大数据

综合处理实验报告

实验课题 -----《金庸的江湖》

131220044 查鹏（组长）

131220034 许金强

目录

一、实验目标

通过一个综合数据分析案例：“金庸的江湖----金庸武侠小说中的人物关系挖掘”，来学习和掌握 MapReduce 课程设计。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。

二、算法原理

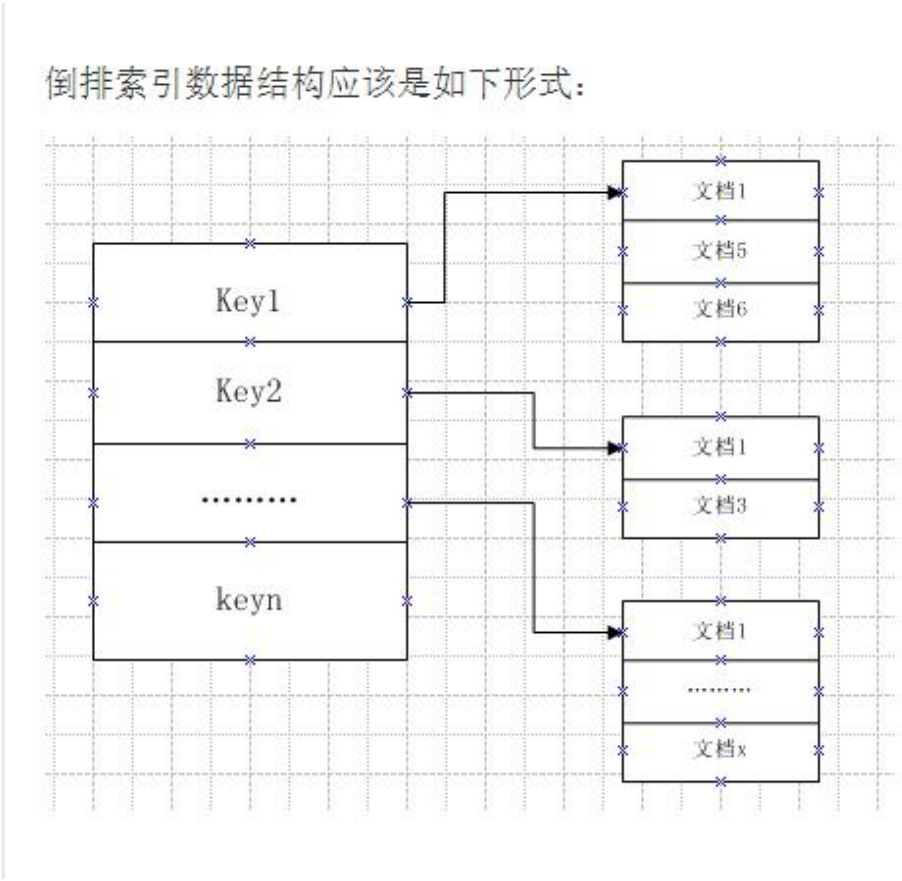
2.1 倒排索引算法

倒排索引源于实际应用中需要根据属性的值来查找记录。这种索引表中的每一项都包括一个属性值和具有该属性值的各记录的地址。由于不是由记录来确定属性值，而是由属性值来确定记录的位置，因而称为倒排索引(inverted index)。带有倒排索引的文件我们称为倒排索引文件，简称倒排文件(inverted file)。

倒排文件（倒排索引），索引对象是文档或者文档集合中的单词等，用来存储这些单词在一个文档或者一组文档中的存储位置，是对文档或者文档集合的一种最常用的索引机制。

搜索引擎的关键步骤就是建立倒排索引，倒排索引一般表示为一个关键词，然后是它的频度（出现的次数），位置（出现在哪一篇文章或

网页中，及有关的日期，作者等信息），它相当于为互联网上几千亿页网页做了一个索引，好比一本书的目录、标签一般。读者想看哪一个主题相关的章节，直接根据目录即可找到相关的页面。不必再从书的第一页到最后一页，一页一页的查找。



如上图，建立各个 key 与所处文档之间的联系，就可以对每一个 key 都能找到它对应的文档，这就是倒排索引的含义。

2.2 单词共现算法

单词共现算法是 MapReduce 可以用来高效解决的一大类问题的抽象化描述。在自然语言处理以及简历语料库上也有着重要的应用。其目的是在海量语料库中发现在固定窗口内单词 a 和单词 b 共同出现的频

率，从而构建单词共现矩阵，这样的矩阵可以是对称的，也可以是不对称的，这要看具体的应用。

这种抽象化的任务的有效解决在实际生活中有着很多的应用。例如电子商家希望发现不同物品被同时购买的情况以便有效安排货物的摆放位置；同时对信息检索领域同义词词典的构建以及文本挖掘等都有着重要的实际应用价值。

设有一个英文句：we are not what we want to be but at least we are not what we used to be.

