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

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1—体育赛事日志分析

- 1. 课程设计目标

  本课程设计通过使用 MapReduce 实现比赛日志分析。

  通过本课程设计的学习，利用 MapReduce 工具实现大数据下的数据分析方法。

- 2. 学习技能

  本次课程设计可以掌握以下 MapReduce 编程技能：

  1. 海量日志数据的统计分析
  2. 基于 MapReduce 的预测模型设计

课程设计1—体育赛事日志分析

- 3. 题目描述
  - 各项体育赛事中，根据运动员在场上的具体表现情况，会产生大量的数据。在职业体育赛事中，对赛事过程中产生的日志进行分析，可以有效分析对手的技战术特点，从而可以帮助教练团队制定相应策略予以应对。
  - 本课程设计数据中记录了一系列篮球赛事的比赛日志，要求学生按照要求进行比赛日志的统计、分析，并根据已有的比赛日志预测后续赛事的比赛结果。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 以现实数据为例

1st Q				
Time	Boston	Score		Golden State
12:00.0	Jump ball: R. Williams vs. K. Looney (M. Smart gains possession)			
11:42.0	J. Tatum misses 2-pt jump shot from 21 ft	0-0		
11:40.0		0-0		Defensive rebound by S. Curry
11:25.0		0-0		A. Wiggins misses 3-pt jump shot from 27 ft
11:23.0	Defensive rebound by A. Horford	0-0		
11:15.0	J. Tatum misses 3-pt jump shot from 26 ft	0-0		
11:15.0		0-0		Defensive rebound by Team
11:06.0		0-0		S. Curry misses 3-pt jump shot from 26 ft
11:03.0		0-0		Offensive rebound by K. Looney
11:02.0		0-3	+3	S. Curry makes 3-pt jump shot from 26 ft (assist by K. Looney)
10:30.0	J. Brown misses 2-pt jump shot from 17 ft	0-3		
10:28.0	Offensive rebound by Team	0-3		
10:19.0	J. Tatum makes 3-pt jump shot from 28 ft (assist by A. Horford) +3	3-3		
10:01.0		3-5	+2	A. Wiggins makes 2-pt layup from 6 ft (assist by K. Looney)
9:50.0	M. Smart makes 3-pt jump shot from 26 ft (assist by J. Tatum) +3	6-5		
9:35.0		6-5		K. Looney misses 2-pt jump shot from 16 ft
9:32.0	Defensive rebound by J. Brown	6-5		
9:23.0	M. Smart misses 3-pt jump shot from 23 ft	6-5		
9:21.0		6-5		Defensive rebound by D. Green
9:17.0		6-8	+3	K. Thompson makes 3-pt jump shot from 23 ft (assist by S. Curry)
9:00.0	Personal foul by A. Wiggins (drawn by J. Tatum)	6-8		

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件的结构如下：
    - Date：比赛日期
    - AwayTeam 和 HomeTeam：参与比赛的球队，区分主客场
    - PlayBy：产生该条日志的球队名称
    - Quarter：事件发生的节次（1~4节，加时赛从5开始递增）
    - SecLeft：事件发生时该节的剩余时间（按秒计算）
    - 其它字段：根据不同的日志类型，会有不同的字段被填入。
  - 日志中的元素
    - 日期：从 2000/1/1 至 2000/6/25 不等。
    - 球队：从 team001 - team030，共 30 支球队。
    - 球员姓名：由日志生成器随机生成。保证所有球员姓名不重复。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 投篮事件
    - 篮板事件
    - 助攻事件
    - 封盖事件
    - 罚球事件
    - 犯规事件
    - 违例事件
    - 失误事件
    - 抢断事件

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务1：统计每场比赛的比赛结果
    - FreeThrowMade 为 make 时，PlayBy 球队得 1 分
    - ShotOutcome 为 make 时：
      - 若 ShotType 为 2-pt ***，PlayBy 球队得 2 分
      - 若 ShotType 为 3-pt ***，PlayBy 球队得 3 分
    - 针对每条日志计算得分情况后按比赛计算总得分
    - 提示：各支球队每天只会有一场比赛
    - 输出格式：日期，主队，主队比分，客队，客队比分

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务2：计算数据集中各项技术统计的前五名球员
    - 提示：根据每条日志产生行为的制造队员进行统计
    - 要求得到得分、篮板、助攻、抢断、盖帽最多的 5 名球员
  - 任务3：预测给定比赛的各队胜率
    - 根据已有的数据作为训练数据，设计预测算法，预测给定几组对阵中主队和客队的胜率。
    - 根据比较数据生成模型本身预测胜率的差值，判断模型的准确度，但本任务更看重算法设计部分。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务4：设计合理的评价标准，评选出表现最好的 5 名球员
    - 可以考虑的因素：
      - 球员技术统计数据
      - 球员所在球队的战绩
      - 实现或定义高阶指标，以评价球员表现
    - 根据评价标准，选出最好的 5 名球员（排名分先后）。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务5：分析 team025 和 team028 的比赛特点（选做）
    - 可以考虑的分析方向：
      - 分析两队球员的出场时间、投篮方式分布
      - 分析两队的轮换策略（各节偏好的出场球员）
      - 分析两队的关键球打法（比分接近时，比赛最后时刻由谁出手）
    - 根据分析出的比赛特点，从其中一队教练的角度出发，尝试提出对抗另一只球队的策略。

