

# MapReduce 课程设计之金庸的江湖

第 11 小组

王 尧 151220112

王一栋 151220113

王宇鑫 151220114

## 实验介绍

此次课程设计，通过对金庸武侠小说中的人物关系挖掘，来学习 MapReduce 程序设计的相关算法并具体实现。实验流程包括对数据的预处理，数据的分析以及数据后处理。

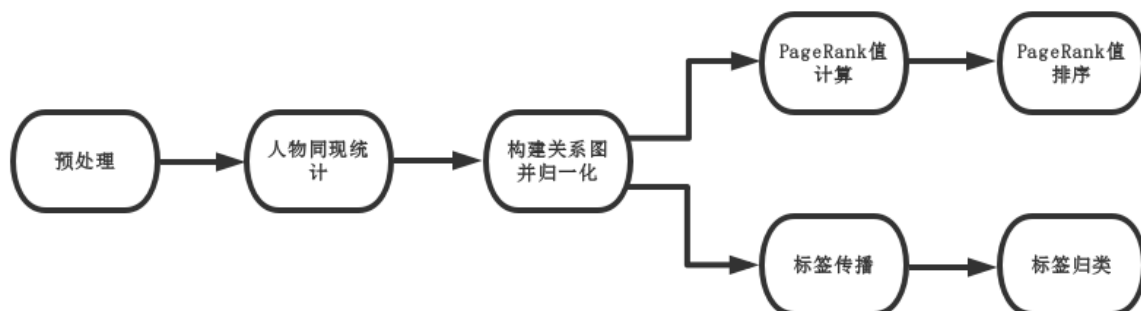
## 实验环境

- Hadoop 2.7.1
- Java 1.7
- 集群
- Python 3.6
- Gephi 0.9.2

## 目录

MapReduce 课程设计之金庸的江湖 .....	1
实验介绍.....	1
实验环境.....	1
实验流程及任务分配.....	3
编译方式及运行方法.....	4
实验过程.....	5
实验结果截图.....	11
实验结果分析.....	19
实验优化.....	20
实验总结.....	21

## 实验流程及任务分配



任务一：数据预处理，对文本进行分词并提取出人名。

任务二：人物同现统计，对出现在同一段落中的人物关系次数进行统计。

任务三：构建任务关系图并归一化，将人物关系用邻接表的形式存储，将共现次数转化为概率。

任务四：计算 PageRank 值，基于任务三的输出算出 PageRank 值。

任务五：标签传播，基于任务三的输出对人物进行聚类，为每个人物打上标签。

任务六：PageRank 排序，按照 PageRank 值从大到小排序，找出金庸小说中的“主角”。

任务七：标签归类，将每个人物按照标签归类在一起。

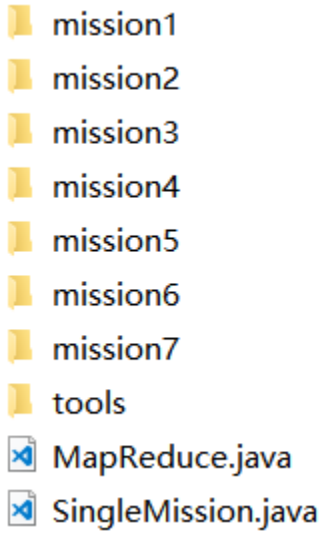
王尧负责：结果测试，完成任务四，任务六，任务七，标签数据可视化。

王一栋负责：完成任务五，任务七，标签数据可视化。

王宇鑫负责：代码框架搭建，结果测试，完成任务一，任务二，任务三，任务五，标签数据可视化。

## 编译方式及运行方法

代码结构如下图



编译方法为在根目录运行 `mvn clean package` 得到 `wuxia.jar`。

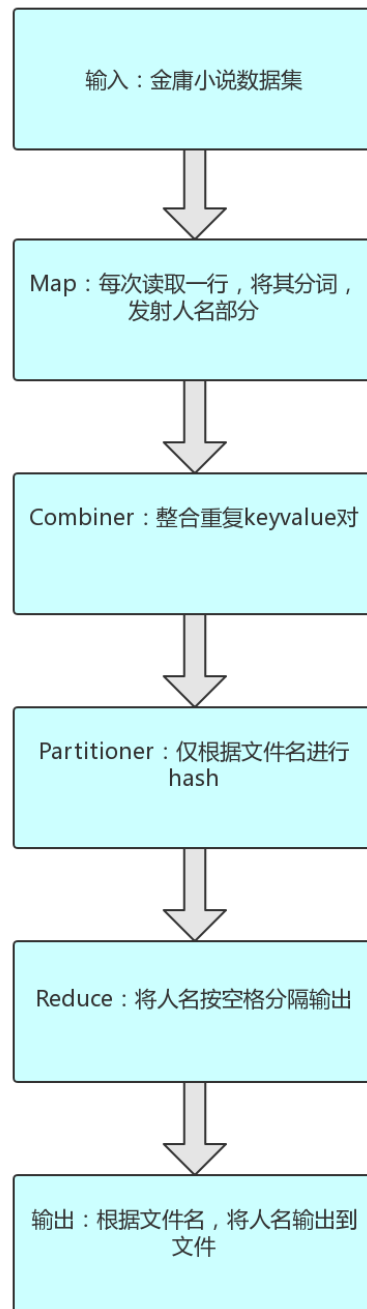
Hadoop 集群上运行 jar 包命令: `hadoop jar wuxia.jar MapReduce ${InputPath} ${OutputPath}`