设共现窗口定义为连续出现的两个单词，则表中给出了上局英文的共现矩阵。

示例英文语句的共现矩阵

	we	are	not	what	we	want	to	be	but	at	least	used
we		2				1						1
are	2		2									
not		2		2								
what			2									
want	1						1					
to						1		1				1
be							1					
but										1		
at									1			
least												
used	1						1					

2.3 PageRank 算法

PageRank，网页排名，又称网页级别、Google 左侧排名或佩奇排名，是一种由根据网页之间相互的超链接计算的技术，而作为网页排名的要素之一，以 Google 公司创办人拉里·佩奇（Larry Page）之姓来命名。Google 用它来体现网页的相关性和重要性，在搜索引擎优化操作中是经常被用来评估网页优化的成效因素之一。Google 的创始人拉里·佩奇和谢尔盖·布林于 1998 年在斯坦福大学发明了这项技术。

PageRank 通过网络浩瀚的超链接关系来确定一个页面的等级。Google 把从 A 页面到 B 页面的链接解释为 A 页面给 B 页面投票，Google 根据投票来源（甚至来源的来源，即链接到 A 页面的页面）和投票目标的等级来决定新的等级。简单的说，一个高等级的页面可以使其他低等级页面的等级提升。

PageRank 让链接来"投票"

一个页面的“得票数”由所有链向它的页面的重要性来决定，到一个页面的超链接相当于对该页投一票。一个页面的 PageRank 是由所有链向它的页面（“链入页面”）的重要性经过递归算法得到的。一个有较多链入的页面会有较高的等级，相反如果一个页面没有任何链入页面，那么它没有等级。

2005 年初，Google 为网页链接推出一项新属性 nofollow，使得网站管理员和网站作者可以做出一些 Google 不计票的链接，也就是说这些链接不算作"投票"。nofollow 的设置可以抵制评论垃圾。

假设一个由 4 个页面组成的小团体：A，B，C 和 D。如果所有页面都链向 A，那么 A 的 PR（PageRank）值将是 B，C 及 D 的 Pagerank 总和。

$$PR(A) = PR(B) + PR(C) + PR(D)$$

继续假设 B 也有链接到 C，并且 D 也有链接到包括 A 的 3 个页面。一个页面不能投票 2 次。所以 B 给每个页面半票。以同样的逻辑，D 投出的票只有三分之一算到了 A 的 PageRank 上。

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

换句话说，根据链出总数平分一个页面的 PR 值。

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

最后，所有这些被换算为一个百分比再乘上一个系数。由于“没有向外链接的页面”传递出去的 PageRank 会是 0，所以，Google 通过数学系统给了每个页面一个最小值：

$$PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\cdots \right) d + \frac{1-d}{N}$$

说明：在 Sergey Brin 和 Lawrence Page 的 1998 年原文中给每一个页面设定的最小值是 1-d，而不是这里的

(1-d)/N。 所以一个页面的 PageRank 是由其他页面的 PageRank 计算得到。Google 不断的重复计算每个页面的 PageRank。如果给每个页面一个随机 PageRank 值（非 0），那么经过不断的重复计算，这些页面的 PR 值会趋向于稳定，也就是收敛的状态。这就是搜索引擎使用它的原因。

2.4 标签传播算法（LPA）

标签传播算法（LPA）是由 Zhu 等人于 2002 年提出，它是一种基于图的半监督学习方法，其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息。利用样本间的关系建立关系完全图模型，在完全图中，节点包括已标注和未标注数据，其边表示两个节点的相似度，节点的标签按相似度传递给其他节点。标签数据就像是一个源头，可以对无标签数据进行标注，节点的相似度越大，标签越容易传播。由于该算法简单易实现，算法执行时间短，复杂度低且分类效果好，引起了国内外学者的关注，并将其广泛地应用到多媒体信息分类、虚拟社区挖掘等领域中。

