R与文本挖掘

文本挖掘简介与系统实现

李舰

Email: lijian.pku@gmail.com

Homepage: www.leejian.name

第三届中国R语言会议

2010 年 6 月

目 录

1 文本挖掘简介
- 概念
- 文本预处理
- 文档模型
- 文本挖掘技术
- 应用范围

目 录

1. 文本挖掘简介
   - 概念
   - 文本预处理
   - 文档模型
   - 文本挖掘技术
   - 应用范围
2. 系统实现
   - 系统架构
   - 实现示例
   - R的优势

目 录

1. 文本挖掘简介
   - 概念
   - 文本预处理
   - 文档模型
   - 文本挖掘技术
   - 应用范围
2. 系统实现
   - 系统架构
   - 实现示例
   - R的优势
3. 云计算
   - 云计算简介
   - MapReduce
   - RHIPE简介

文本挖掘的概念

- 名称
  - Text Mining
  - Text Data Mining
  - Knowledge Discovery in Text
  - Knowledge Discovery in Textual Data(bases)
- 定义
  - 文本挖掘是从大量文本数据中抽取隐含的，未知的，可能有用的信息。

数据清洗

- 预处理
  - 字符编码转换："UTF-8"
  - 正则表达式：gsub("[^\u4e00-\u9fa5]","",x)
- 中文分词
  - 乒乓/球拍/卖/完/了
  - 乒乓球/拍卖/完/了

中文分词常用方法

- 最大匹配法
  - 设定最大词长，从左到右匹配
  - 最大匹配法、双相匹配法、最佳匹配法、联想回溯法
  - 示例：长春市长春节致词
- 最大概率法
  - 一个待切分的汉字串可能包含多种分词结果，将其中概率最大的那个作为该字串的分词结果
  - 条件概率近似公式
  - 示例：乒乓球拍卖完了
- 最短路径分词方法
  - 在词图上选择一条词数最少的路径
  - 示例：他说的确实在理
- 隐马尔可夫模型
  - HMM，基于马尔可夫过程
  - 利用转移概率分词

中文分词工具简介

- ICTCLAS
  - 中文分词、词性标注、命名实体识别、新词识别
  - 分词正确率高达97.58％
- 基于Lucene的中文分词器
  - Paoding
  - imdict，使用ICTCLAS HMM隐马尔科夫模型
  - mmseg4j，MMSeg算法
  - ik，正向迭代最细粒度切分算法

文档模型简介

- 布尔模型
  - 以集合论和布尔代数为基础
  - 进行布尔逻辑运算
- 向量空间模型
  - 基于概率论和信息论
  - 将文档转化为向量，看作向量空间的一个点
- 文档概率模型
  - 基于贝叶斯方法

文本分类

- 中文分词
  - 将网络信息处理为标准化文本，进行分词操作
- 文档建模
  - 将分词后的文章转化成向量模型（Term Vector）
  - 计算文章d中每个词w在t时刻取词的权数:

$$weight_t(d,w) = \frac{tf(d,w)log((W_t+1)/(wf_t(w)+0.5))}{\sqrt{\sum_{w^1 \in d}(tf(d,w^1)log((W_t+1)/(wf_t(w^1)+0.5)))^2}}$$

- 按照话题聚类
  - 将t时刻采集的文本向量按相似度（Similarity）聚类
  - 使用余弦夹角的方式计算相似度
- 判别分析
  - 将类簇归入某个已知的类别
- 评价和检验
  - 计算precision（查准）、recall（查全）
- 更新训练集

其他挖掘技术（一）

- 文本智能检索
  - 网络搜索
  - 全文检索
- 话题检测跟踪
  - Topic Detection and Tracking (TDT)
  - 话题检测，将新闻分为话题类簇
  - 话题跟踪，监控新闻报道信息流以便发现与某一已知话题有关的新报道
- 文本过滤
  - 信息过滤(IF)，从动态的信息流中将满足用户兴趣的信息挑选出来
  - 关注用户建模

其他挖掘技术（二）

- 关联分析
  - 类似于DM中的关联规则
  - 关键词-性能指标矩阵
- 文档自动摘要
  - 利用计算机自动地从原始文档中提取全面准确地反映该文档中心内容的简单连贯的短文
  - 摘要方法：位置法、提示字串法、频率统计法、文章框架法、仿人算法
  - 评价方式：利用文档摘要代替原文档执行某个文档相关的应用（检索、分类等）