项目地址可见: <https://github.com/NJUA422Hadoop/Assignment3>

# 实验过程

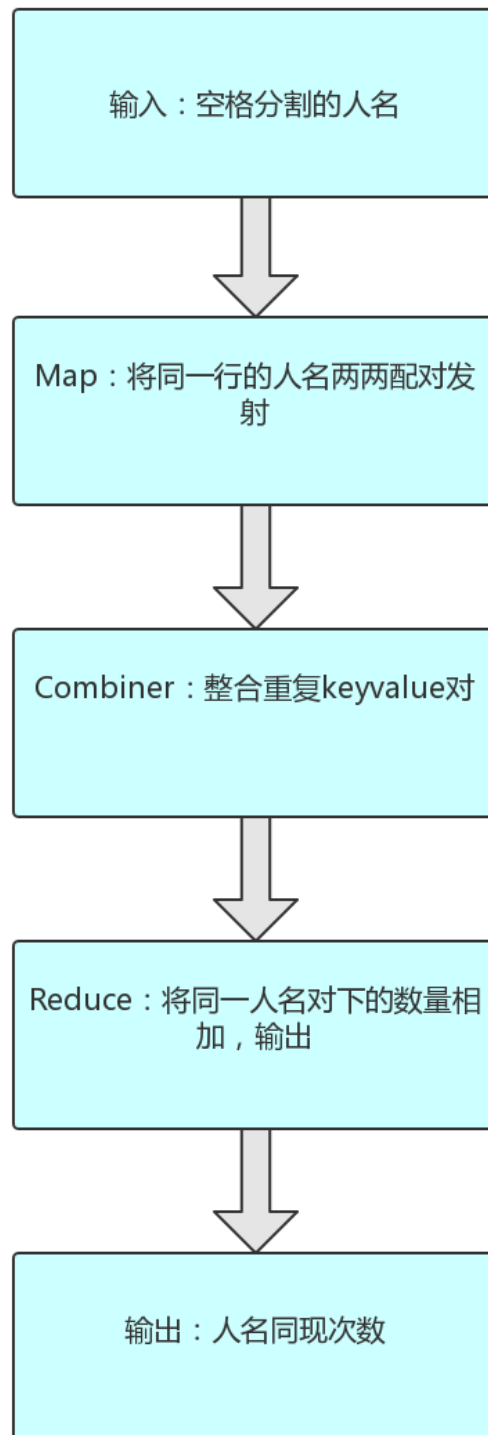
## 1. 数据预处理

实现于 Mission1 中。



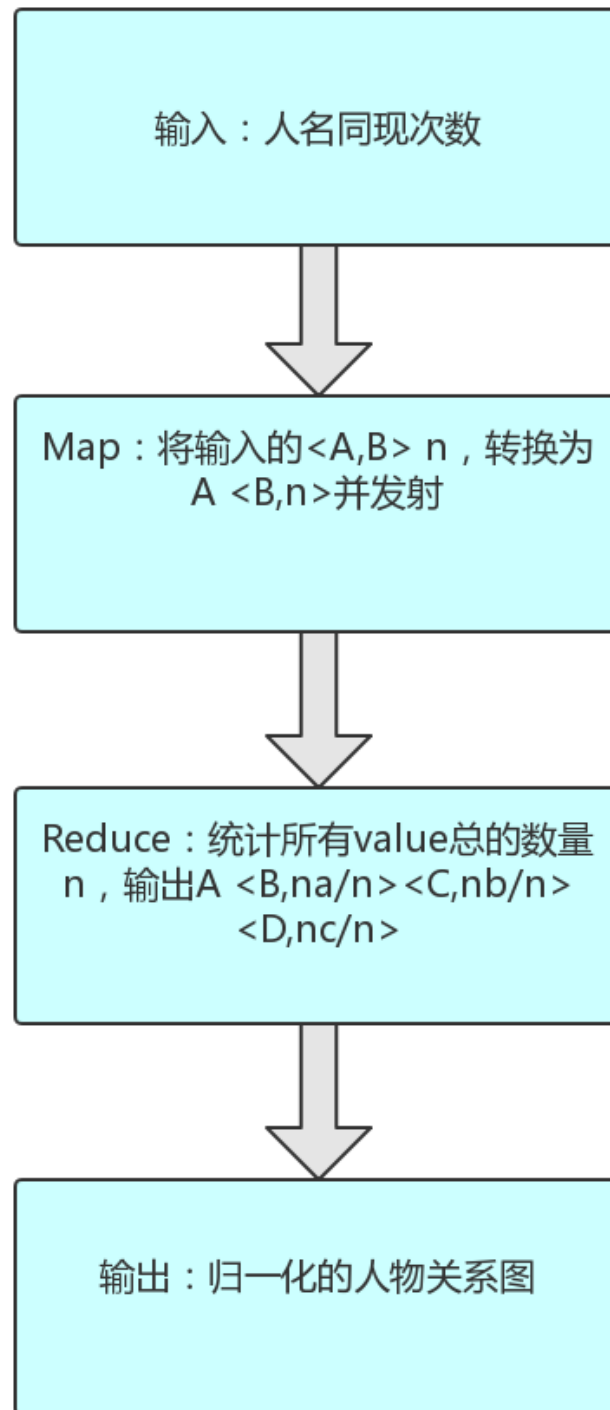
## 2. 人物同现统计

实现于 Mission2 中。



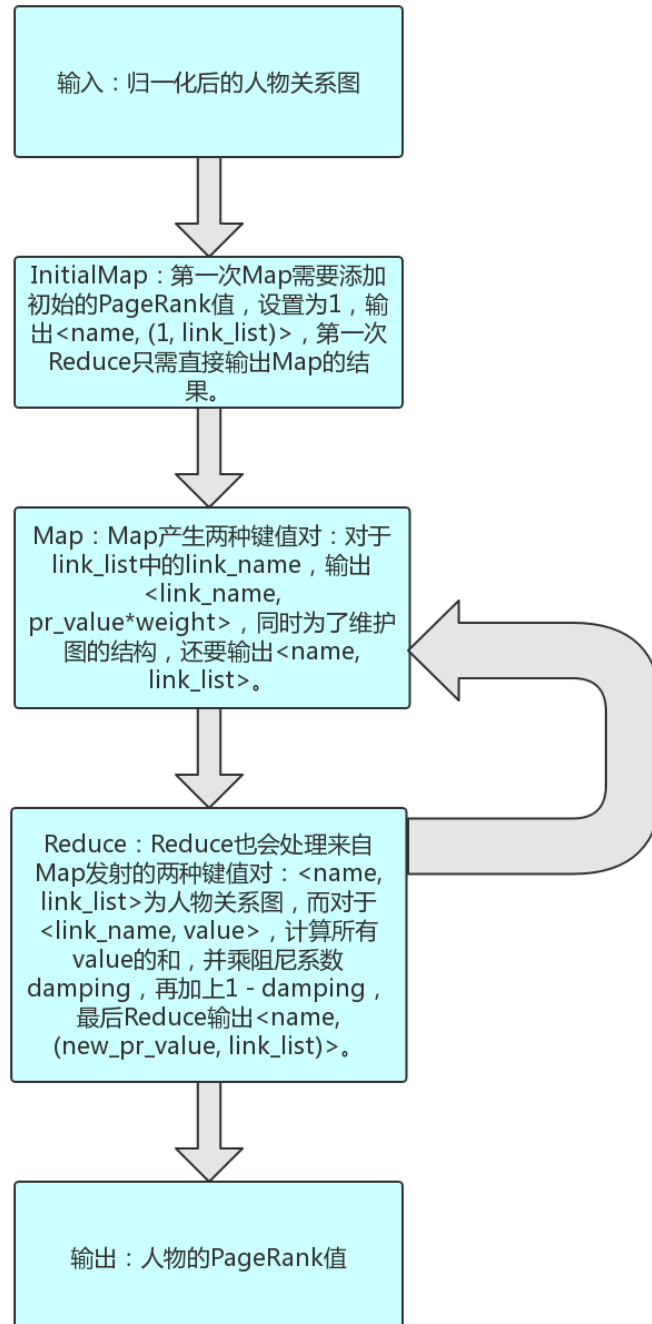
### 3. 构建人物关系图并归一化

实现于 Mission3 中。



#### 4. PageRank 值计算

实现于 Mission4 中。

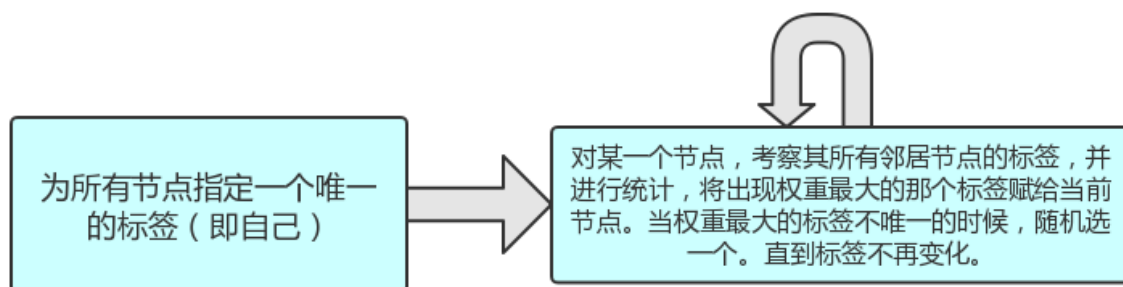


关于迭代次数，事先设置为 30 轮，在最后查看结果文件时发现在进行到第 5 轮时 PageRank 值就已经基本收敛。



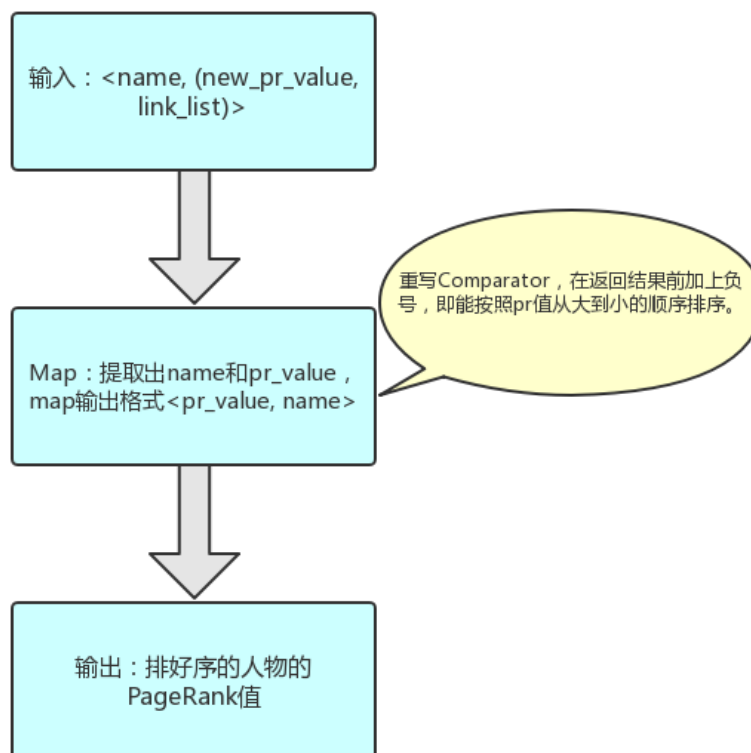
