



大数据处理综合实验 课程设计（2023）

南京大学 计算机科学与技术系

MapReduce课程设计选题



- 课程设计1 - 网站访问日志分析
- 课程设计2 - 哈利波特的魔法世界
- 课程设计3 - 新闻文本情感分类



课程设计2 - 哈利波特的魔法世界

1 课程学习目标

通过一个综合数据分析案例：“哈利波特的魔法世界（英文版）系列小说中的人物关系挖掘”，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 - 哈利波特的魔法世界

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：

1. 在 Hadoop 中使用第三方的 Jar 包来辅助分析；
2. 掌握简单的 MapReduce 算法设计：
 - a) 单词同现算法；
 - b) 数据整理与归一化算法；
 - c) 数据排序（选做）；
3. 掌握带有迭代特性的 MapReduce 算法设计：
 - a) PageRank 算法；
 - b) 标签传播（Label Propagation）算法（选做）。



课程设计2 - 哈利波特的魔法世界

3 任务描述

本课程设计包括如下的若干任务。这些任务组合起来，就构成了一个完整的人物关系分析流程。

任务 1 数据预处理

1) 本任务的主要工作是从英文版哈利波特系列小说的文本中，抽取与人物互动相关的数据，而屏蔽掉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。

2) **数据输入**：系列小说文集（未分词）；小说中出现的人名列表

3) **数据输出**：保留人名

4) **示例**：

输入：（某段内容）Harry Potter opened the door and Ron Weasley came in.

输出：Harry Potter, Ron Weasley



课程设计2 - 哈利波特的魔法世界

注：小说全文中对于人物名称的使用并不统一，例如部分章节使用

“Harry Potter”，部分章节使用“Harry”、“Potter”。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，次要人物可不进行处理。例如将“Harry Potter”、“Harry”、“Potter”统一处理为“Harry Potter”。（也可处理为“Harry”或“Potter”，具体的统一情况处理可自己选择）



课程设计2 - 哈利波特的魔法世界

任务 2 特征抽取：人物同现统计

本任务的重要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，则说明两人的关系越密切。

输入输出

输入：任务 1 的输出；

输出：在哈利波特系列小说中，人物之间的同现次数。

示例：

输入：Harry Ron Ron Hermione

输出：<Harry, Ron> 2 <Ron, Harry> 2
 <Harry, Hermione> 1 <Hermione, Harry> 1
 <Ron, Hermione> 2 <Hermione, Ron> 2



课程设计2 - 哈利波特的魔法世界

任务 3 特征处理：人物关系图构建与特征归一化

当获取了人物之间的共现关系之后，我们就可以根据共现关系，生成人物之间的关系图了。人物关系图使用邻接表的形式表示，方便后面的 PageRank 计算。在人物关系图中，人物是顶点，人物之间的互动关系是边。人物之间的互动关系靠人物之间的共现关系确定。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。边的权重越高则体现了两个人的关系越密切。

为了使后面的方便分析，还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，具体的过程见后面的示例。

输入输出

输入：任务 2 的输出

输出：归一化权重后的人物关系图



课程设计2 - 哈利波特的魔法世界

示例：

输入： <Harry, Ron> 2 <Ron, Harry> 2 <Harry, Hermione> 1
 <Hermione, Harry> 1 <Ron, Hermione> 2 <Hermione, Ron> 2

输出： Harry [Ron, 0.66667 | Hermione, 0.33333]
 Ron [Harry, 0.5 | Hermione, 0.5]
 Hermione [Harry, 0.33333 | Ron, 0.66667]

说明：

首先是将统计出的人物共现次数结果，转换成邻接表的形式表示，每一行表示一个邻接关系。Ron [Harry, 0.5 | Hermione, 0.5]表示顶点Ron有两个邻接点，分别是Harry和Hermione，对应两条邻接边，每条边有各自的权重。该权重是根据某个人与其他人共现的“概率”得到的，以Ron为例，他分别和Harry、Hermione共现2次，因此共现概率分别为 $2/(2+2)=0.5$, $2/(2+2)=0.5$ 。通过这种归一化，我们确保了某个顶点的出边权重的和为1。



课程设计2 - 哈利波特的魔法世界

任务 4 数据分析：基于人物关系图的 PageRank 计算

在给出人物关系图之后，我们就可以对人物关系图进行一个数据分析。其中一个典型的分析任务是：PageRank 值计算。通过计算 PageRank，我们就可以定量地分析出哈利波特系列小说的“主角”们是哪些。

输入输出

输入：任务 3 的输出

输出：人物的 PageRank 值

该任务默认的输出内容是杂乱的，从中无法直接的得到分析结论。可以对人物的 PageRank 值进行全局排序，从而很容易地发现 PageRank 值最高的几个人物。排序工作可以通过一个排序 MapReduce 程序完成，也可以将 PageRank 值导入 Hive 中，然后利用 Hive 完成排序。



课程设计2 - 哈利波特的魔法世界

任务 5 数据分析：在人物关系图上的标签传播（选做）

标签传播 (Label Propagation) 是一种半监督的图分析算法，他能为图上的顶点打标签，进行图顶点的聚类分析，从而在一张类似社交网络图中完成社区发现 (Community Detection)。图 1 中人物顶点的颜色就是根据标签传播的结果进行的染色。

