MapReduce课程设计选题

- 课程设计1 - 体育赛事日志分析
- 课程设计2 - 人物关系挖掘
- 课程设计3 - 新闻自动分类

课程设计2—人物关系挖掘

- 1. 课程设计目标

通过一个综合数据分析案例：“西游释厄传——西游记中的人物关系挖掘”，来学习和掌握 MapReduce 程序设计。通过本课程设计的学习，可以体会如何使用 MapReduce 完成一个综合性的数据挖掘任务，包括全流程的数据预处理、数据分析、数据后处理等。

课程设计2—人物关系挖掘

- 2. 学习技能

通过本课程设计，可以熟悉和掌握以下 MapReduce 编程技能：

a) 在 Hadoop 中使用第三方的 JAR 包来辅助分析

b) 掌握简单的 MapReduce 算法设计

① 单词同现算法

② 数据整理与归一算法

③ 数据排序

c) 掌握带有迭代特性的 MapReduce 算法设计

① PageRank 算法

② 标签传播算法（选做）

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 1：数据预处理
    - 从原始的西游记小说的文本中，抽取出与人物互动相关的数据。需要屏蔽与人物关系无关的文本内容，为后面的基于人物共现的分析做准备。
    - 数据输入：西游记系列小说文集（未分词）；西游记系列小说中的人名列表。
    - 数据输出：分词后保留人名。

输入：（西游记中的某一段内容）将近天门，金星高叫道：“那天门天将，大小吏兵，放开路者。此乃下界仙人，我奉玉帝圣旨，宣他来也。”这增长天王与众天丁俱才敛兵退避。猴王始信其言。同金星缓步入里观看…

输出：金星 玉帝 增长天王 猴王 金星

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 2：人物同现统计
    - 完成基于单词同现算法的人物同现统计。在人物同现分析中，如果两个人在原文的同一段落中出现，则认为两个人发生了一次同现关系。我们需要对人物之间的同现关系次数进行统计，同现关系次数越多，说明两人之间的关系越密切。
    - 数据输入：任务 1 的输出
    - 数据输出：人物之间的同现次数

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 2：人物同现统计
    - 注意：小说对于人物名称的使用并不统一。例如某些段落使用全名、某些段落使用不带姓氏的名字、某些段落使用称号等。为了提高分析结果的准确性，请将小说中的主要人物的名称进行统一（主要人物及其别名已经在人名列表文件的开头给出）。

输入：

```
唐僧 悟空 猴王 八戒
悟空 八戒
```

输出：

```
<唐僧，悟空> 1
<唐僧，八戒> 1
<悟空，唐僧> 1
<悟空，八戒> 2
<八戒，唐僧> 1
<八戒，悟空> 2
```

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 3：人物关系图构建与特征归一化
    - 根据共现关系，生成人物之间的关系图。任务关系图使用邻接表形式表示，方便后续的 PageRank 计算。人物关系图中，人物是顶点，人物间的互动关系是边，人物互动关系靠人物间的共现关系决定。如果两个人之间具有共现关系，则两个人之间就具有一条边。两人之间的共现次数体现出两人关系的密切程度，反映到共现关系图上就是边的权重。权重越高，两人关系越密切。

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 3：人物关系图构建与特征归一化
    - 数据输入：任务 2 的输出。
    - 数据输出：人物关系图。
    - 注意：为了使后面的分析方便，需要对共现次数进行归一化处理：将共现次数转换为共现概率。

输入：

```
<唐僧，悟空> 1
<唐僧，八戒> 1
<悟空，唐僧> 1
<悟空，八戒> 2
<八戒，唐僧> 1
<八戒，悟空> 2
```

输出：

```
唐僧 [悟空，0.5|八戒，0.5]
悟空 [唐僧，0.33333|八戒，0.66666]
八戒 [唐僧，0.33333|悟空，0.66666]
```

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 4：基于人物关系图的 PageRank 计算
    - 计算 PageRank，定量分析小说的主角。
    - 数据输入：任务 3 的输出
    - 数据输出：各人物的 PageRank 值
    - 注意：该任务默认的输出是杂乱的，从中无法直接得到分析结论。需要对 PageRank 值进行全局排序，确定 PageRank 值最高的任务。排序工作可用一个 MapReduce 程序完成，也可导入 Hive 中，利用 Hive 完成排序。

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 5：人物关系图上的标签传播（选做）
    - 实现标签传播算法。标签传播是一种半监督图分析算法，通过在图上顶点打标签，进行图顶点的聚类分析，从而在一张社交网络图中完成社区发现。
    - 数据输入：任务 3 的输出
    - 数据输出：人物标签信息
    - 注意：对于该任务的输出，可通过一个 MapReduce 程序将属于同一标签的人物输出到一起，以便查看标签传播结果。

课程设计2—人物关系挖掘

- 3. 题目描述
  - 任务 5：人物关系图上的标签传播（选做）
    - 参考文献：
      - https://journals.aps.org/pre/abstract/10.1103/PhysRevE.76.036106
      - http://www.cnphp6.com/archives/24136

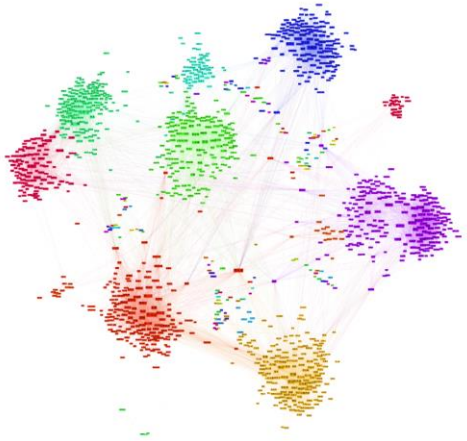

图 1|标签传播的结果展示

注：人物名字的大小由人物顶点的度数确定,人物标签的颜色根据标签传播算法的分析结果确定。

课程设计2—人物关系挖掘

- 4. 提交材料
  - 程序源代码，要求提供包含完整目录结构的 src 代码包，并提供编译和执行方法说明
  - 程序可执行 jar 包以及 jar 包的执行方式。本课程设计的运行环境为 hadoop-2.7、jdk-1.7 或以上环境
  - 程序设计报告。报告内容包括程序设计的主要流程、程序采用的主要算法、进行的优化工作、优化取得的效果、程序的性能分析以及程序运行截图等。