理技术
  - 编码转换
  - 分词
  - 词汇过滤
    - 去停用词
    - 去高频词
    - 去低频词

常用数据处理技术（续）

- 中文预处理技术
  - 编码转换
  - 分词
  - 词汇过滤
    - 去停用词
    - 去高频词
    - 去低频词

常见中文字符编码包括:

GB2312
BIG5
GBK
UTF-8
…

需要提前进行统一转换

常用数据处理技术（续）

- 中文预处理技术
  - 编码转换
  - 分词
  - 词汇过滤
    - 去停用词
    - 去高频词
    - 去低频词

中文文本为字符串，没有词汇边界。
需要提前进行中文切词

我是一个中国人

我_是_ 一个_中国人

常用数据处理技术（续）

- 中文预处理技术
  - 编码转换
  - 分词
  - 词汇过滤
    - 去停用词
    - 去高频词
    - 去低频词

停用词通常包括: 的、地、得。。。

很多情况下，超高频词 和 超低频词 也可以考虑去掉

超低频词 可能来自于笔误或者新词，去掉后可以减少词典规模

常用数据处理技术（续）

常用文本预处理技术

常用文本语义分析技术

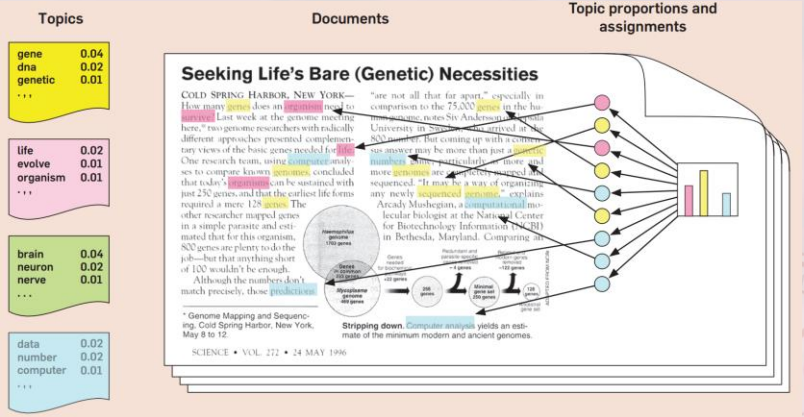
常用网络关系建模技术

常用数据处理技术（续）

- 文本语义分析技术
  - 主题模型
  - 词嵌入模型

主题模型 (LDA):

输入： 大规模文档语料
输出： 文档-主题分布、主题-词汇分布

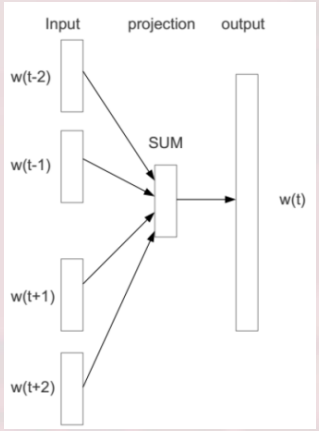


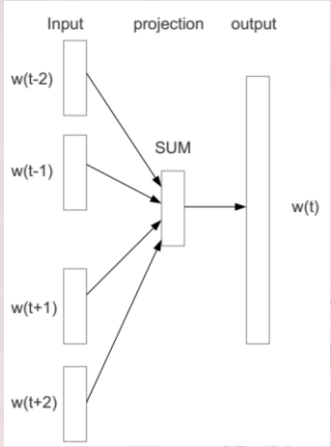
常用数据处理技术（续）

- 文本语义分析技术
  - 主题模型
  - 词嵌入模型

词嵌入模型 (word2vec):

输入： 大规模文档语料
输出： 词汇隐含空间内的低维表示





常用数据处理技术（续）

常用文本预处理处理技术

常用文本语义分析技术

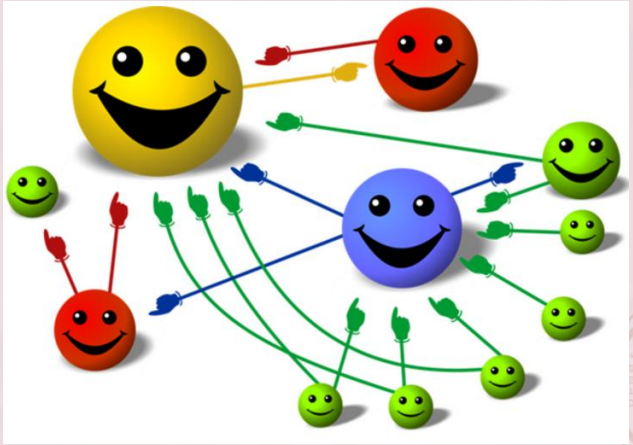
常用网络关系建模技术

常用数据处理技术（续）

- 网络关系建模技术
  - PageRank
  - 图正则化
  - 网络嵌入

PageRank模型:

输入：大规模网络关系图
输出：网络节点的“重要度”分数

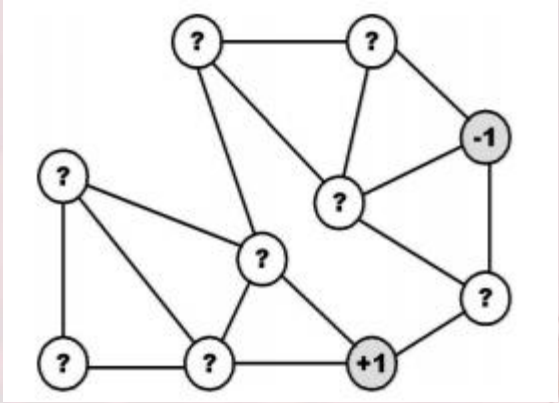


常用数据处理技术（续）

- 网络关系建模技术
  - PageRank
  - 图正则化
  - 网络嵌入

图正则化模型:

输入： 大规模网络图以及部分节点标签
输出： 全局预测出来的节点标签

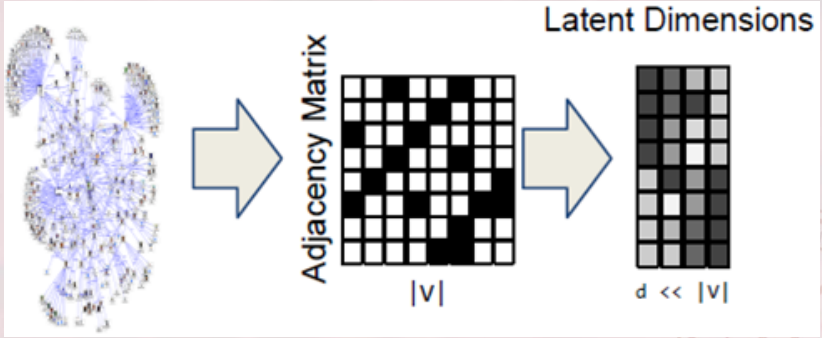


常用数据处理技术（续）

- 网络关系建模技术
  - PageRank
  - 图正则化
  - 网络嵌入

网络嵌入模型:

输入：大规模网络图
输出：每个节点的隐含空间低维表示



讲述提纲

社交媒体概述

常用数据处理技术

典型任务与解决方法

研究方向展望

小结

小结

了解常用的社交媒体数据处理技术