## 5. 标签传播

实现于 Mission5 中。



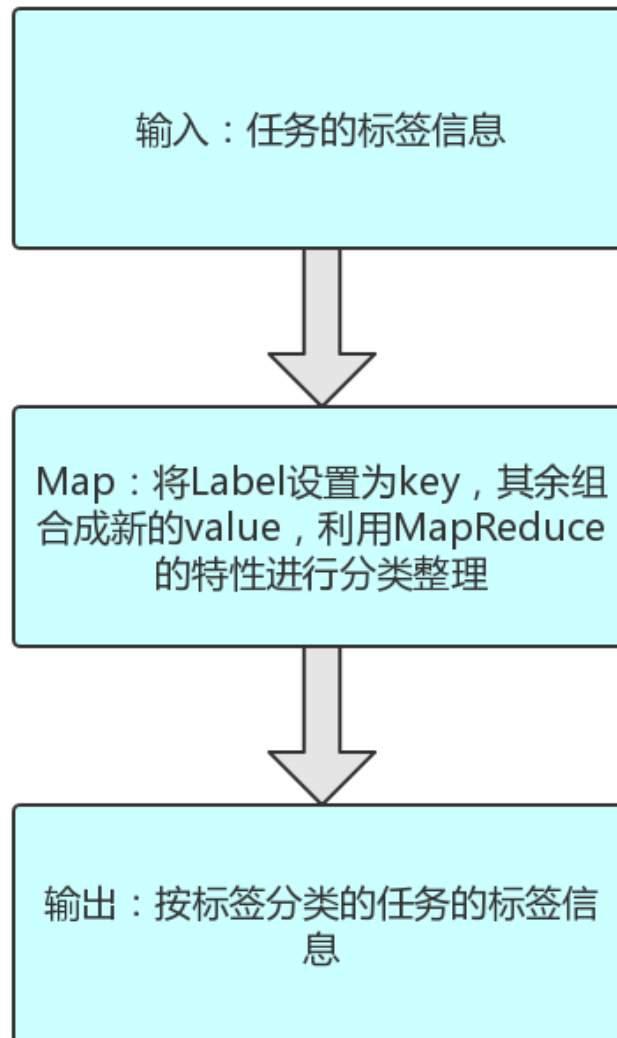
## 6. PageRank 值排序

实现于 Mission6 中。



## 7. 标签归类

实现于 Mission7 中。



## 8. 可视化处理

实现于 node.py 中。

首先对 mission7 的结果用 python 处理（由于该部分只涉及对数据处理而和 MapReduce 设计无关，故使用 python 编写），得到两个文件分别代表结点文件和边文件。结点文件的信息包括人名及其所属标签。边文件的信息包括起点，终点以及边的权重，其中权重是由入度乘边的权重获得。

```
maps[v_name] += v_weight
```

将结点文件和边文件导入 Gephi 软件，选择以标签分类染色，并采用 Force Atlas 和 Yifan Hu 布局，将斥力强度调整为 500，输出最后的可视化图，对结点大小以及线的粗细进行处理后得到最终结果。