根据 LPA 算法基本理论，每个节点的标签按相似度传播给相邻节点，在节点传播的每一步，每个节点根据相邻节点的标签来更新自己的标签，与该节点相似度越大，其相邻节点对其标注的影响权值越大，相似节点的标签越趋于一致，其标签就越容易传播。在标签传播过程中，保持已标注数据的标签不变，使其像一个源头把标签传向未标注数据。最终，当迭代过程结束时，相似节点的概率分布也趋于相似，可以划分到同一个类别中，从而完成标签传播过程。

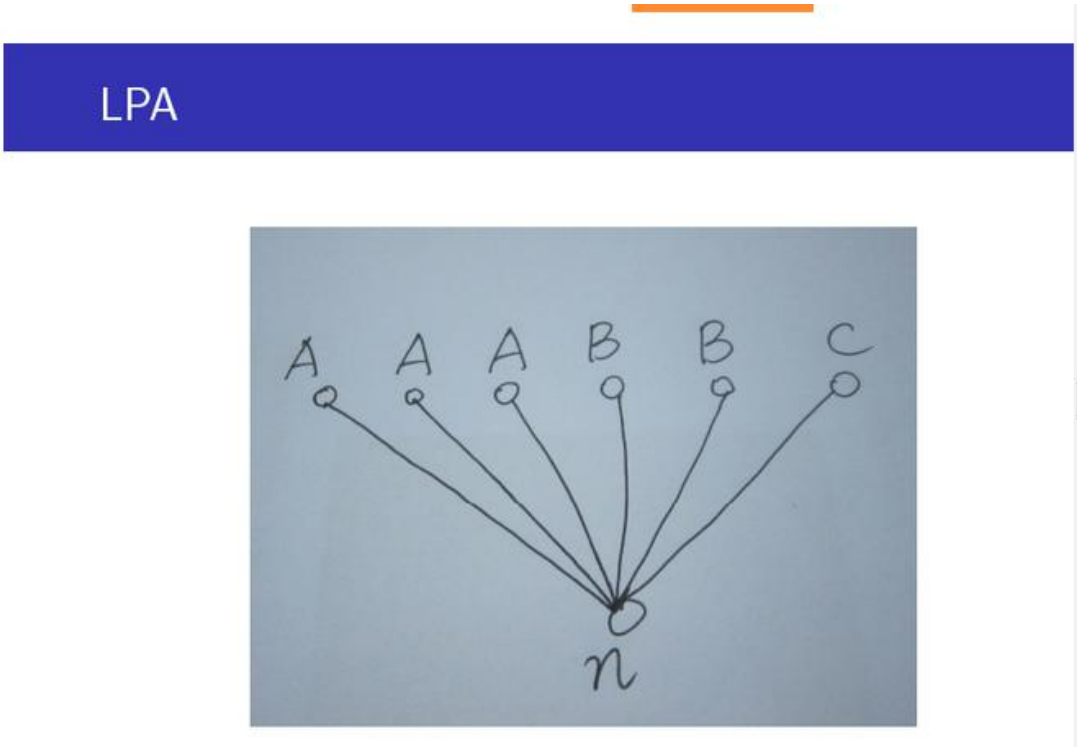
算法过程

第一步: 为所有节点指定一个唯一的标签；

第二步: 逐轮刷新所有节点的标签，直到达到收敛要求为止。对于每一轮刷新，节点标签刷新的规则如下:

对于某一个节点，考察其所有邻居节点的标签，并进行统计，

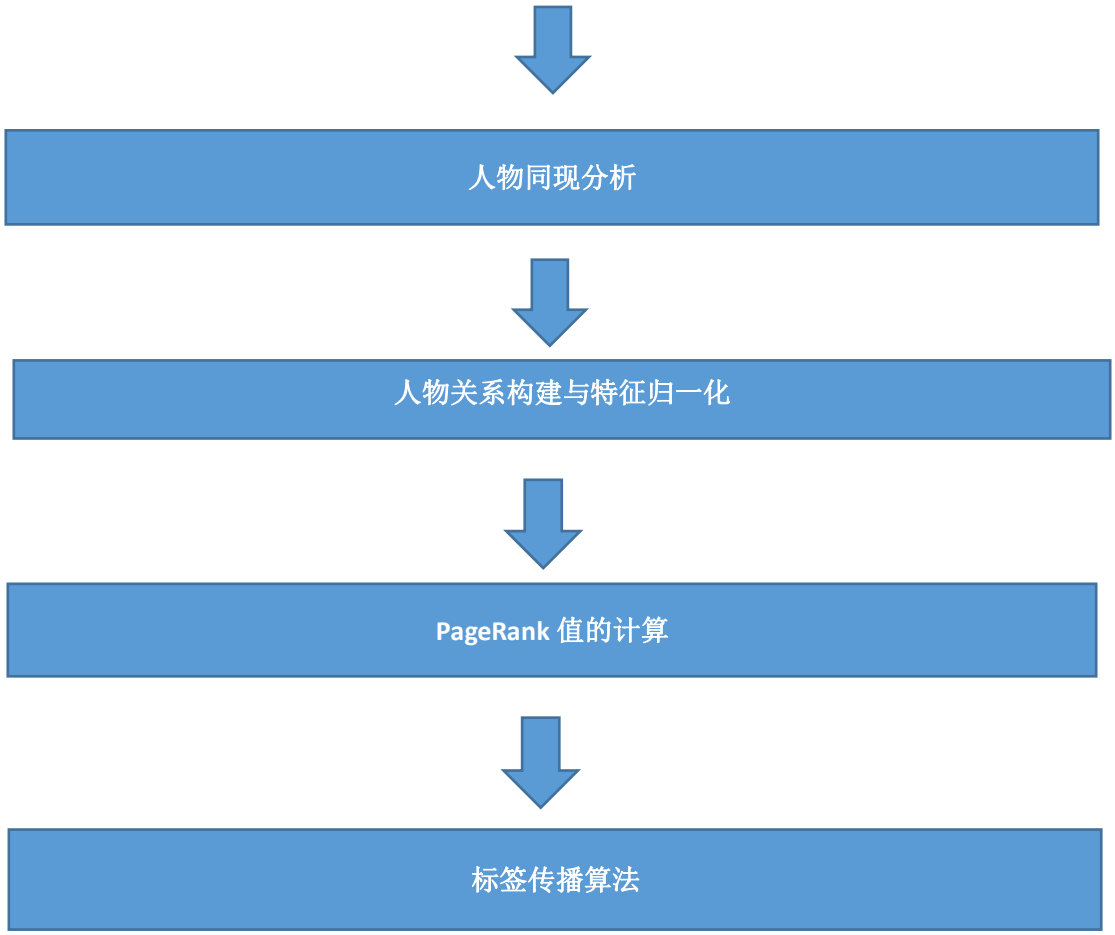
将出现个数最多的那个标签赋给当前节点。当个数最多的标签不唯一时，随机选一个。



如上图（一个图的局部），对于节点 n，与它相邻的一共有六个节点，它们的标签分别为 A、A、A、B、B、C，根据标签传播的理论，此时 n 的标签应该为与它相邻的节点中标签最多的那个标签，如果最多的标签不止一个，随机取一个，所以节点 n 的标签应该为 A。其它的节点也是如此这般，最终整个图的标签趋于稳定。