文本挖掘的应用范围

- 智能信息检索
  - 同义词、简称词、异形词、同音字、赘字移除等
- 网络内容安全
  - 内容监控
  - 内容过滤
- 内容管理
  - 自动分类
  - 检测和追踪
- 市场监测
  - 口碑监测
  - 竞争情报系统
  - 市场分析

系统环境

- 开发环境
  - 数据库Oracle
  - 数据层iBatis
  - 控制层Spring
  - 展现层JSP
  - 运算引擎R
- 数据采集
  - Lucene + Nutch
  - JAVA定制开发
- 文本挖掘
  - R语言（rtm包）
  - rJava
  - 中文分词工具imdict-chinese-analyzer

系统架构示例


应用层
新闻出版
图书馆档案馆
舆情预警
企业竞争情报CIS
电子政务
CRM/ERP
核心层
智能获取
网络雷达
元搜索
数据网关
智能处理
自动分类
自动消重
自动摘要
智能分析
主题检测
关联分析
知识地图
智能检索
全文检索
相关推荐
图片检索
自然语言理解技术与数据挖掘技术
支撑层
内容管理平台（文本、图像、视音频、数据库）
软件硬件平台

中文分词

- 先将文本进行预处理，然后进行中文分词，使得每一篇文档转化为词的集合。

```
>
> VsegWord("R是一门用于统计计算和作图的语言，其官方机构每年
+ 都会举办useR!会议，但会议地点主要局限在欧美地区。",newws)
 [1] "r"    "是"   "一"   "门"   "用于" "统计" "计算" "和"   "作"   "图"
[11] "的"   "语言" "其"   "官方" "机构" "每年" "都"   "会"   "举办" "user"
[21] "会议" "但"   "会议" "地点" "主要" "局限" "在"   "欧"   "美"   "地区"
>
```

向量化处理

- 文本向量化处理，计算每篇文档每个词的频数和权数。

```
> Lweight[1:10,]
    no   V1 V2         V3 istory
1    1    一  1 0.02950499     814
2   40 一律  1 0.06534208     814
3  129    上  1 0.03248047     814
4  180 上限  1 0.06333967     814
5  286 不得  3 0.15592462     814
6  389    且  2 0.10098861     814
7  445    两  1 0.03706464     814
8  465 严格  1 0.04849784     814
9  500 中国  1 0.03097695     814
10 577    为  1 0.03099827     814
>
```

R语言在TM中的优势

- 中文分词
  - rJava + imdict-chinese-analyzer
  - R语言开发or JAVA数据交换
- 文本向量模型
  - RSQLite + DB索引优化+ R语言下标计算
  - 矩阵和data.frame
- 文本挖掘
  - 矩阵运算
  - 聚类模型

什么是云计算

- 维基百科
  - 云计算将IT相关的能力以服务的方式提供给用户，允许用户在不了解提供服务的技术、没有相关知识以及设备操作能力的情况下，通过Internet获取需要服务。
- 中国云计算网
  - 云计算是分布式计算（Distributed Computing）、并行计算（Parallel Computing）和网格计算（Grid Computing）的发展，或者说是这些科学概念的商业实现。
- Forrester Research 的分析师James Staten
  - 云计算是一个具备高度扩展性和管理性并能够胜任终端用户应用软件计算基础架构的系统池。