## 实验结果截图

### 1. 数据预处理

```
1 胡一刀
2 苗人凤
3 胡一刀 苗人凤 大汉 汉子
4 苗人凤 胡一刀
5 汉子
6 汉子 汉子 汉子 胡一刀 苗人凤
7 汉子
8 汉子 汉子
9 汉子 脚夫
10 马春花 汉子 马行空
11 徐铮
12 徐铮
13 徐铮 马行空 马行空 徐铮 马春花
14 马春花
15 少妇 马春花 少妇 少妇 少妇
16 少妇 马春花 少妇 马春花 少妇 马春花
17 少妇
18 徐铮 马行空 脚夫
19 徐铮
20 徐铮 徐铮
```

application_1528693522781_1334	2018st11	Mission1_1	MAPREDUCE	root.2018st11	Wed Jul 11 14:25:38 +0800 2018	Wed Jul 11 14:26:06 +0800 2018	FINISHED SUCCEEDED	<div></div>	History
--------------------------------	----------	------------	-----------	---------------	--------------------------------------	--------------------------------------	--------------------	-------------	---------

## 2. 人物同现统计

1	<一灯大师,上官>	1
2	<一灯大师,丘处机>	5
3	<一灯大师,乔寨主>	1
4	<一灯大师,农夫>	16
5	<一灯大师,华筝>	1
6	<一灯大师,卫璧>	2
7	<一灯大师,吕文德>	1
8	<一灯大师,周伯通>	26
9	<一灯大师,哑巴>	1
10	<一灯大师,哑梢公>	1
11	<一灯大师,大汉>	1
12	<一灯大师,天竺僧>	1
13	<一灯大师,天竺僧人>	3
14	<一灯大师,完颜萍>	2
15	<一灯大师,小沙弥>	3
16	<一灯大师,小龙女>	8
17	<一灯大师,尹克西>	1
18	<一灯大师,尼摩星>	3
19	<一灯大师,张无忌>	2
20	<一灯大师,无色>	1

application_1528693522781_1335	2018st11	Mission2_1	MAPREDUCE	root.2018st11	Wed Jul 11 14:26:10 +0800 2018	Wed Jul 11 14:26:34 +0800 2018	FINISHED SUCCEEDED	<div></div>	History
--------------------------------	----------	------------	-----------	---------------	--------------------------------------	--------------------------------------	--------------------	-------------	---------

## 3. 构建人物关系图并归一化

1	一灯大师	[丘处机:0.012224939 农夫:0.039119806 华筝:0.002444988 卫璧:0.004889976 吕文德:0.00244
2	丁不三	[小翠:0.008403362 柯万钧:0.008403362 汉子:0.016806724 白万剑:0.20168068 白自在:0.01680672
3	丁不四	[龙岛主:0.015748031 大汉:0.007874016 天虚:0.003937008 封万里:0.011811024 小翠:0.031496063
4	丁典	[桃红:0.004950495 万圭:0.03960396 万震山:0.00990099 农夫:0.004950495 凌退思:0.05940594 凌
5	丁勉	[丛不弃:0.028169014 令狐冲:0.07042254 任我行:0.014084507 何足道:0.014084507 刘正风:0.1126
6	丁同	[老头子:0.125 李三:0.25 霍元龙:0.125 李文秀:0.5]
7	丁坚	[左冷禅:0.01724138 丹青生:0.13793103 令狐冲:0.22413793 任我行:0.01724138 向问天:0.1034482
8	丁大全	[宋五:0.0625 大汉:0.25 小王将军:0.125 汉子:0.0625 王惟忠:0.1875 郭襄:0.0625 陈大方:0.25]
9	丁敏君	[俞岱岩:0.004878049 卫璧:0.009756098 吊死鬼:0.004878049 周芷若:0.10243902 宋远桥:0.009756
10	丁春秋	[康广陵:0.01983471 不平道人:0.0049586776 乌老大:0.008264462 乔峰:0.0066115703 何足道:0.00
11	丁游	[刘培生:0.083333336 大汉:0.083333336 孟伯飞:0.083333336 孟铮:0.083333336 孟铸:0.083333336
12	万圭	[沈城:0.039007094 丁典:0.028368793 万震山:0.18439716 冯坦:0.0070921984 凌退思:0.007092198
13	万大平	[刘菁:0.2 汉子:0.2 刘正风:0.4 史登达:0.2]
14	万庆澜	[余鱼同:0.015267176 卫春华:0.053435113 吴国栋:0.007633588 骆冰:0.053435113 周仲英:0.13740
15	万里风	[何铁手:0.041666668 刘培生:0.083333336 梅剑和:0.125 洞玄:0.083333336 焦公礼:0.125 袁承志:
16	万震山	[万圭:0.19259259 冯坦:0.0074074073 凌退思:0.0074074073 卜垣:0.033333335 吴坎:0.107407406
17	上官	[王进宝:0.0046296297 一灯大师:0.0046296297 万庆澜:0.0046296297 上官云:0.06481481 上官毅山
18	上官云	[不戒和尚:0.005076142 东方不败:0.1573604 于嫂:0.005076142 仇松年:0.005076142 令狐冲:0.157
19	上官毅山	[阎世章:0.02173913 上官:0.10869565 余鱼同:0.10869565 兆惠:0.02173913 周仲英:0.0217391
20	上官虹	[霍元龙:0.21052632 史仲俊:0.36842105 陈达海:0.15789473 李三:0.2631579]

application_1528693522781_1336	2018st11	Mission3_1	MAPREDUCE	root.2018st11	Wed Jul 11 14:26:42 +0800 2018	Wed Jul 11 14:27:12 +0800 2018	FINISHED SUCCEEDED	<div></div>	History
--------------------------------	----------	------------	-----------	---------------	--------------------------------------	--------------------------------------	--------------------	-------------	---------