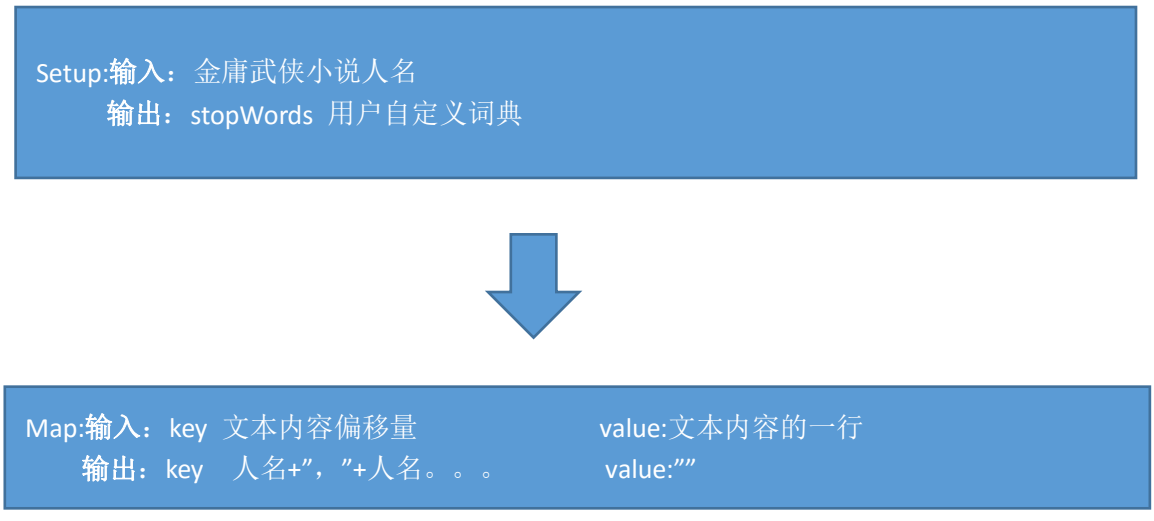
三、实验流程

实验总的流程图



3.1 数据预处理

本任务的主要工作是从金庸小说文本中的每一段，抽取出出现的人名，舍弃其它内容，为后面的人物同现分析做准备。



Reduce:输入：key　人名+”，”+人名。。。　　value:””
　　输出：key　人名+”，”+人名。。。　　value:””

具体思路：

setup 中先把输入文件（金庸武侠小说中的人名停词表）中的所有内容存储到 stopWords 中（作为用户自定义词典）；

map 阶段读取武侠小说中的每一行，用 pdf 提供的 Ansj_seg 分词工具分词，得到分到的词表，然后遍历这个词表，判断词语的词性是否是名字（这个用到分词工具里的词语属性，名字的属性是 nr），对每一个是名字的词语再判断是否存在于之前的停词表中，如果存在，说明这是一个金庸小说中的人物名字，输出，否则继续遍历。

reduce 阶段不做特殊操作，直接输出，作为下一个任务输入。

3.2 特征抽取：人物同现统计

本任务的主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。

Map：输入：key 任务一的输出文本偏移量　value 人名 A，人名 B，。。。。。。。
　　输出：key 人名 A，人名 B　　value　1
　　。。。。。。

Reduce:**输入**：key 人名 A，人名 B　　value 1,1,1。。。。。。
　　输出：key 人名 A，人名 B　　value sum（输入的 1 的总数）

具体思路：

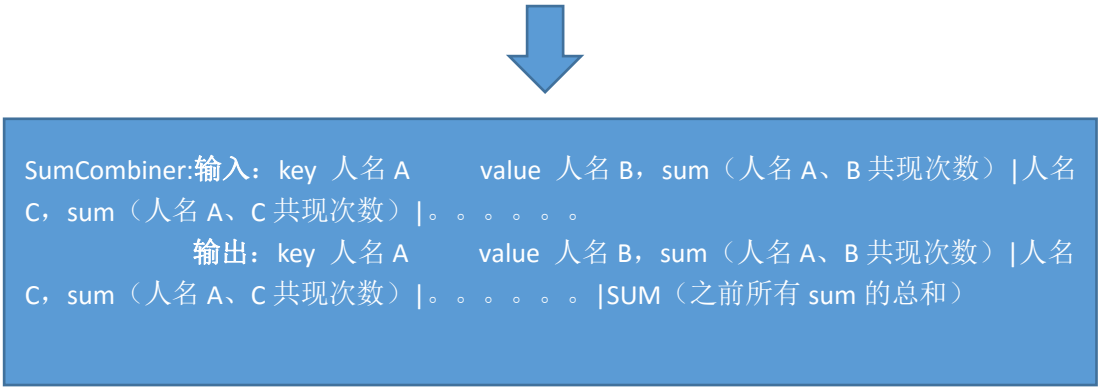
map 阶段直接读取一行中的所有出现的人名，然后看看两两组合的情况。map 阶段输出 key（人名 A,人名 B）,value 1。