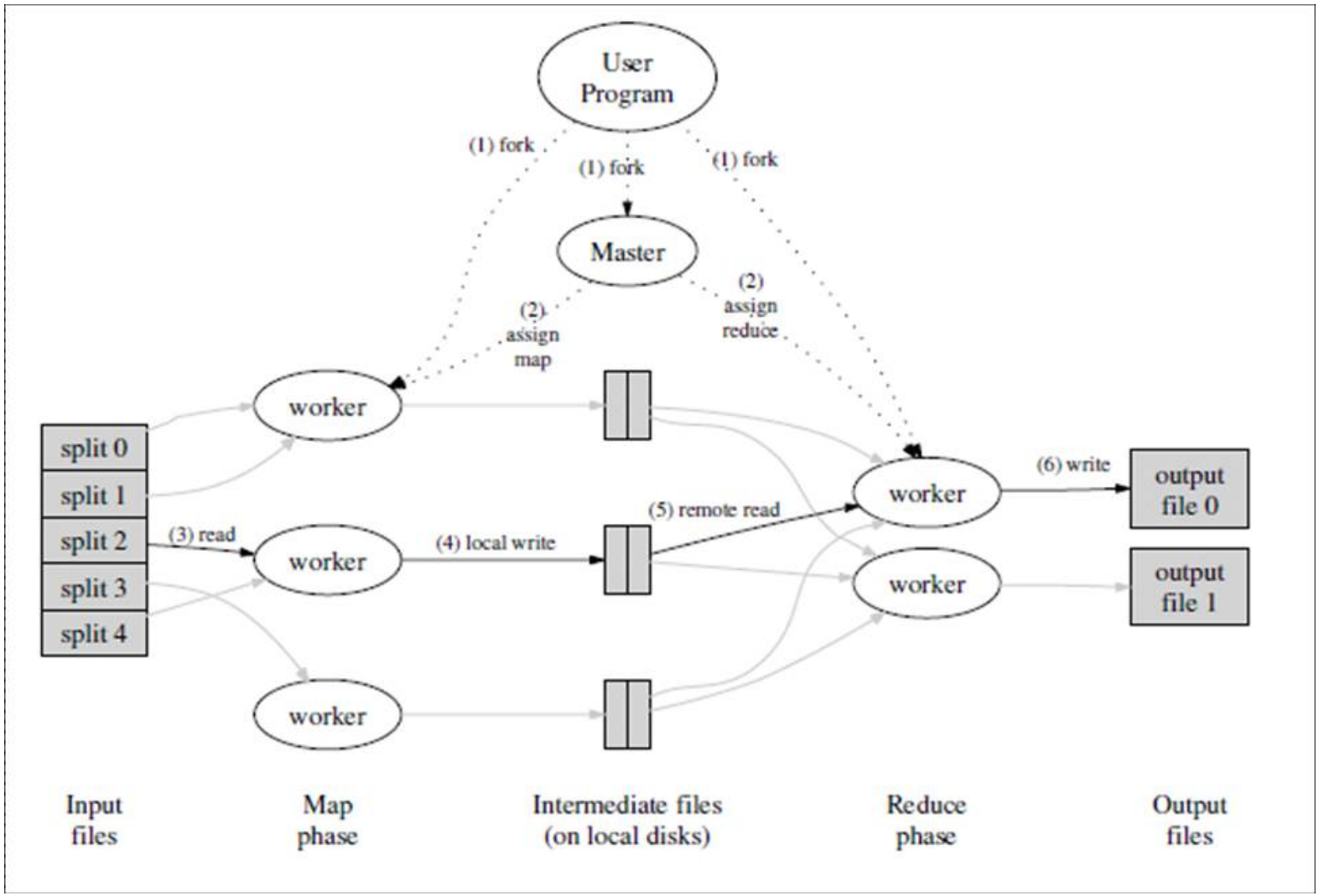
云计算的特点

- 云计算系统提供的是服务
  - 服务的实现机制对用户透明，用户无需了解云计算的具体机制，就可以获得需要的服务。
- 用冗余方式提供可靠性
  - 云计算系统由大量商用计算机组成机群向用户提供数据处理服务。采用软件的方式，即数据冗余和分布式存储来保证数据的可靠性。
- 高可用性
  - 云计算系统可以自动检测失效节点，并将失效节点排除，不影响系统的正常运行。
- 高层次的编程模型
  - 云计算系统提供高级别的编程模型。用户通过简单学习，就可以编写自己的云计算程序，在“云”系统上执行，满足自己的需求。现在云计算系统主要采用MapReduce模型。
- 经济性
  - 组建一个采用大量的商业机组成的机群相对于同样性能的超级计算机花费的资金要少很多。

MapReduce

- Google的一个专利申请
  - 2010年1月获批，编号为7 650 331，名为System and method for efficient large-scale data processing（高效大规模数据处理）。是Google最引为自豪的成果之一，也是云计算最重要的核心技术之一。
- MapReduce的应用
  - Google基础应用
  - 雅虎搜索
  - Amazon的Elastic MapReduce服务
  - 开源项目Apache Hadoop

Google的MapReduce执行方式



R Package： mapReduce

- MapReduce思路的简单实现
  - 基于apply系列函数
  - 功能和by以及aggregate相似
- 实现方式
  - 使用split函数将矩阵进行拆分
  - 使用apply函数并行处理
  - 汇总输出

RHIPE简介

- 开源的MapReduce：Hadoop
  - Hadoop 是Google MapReduce 的一个Java实现
  - 定义Mapper，处理输入的Key-Value对，输出中间结果。定义Reducer，可选，对中间结果进行规约，输出最终结果。定义main函数。
  - 提交JOB，系统自动完成
- R和Hadoop的整合：RHIPE
  - 开源项目，将R和Hadoop集成在一起
  - 目前只有Linux和Mac OS版本

Thank you!

Email: lijian.pku@gmail.com
Homepage: www.leejian.name