## 4. PageRank 值计算

1	一灯大师	1.4149134449394731	[丘处机:0.012224939 农夫:0.039119806 华筝:0.002444988 卫璧:0.004889
2	丁不三	1.1799750714684967	[小翠:0.008403362 柯万钧:0.008403362 汉子:0.016806724 白万剑:0.20168068
3	丁不四	2.1699355718393956	[龙岛主:0.015748031 大汉:0.007874016 天虚:0.003937008 封万里:0.01181102
4	丁典	1.6126911775481503	[桃红:0.004950495 万圭:0.03960396 万震山:0.00990099 农夫:0.004950495 凌
5	丁勉	0.6509027223562488	[丛不弃:0.028169014 令狐冲:0.07042254 任我行:0.014084507 何足道:0.01408
6	丁同	0.2284500979537844	[老头子:0.125 李三:0.25 霍元龙:0.125 李文秀:0.5]
7	丁坚	0.4307438615046898	[左冷禅:0.01724138 丹青生:0.13793103 令狐冲:0.22413793 任我行:0.0172413
8	丁大全	0.5752989087595519	[宋五:0.0625 大汉:0.25 小王将军:0.125 汉子:0.0625 王惟忠:0.1875 郭襄:0.
9	丁敏君	1.0357788423738132	[俞岱岩:0.004878049 卫璧:0.009756098 吊死鬼:0.004878049 周芷若:0.102439
10	丁春秋	3.2940014411162934	[康广陵:0.01983471 不平道人:0.0049586776 乌老大:0.008264462 乔峰:0.0066
11	丁游	0.23696598611510428	[刘培生:0.083333336 大汉:0.083333336 孟伯飞:0.083333336 孟铮:0.08333333
12	万圭	1.9190502437297141	[沈城:0.039007094 丁典:0.028368793 万震山:0.18439716 冯坦:0.0070921984
13	万大平	0.24055130383080509	[刘菁:0.2 汉子:0.2 刘正风:0.4 史登达:0.2]
14	万庆澜	0.6442932761831222	[余鱼同:0.015267176 卫春华:0.053435113 吴国栋:0.007633588 骆冰:0.053435
15	万里风	0.30647847473906664	[何铁手:0.041666668 刘培生:0.083333336 梅剑和:0.125 洞玄:0.083333336 焦
16	万震山	1.8698632232834287	[万圭:0.19259259 冯坦:0.0074074073 凌退思:0.0074074073 卜垣:0.033333335
17	上官	1.217383114558634	[王进宝:0.0046296297 一灯大师:0.0046296297 万庆澜:0.0046296297 上官云:0
18	上官云	1.0467627466961522	[不戒和尚:0.005076142 东方不败:0.1573604 于嫂:0.005076142 仇松年:0.0050
19	上官毅山	0.3293859206252841	[阎世章:0.02173913 上官:0.10869565 余鱼同:0.10869565 兆惠:0.0217391
20	上官虹	0.43174708763786845	[霍元龙:0.21052632 史仲俊:0.36842105 陈达海:0.15789473 李三:0.2631579]

application_1528693522781_1377	2018st11	Mission4_30	MAPREDUCE	root.2018st11	Wed Jul 11 14:46:05 +0800 2018	Wed Jul 11 14:46:34 +0800 2018	FINISHED SUCCEEDED	<div></div>	History
--------------------------------	----------	-------------	-----------	---------------	--------------------------------------	--------------------------------------	--------------------	-------------	---------

## 5. 标签传播

1	一灯大师	黄蓉	[丘处机,黄蓉:0.012224939 农夫,黄蓉:0.039119806 华筝,黄蓉:0.002444988 卫璧,赵敏:0.004889976
2	丁不三	石破天	[小翠,石破天:0.008403362 柯万钧,闵柔:0.008403362 汉子,韦小宝:0.016806724 白万剑,石破天:0.20168
3	丁不四	石破天	[龙岛主,龙岛主:0.015748031 大汉,韦小宝:0.007874016 天虚,闵柔:0.003937008 封万里,石破天:0.01181
4	丁典	狄云	[桃红,狄云:0.004950495 万圭,狄云:0.03960396 万震山,狄云:0.00990099 农夫,黄蓉:0.004950495 凌退思
5	丁勉	岳不群	[丛不弃,岳不群:0.028169014 令狐冲,岳不群:0.07042254 任我行,岳不群:0.014084507 何足道,杨过:0.01
6	丁同	苏普	[老头子,岳不群:0.125 李三,史仲俊:0.25 霍元龙,苏普:0.125 李文秀,苏普:0.5]
7	丁坚	岳不群	[左冷禅,岳不群:0.01724138 丹青生,岳不群:0.13793103 令狐冲,岳不群:0.22413793 任我行,岳不群:0.01
8	丁大全	韦小宝	[宋五,杨过:0.0625 大汉,韦小宝:0.25 小王将军,韦小宝:0.125 汉子,韦小宝:0.0625 王惟忠,韦小宝:0.18
9	丁敏君	赵敏	[俞岱岩,赵敏:0.004878049 卫璧,赵敏:0.009756098 吊死鬼,史季强:0.004878049 周芷若,赵敏:0.1024390
10	丁春秋	王语嫣	[康广陵,王语嫣:0.01983471 不平道人,王语嫣:0.0049586776 乌老大,王语嫣:0.008264462 乔峰,阿朱:0.0
11	丁游	袁承志	[刘培生,袁承志:0.083333336 大汉,韦小宝:0.083333336 孟伯飞,袁承志:0.083333336 孟铮,袁承志:0.083
12	万圭	狄云	[沈城,狄云:0.039007094 丁典,狄云:0.028368793 万震山,狄云:0.18439716 冯坦,狄云:0.0070921984 凌退
13	万大平	岳不群	[刘菁,岳不群:0.2 汉子,韦小宝:0.2 刘正风,岳不群:0.4 史登达,岳不群:0.2]
14	万庆澜	陈家洛	[余鱼同,陈家洛:0.015267176 卫春华,陈家洛:0.053435113 吴国栋,陈家洛:0.007633588 骆冰,陈家洛:0.0
15	万里风	袁承志	[何铁手,袁承志:0.041666668 刘培生,袁承志:0.083333336 梅剑和,袁承志:0.125 洞玄,袁承志:0.08333333
16	万震山	狄云	[万圭,狄云:0.19259259 冯坦,狄云:0.0074074073 凌退思,狄云:0.0074074073 卜垣,狄云:0.033333335 吴
17	上官	岳不群	[王进宝,韦小宝:0.0046296297 一灯大师,黄蓉:0.0046296297 万庆澜,陈家洛:0.0046296297 上官云,岳不群
18	上官云	岳不群	[不戒和尚,岳不群:0.005076142 东方不败,岳不群:0.1573604 于嫂,岳不群:0.005076142 仇松年,张夫人:0
19	上官毅山	陈家洛	[阎世章,陈家洛:0.02173913 上官,岳不群:0.10869565 余鱼同,陈家洛:0.10869565 兆惠,陈家洛:0.02
20	上官虹	史仲俊	[霍元龙,苏普:0.21052632 史仲俊,史仲俊:0.36842105 陈达海,苏普:0.15789473 李三,史仲俊:0.2631579]

application_1528693522781_1351	2018st11	Mission5_6	MAPREDUCE	root.2018st11	Wed Jul 11 14:34:18 +0800 2018	Wed Jul 11 14:34:46 +0800 2018	FINISHED SUCCEEDED	<div></div>	History
--------------------------------	----------	------------	-----------	---------------	--------------------------------------	--------------------------------------	--------------------	-------------	---------