课程设计1—体育赛事日志分析

- 4. 提交作业
  - 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。

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。

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

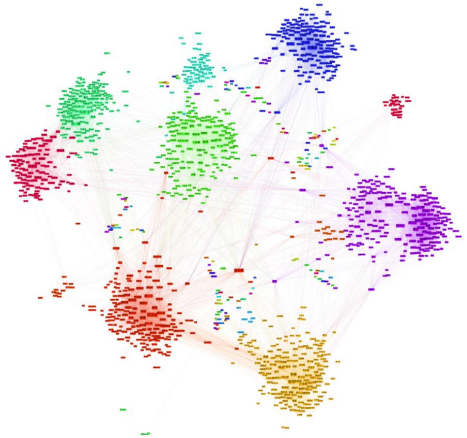


图 1|标签传播的结果展示

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。

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3—新闻自动分类

- 1. 课程设计目标

本课程设计的目标是通过 MapReduce 和基本的机器学习方法来实现对新闻的自动分类。通过本课程设计，可以学习如何使用 MapReduce 完成一个综合的数据挖掘任务，包括数据预处理，机器学习建模、样本预测等。

课程设计3—新闻自动分类

- 2. 课程设计目标

  通过本课程设计，可以熟悉或掌握以下 MapReduce 编程技巧：

  – 在 Hadoop 中使用第三方的 Jar 包来辅助分析

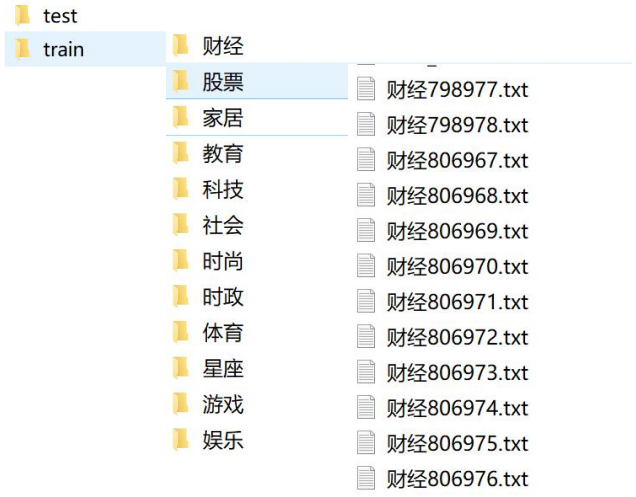
  – MapReduce 算法设计

    - 文本特征选择算法
    - 文本特征表示算法
    - 文本分类算法

课程设计3—新闻自动分类

- 3. 任务描述

在日常生活中，我们所看到的新闻通常伴随着相应类别，例如政治、经济、科教等等。不同的新闻包含不同的主题特征。本实验通过 MapReduce 技术实现新闻文本的自动分类。



新闻分类目录结构，每个新闻文件的文件名为新闻的类别+id；

课程设计3—新闻自动分类

- 3. 任务描述

  任务 1：文本特征选择

  - 对原始新闻中的文本进行特征选择，选择能够表征新闻特性的特征词，为后续的文本分类做准备。
  - 输入：新闻文本训练数据和测试数据；中文停用词表
  - 输出：新闻文本特征
  - 注意：需要过滤停用词表中的词

```
793    一下来  1
794    一丘之貉    1
795    一丝    1
796    一丝一毫    1
797    一个    31
798    一个个  1
799    一个人  5
800    一个又一个  1
801    一举    2
802    一举一动    1
803    一举两得    1
804    一件    1
805    一件事  1
806    一份    5
807    一会    2
808    一会儿  1
809    一位    11
810    一体    1
811    一倍    1
812    一元    1
813    一再    1
814    一再强调    1
815    一分钟  2
816    一切照旧    1
817    一切都  2
```

课程设计3—新闻自动分类

- 3. 任务描述

  任务 2：关键词提取

  - 在任务 1 的基础上，每个新闻类别提取 20 个最能够代表该类别的关键词。
  - 数据输入：新闻文本数据，以及任务 1 的输出
  - 数据输出：每个新闻类别的关键词

课程设计3—新闻自动分类

- 3. 任务描述

  任务 3：文本特征表示

  - 基于任务 1 得到的特征词，为每条新闻文本计算特征表示。
  - 数据输入：任务 1 的输出；新闻文本数据
  - 数据输出：每条新闻文本的特征向量

1	体育	天气:0.007142857142857143 奥运:0.007142857142857143 稀:0.0035714285714285713 正式:0.024999999999999
2	体育	堂:0.006666666666666667 占据:0.013333333333333334 热身赛:0.006666666666666667 信心:0.03 比赛:0.02 上
3	体育	仍然:0.006493506493506494 刻苦:0.0012987012987012987 正式:0.01818181818181818 以前:0.010389610389610
4	体育	言:0.0017921146953405018 20岁:0.0017921146953405018 战略性:0.0035842293906810036 迷茫:0.00179211469

样本特征表示

课程设计3—新闻自动分类

- 3. 任务描述

  任务 4：文本分类

  - 利用机器学习分类算法实现新闻文本的分类。具体采用何种分类算法，请同学们自行选择，也可以验证多种分类算法的优劣。
  - 基于以上得到的分类模型，对测试数据中的新闻文本进行预测，输出预测结果，并统计预测的正确率。

  任务 5：新闻自动分类的 Spark 实现（选做）

  - 考虑另外使用 Spark 实现新闻自动分类算法，并对新闻文本进行分类，输出分类结果。

课程设计3—新闻自动分类

- 4. 提交材料

  请各位同学提交如下材料：

  – 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。

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

谢 谢！