参考资料

1. 英文资料：标签传播算法英文原始文献可参考[原始英文论文](#)中的 III. COMMUNITY DETECTION USING LABEL PROPAGATION 一节内容。
2. 中文资料：<http://www.cnphp6.com/archives/24136>

输入输出

输入：任务 3 的输出

输出：人物的标签信息

对于该任务的输出，可以通过写一个 MapReduce 程序，将属于同一个标签的人物输出到一起，便于人来查看标签传播的结果。

原始英文论文：<https://journals.aps.org/pre/abstract/10.1103/PhysRevE.76.036106>



课程设计2 - 哈利波特的魔法世界

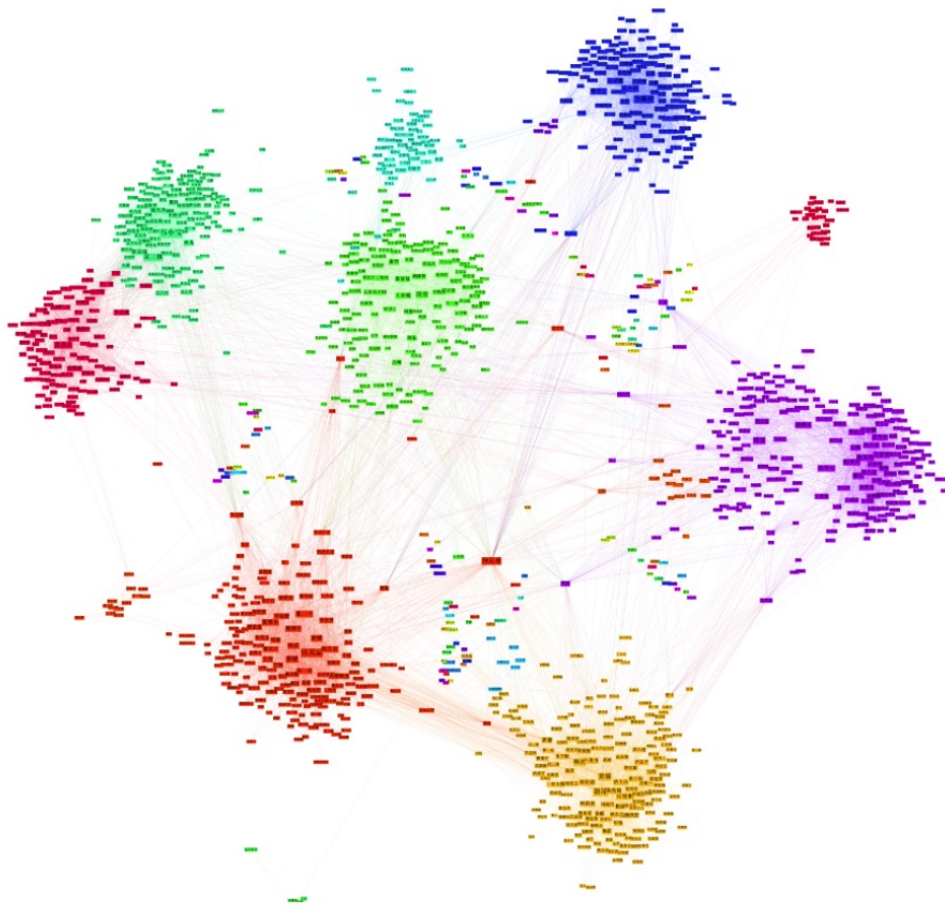


图 1|标签传播的结果展示

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 - 哈利波特的魔法世界

4 提交材料

请各位同学提交如下材料。

- 1、程序源代码，要求提供包含完整目录结构的 src 代码包，并且提供编译方法说明。
- 2、程序可执行 jar 包以及 jar 包的执行方式。本题目的运行环境在 hadoop-2.7、jdk-1.7 及以上的环境下，必须采用 MapReduce 编程模型。
- 3、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。


MapReduce课程设计

• 最终课题完成与提交

■ 课程设计结果提交（以下内容打包提交）

- **课程设计报告**，内容包括

1. 小组信息（人员，学号，联系信息，导师及研究领域）
 2. 课题小组分工：需要明确说明各成员在整个课题中分工负责完成的内容
 3. 课程设计题目
 4. 摘要
 5. 研究问题背景
 6. 主要技术难点和拟解决的问题，尤其要解释说明哪些地方、为什么需要采用MapReduce
 7. 主要解决方法和设计思路，尤其要解释说明如何采用MapReduce并行化算法解决问题
 8. 详细设计说明，包括详细算法设计、程序框架、功能模块、主要类的设计说明，包括主要类、函数的输入输出参数、**尤其是map和reduce函数的输入输出键值对详细数据格式和含义**，主要功能和算法代码中加清晰的注释说明。对于引用的部分，需要给出参考文献。
 9. **输入文件数据和详细输入数据格式**，输出结果文件数据片段和详细输出数据格式（**必须清晰描述**）
 10. 程序运行实验结果说明和分析
 11. 总结：特点总结，功能、性能、扩展性等方面存在的不足和可能的改进之处
 12. 参考文献
- **带注释的源程序（必须提交源程序以备检查实现情况，无源程序的以未完成课程设计处理）**
 - **输入数据文件和运行结果文件（必须提交输入输出文件数据，数据量太大可取部分数据）**
 - **执行程序**



严禁抄袭开源项目
或其他同学的课设
代码，违者本课程
一律0分计算!!!