## 6. PageRank 值排序

1	34.843025	韦小宝
2	20.783754	令狐冲
3	20.242754	张无忌
4	15.86831	郭靖
5	13.933192	黄蓉
6	13.905119	袁承志
7	13.8790655	段誉
8	13.833133	杨过
9	13.174056	汉子
10	12.927321	胡斐
11	9.442011	陈家洛
12	7.7461405	吴三桂
13	7.3062286	岳不群
14	7.2682424	石破天
15	6.963841	赵敏
16	6.425451	谢逊
17	6.254323	小龙女
18	6.0788293	狄云
19	5.7846775	虚竹
20	5.627946	徐天宏

application\_1528693522781\_1378 2018st11 Mission6\_1 MAPREDUCE root.2018st11 Wed Jul 11 14:46:37 +0800 2018 Wed Jul 11 14:46:56 +0800 2018 FINISHED SUCCEEDED History

## 7. 标签归类

1	倪不小小	倪不大	[文醉翁,福康安:0.025641026 余鱼同,陈家洛:0.025641026 倪不小小,倪不小小:0.20]
2	倪不小小	西灵道人	[倪不大大,倪不小小:0.16666667 倪不小小,倪不小小:0.16666667 哈赤大师,福康安:0.025641026 常伯志,常赫志:0.025641026 余鱼同,陈家洛:0.04347826 大痴,陈家洛:0.04347826 大癡,陈家洛:0.04347826]
3	倪不小小	倪不小小	[哈赤大师,福康安:0.025641026 常伯志,常赫志:0.025641026 余鱼同,陈家洛:0.04347826 于万亨,陈家洛:0.04347826 余鱼同,陈家洛:0.04347826 元伤,元痛:0.08695652]
4	元痛	元悲	[周仲英,陈家洛:0.04347826 大痴,陈家洛:0.04347826 大癡,陈家洛:0.04347826]
5	元痛	元痛	[于万亨,陈家洛:0.04347826 余鱼同,陈家洛:0.04347826 元伤,元痛:0.08695652]
6	元痛	元伤	[余鱼同,陈家洛:0.04347826 于万亨,陈家洛:0.04347826 元悲,元痛:0.08695652]
7	冯不破	冯不摧	[石骏,袁承志:0.0754717 何惕守,袁承志:0.0754717 冯不破,冯不破:0.16981132]
8	冯不破	冯不破	[何惕守,袁承志:0.06382979 冯不摧,冯不破:0.19148937 冯难敌,袁承志:0.1276]
9	史仲俊	李三	[史仲俊,史仲俊:0.1875 大汉,韦小宝:0.03125 少妇,韦小宝:0.03125 明珠,韦小宝:0.21052632 史仲俊,史仲俊:0.36842105 陈达海,苏普:0.15789473]
10	史仲俊	上官虹	[霍元龙,苏普:0.21052632 史仲俊,史仲俊:0.36842105 陈达海,苏普:0.15789473]
11	史仲俊	史仲俊	[少妇,韦小宝:0.045454547 上官虹,史仲俊:0.3181818 李三,史仲俊:0.27272728]
12	史季强	讨债鬼	[史仲猛,史季强:0.33333334 催命鬼,史季强:0.33333334 丧门鬼,史季强:0.33333334]
13	史季强	丧门鬼	[笑脸鬼,杨过:0.18181819 俏鬼,杨过:0.09090909 催命鬼,史季强:0.18181819 史仲猛,史季强:0.25 大汉,韦小宝:0.25 何足道,杨过:0.25 杨过,杨过:0.25]
14	史季强	史季强	[杨过,杨过:0.094339624 丧门鬼,史季强:0.018867925 俏鬼,杨过:0.03773585 催命鬼,史季强:0.055555556 俏鬼,杨过:0.027777778 催命鬼,史季强:0.055555556]
15	史季强	史仲猛	[丧门鬼,史季强:0.024390243 俏鬼,杨过:0.048780486 韦小宝,韦小宝:0.097560]
16	史季强	吊死鬼	[丧门鬼,史季强:0.024390243 俏鬼,杨过:0.048780486 韦小宝,韦小宝:0.097560]
17	史季强	煞神鬼	[史仲猛,史季强:0.25 大汉,韦小宝:0.25 何足道,杨过:0.25 杨过,杨过:0.25]
18	史季强	催命鬼	[丧门鬼,史季强:0.22222222 俏鬼,杨过:0.11111111 史仲猛,史季强:0.22222222]
19	夏青	欧阳牧之	[夏青,夏青:0.06666667 司徒千钟,夏青:0.13333334 唐文亮,赵敏:0.06666667]
20	夏青	叶长青	[谢逊,赵敏:0.2 空智,空闻:0.2 汉子,韦小宝:0.2 司徒千钟,夏青:0.4]

## 8. 可视化

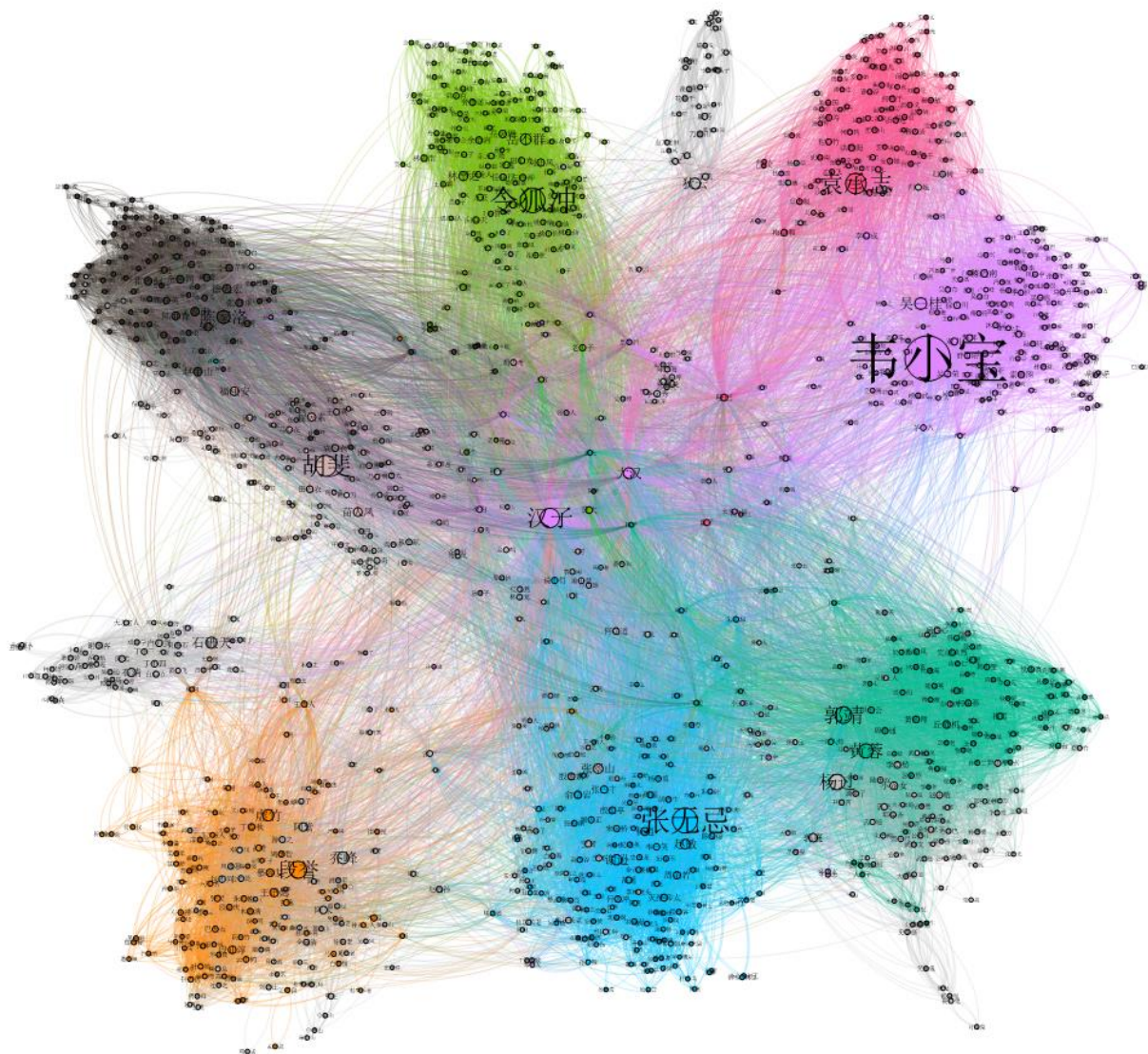
1	ID,label,name,weight
2	文醉翁,文醉翁,福康安,0.711113219
3	余鱼同,余鱼同,陈家洛,3.906255030999999
4	倪不大,倪不大,倪不大,0.76135364005
5	哈赤大师,哈赤大师,福康安,0.339060635
6	常伯志,常伯志,常赫志,0.8809688637000003
7	常赫志,常赫志,常赫志,0.9179330415000001
8	德布,德布,胡斐,0.10666922595
9	文泰来,文泰来,陈家洛,4.889691563099998
10	桑飞虹,桑飞虹,胡斐,1.1441501841

上图为结点文件格式

1	source,target
2	倪不大,文醉翁
3	倪不大,余鱼同
4	倪不大,倪不大
5	倪不大,哈赤大师
6	倪不大,常伯志
7	倪不大,常赫志
8	倪不大,德布
9	倪不大,文泰来
10	倪不大,桑飞虹
11	倪不大,汤沛
12	倪不大,海兰弼
13	倪不大,福康安
14	倪不大,程灵素
15	倪不大,胡斐
16	倪不大,西灵道人
17	倪不大,赵半山
18	倪不大,郭玉堂
19	倪不大,陆菲青
20	倪不大,陈家洛

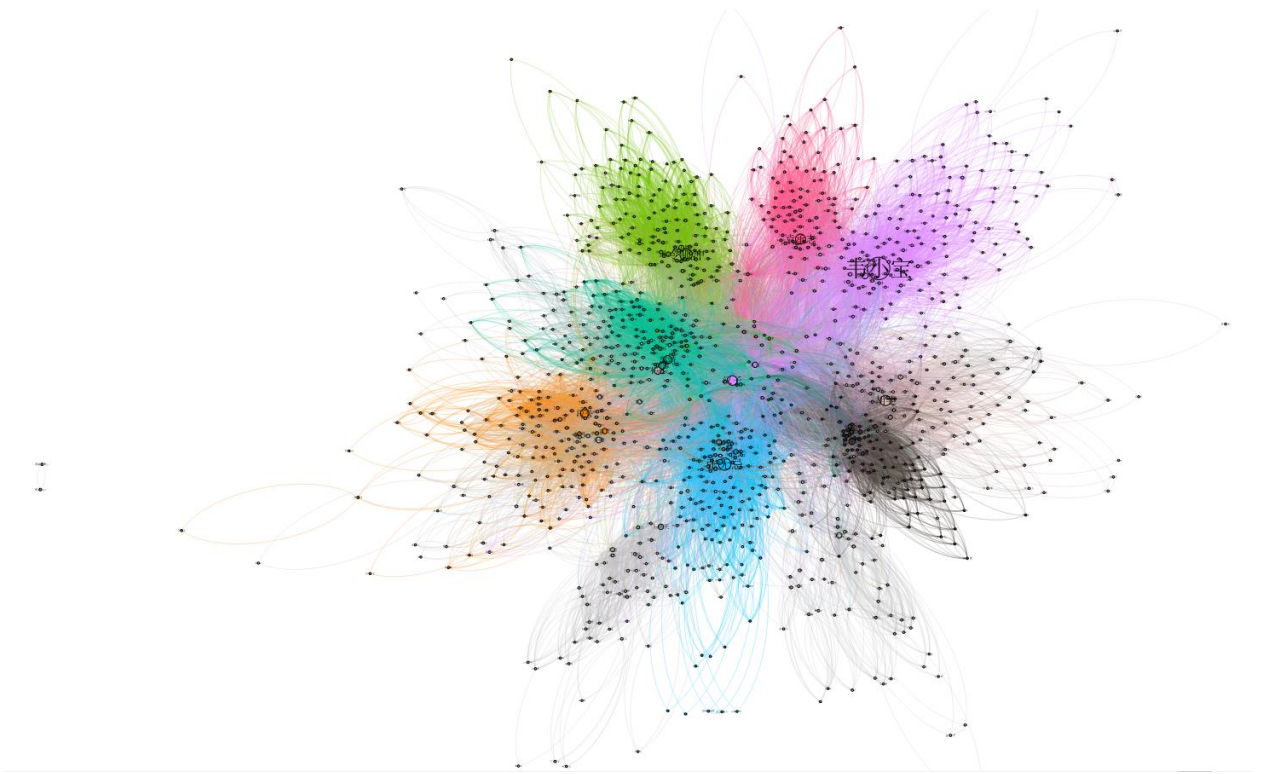
上图为边结点格式





上图为 Force Atlas 布局结果，图中具体细节可参见 Force Atlas.pdf。



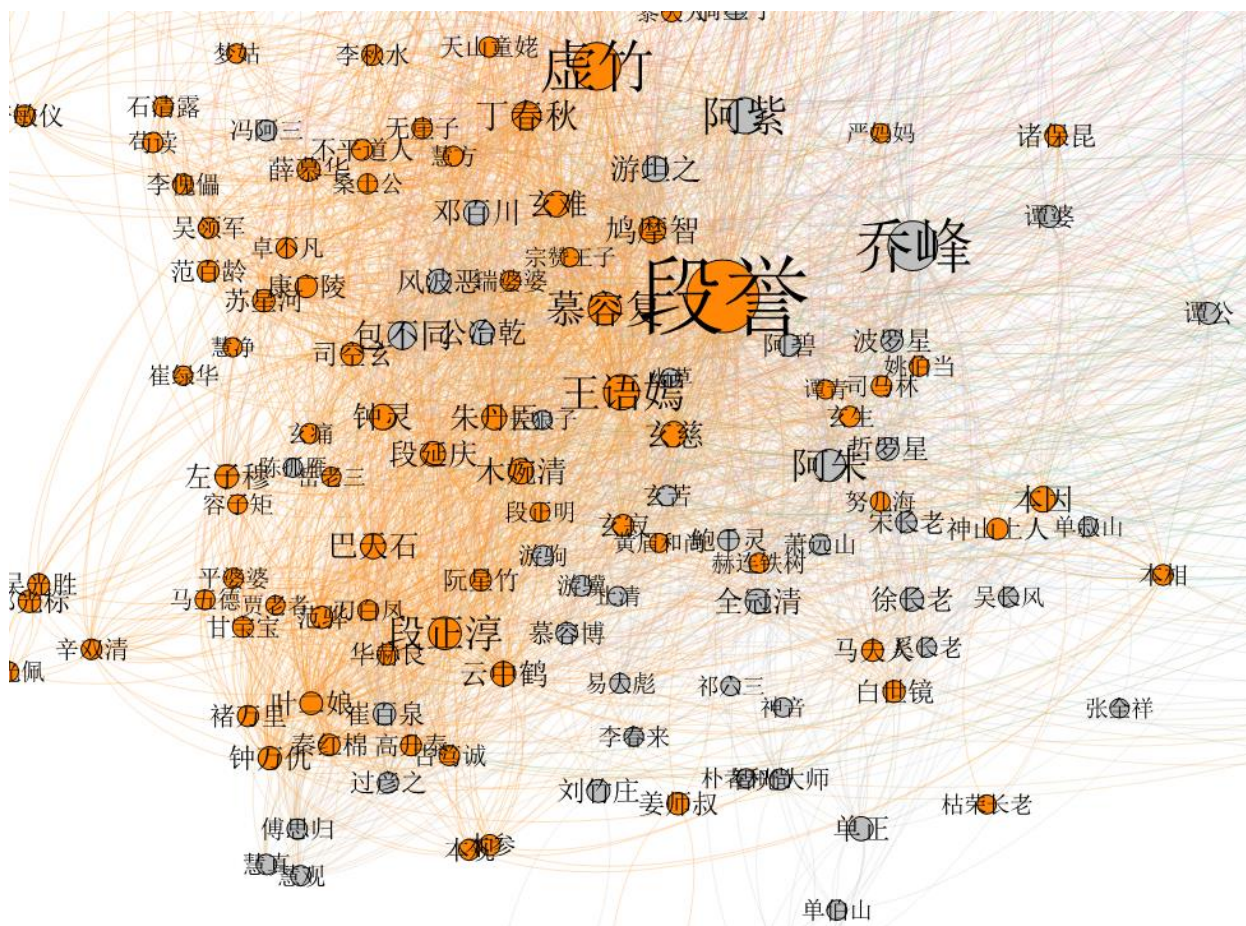


上图为 Yifan Hu 布局结果，图中具体细节可参见 Yifan Hu.pdf。









以上为局部截图。

## 实验结果分析

根据 PageRank 的结果，可以看出各大小小说的主角 PageRank 值都排在前列，符合常识。其中“汉子”是金庸对某些男性的代称，基本出现在每部小说中，因此 PageRank 值也较高。

根据标签传播可视化的结果，可以看出各本小说的人物基本都归在同一类。

图的中心基本都是金庸对某些人物的代称，例如“汉子”，“大汉”，“胖子”，“少妇”等。

《神雕侠侣》和《射雕英雄传》两部小说中由很多人物交叉，但根据可视化结果来看，两部小说的主要人物都被正确地分在了两个不同类别中。

《天龙八部》小说中，段誉和乔峰是两大主角，在同一部小说中却被分为了不同类别。根据网上资料及自己猜测，可能是因为小说中两者戏份相当，网上对两者谁是主角争议也较大。

















## 实验优化

Mission1: 由于有 15 个文件，并且相互无关，如果都输出到一个文件里，会降低并行度，所以将其输出到 15 个文件中，这样 Mission2 处理可以并行进行。

lt > mission1

名称

^

-  \_SUCCESS
-  金庸01飞狐外传.txt
-  金庸02雪山飞狐.txt
-  金庸03连城诀.txt
-  金庸04天龙八部.txt
-  金庸05射雕英雄传.txt
-  金庸06白马啸西风.txt
-  金庸07鹿鼎记.txt
-  金庸08笑傲江湖.txt
-  金庸09书剑恩仇录.txt
-  金庸10神雕侠侣.txt
-  金庸11侠客行.txt
-  金庸12倚天屠龙记.txt
-  金庸13碧血剑.txt
-  金庸14鸳鸯刀.txt
-  金庸15越女剑.txt

Mission5: 一开始标签传播是将标签附加在字符串中: A A [B,B:0.1|C,C:0.3]

打印在文件里, 每次从字符串中寻找标签值, 并且同时将更新的值存储在 map 里, 不断更新。

但由于这样单次 map 后保存下来的结果不是全局范围的, 后更新的 label 无法更新到先更新的 label 中, 所以考虑将 map 持久化在 hdfs 上, 每次将该 map 读入内存, 保证 label 最新。

收敛加快许多。

## 实验总结

通过此次实验, 我们理解了如何处理一个综合性大数据问题的步骤, 包括对数据的预处理, 计算, 以及后续处理可视化等等。在这个过程中, 我们锻炼了 MapReduce 程序设计的能力, 同时学会了使用各种辅助工具。

### 参考资料

[1] 《深入理解大数据 大数据处理与编程实践》黄宜华

[2] Gephi 的 wiki 主页: <https://github.com/gephi/gephi/wiki>

[3] 标签传播算法介绍: <https://blog.csdn.net/sejuezai9708/article/details/78327204>