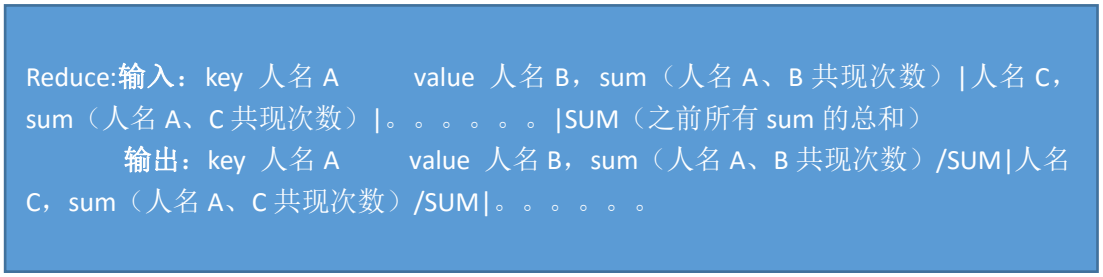
reduce 阶段由于经过 combiner，相同 key 的输入已经全部集中到一起了，所以只要把相同 key 的输入的 value 相加就行了，比如（人名 A,人名 B）1,1,1 直接变成 （人名 A,人名 B）3 输出。

3.3 特征处理：人物关系构建与特征归一化

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。

Map：**输入**：任务二的输出文本偏移量　　value 人名 A，人名 B　sum（人名 A、B 共现次数）
　　输出：key 人名 A　　value 人名 B，sum（人名 A、B 共现次数）





具体思路：

map 阶段对输入进行处理，因为要知道概率，必须知道总数，比如（A,B）3，我们必须知道所有（A，...）...，把所有总数给求出来，然后才能知道概率，所以，我们以 A 作为 key，（B，3）作为 value 输出，sumCombiner 阶段，同样所有以 A 为 key 的输出全部集中到一起了，我们直接计算总数，最后输出 key 为 A，value 为 B，3|...|sum。reduce 阶段，取 value 的最后一项就是总数，然后 value 前面的每一项的个数除以总数就是概率，输出 key 是 A，value 是 B，0.333|...|。

3.4 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，我们就可以定量地金庸武侠江湖中的“主角”们是哪些。

第一阶段：
map: **输入**:key 任务三的输出文本偏移量　　value 人名 A　人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

输出：key 人名 B　　value 人名 A，pagerank*sum（人名 A、B 共现次数）/SUM

。。。。。。

key　人名 A　　value　#人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

Reduce: **输入**：key 人名 B　　value 人名 A，pagerank*sum（人名 A、B 共现次数）/SUM

。。。。。。

key　人名 A　　value　#人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

输出：key　　value　pagerank（新的）　links

第二阶段：
map: **输入**:第一阶段输出文本偏移量　value 人名 A　pagerank　人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

输出：key 人名 B　　value 人名 A，pagerank*sum（人名 A、B 共现次数）/SUM

。。。。。。

key　人名 A　　value　#人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

Reduce: 输入：key 人名 B value 人名 A，pagerank*sum（人名 A、B 共现次数）/SUM
。。。。。。
key 人名 A value #人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。
输出：key value pagerank（新的） links

具体思路：

第一阶段：map 阶段，输入是 value：A B，0.333|...|，处理该 value，定义一个初始 pagerank 1.0，对每一个 value 输出为 key 是 B，value 是 A，0.333*1.0，最后还要输出 key 是 A，value 是 #B，0.333|...|，（这个是为了之后的 pageRank 做准备的）。

reduce 阶段定义一个 links，对每一个 value，先判断是否是以#开头，如果是，links 为 value 从#开始之后的，就是 B,0.333|...|，如果不是#开头的，计算 pagerank 为 所有 key 为 B 的输入的值的总和（key 为 B，value 为 A，0.333*1.0 等等）最后，pagerank = 0.15 + 0.85 * pagerank，输出 key，value 是 pagerank links。

第二阶段，输入为 value 为 A 0.25 B,0.333|...|，注意这里和第一阶段的差别，多了一个 0.25 pagerank 值，这就是经过第一阶段的预处理得到的输入。map 阶段处理 value，同上，只是这里不要定义一个 pagerank 1.0 了，而是调用输入的 0.25，输出 B A，0.25*0.333|...|，reduce 阶段同上，最后输出 key pagerank links。

3.5 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）。

第一阶段：
Map：**输入**：key 任务三的输出文本偏移量 value 人名 A 人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。
输出： key 人名 B value 人名 A(A 的初始标签)
key 人名 C value 人名 A
。。。。。。
key 人名 A value #人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

Reduce：**输入**：key 人名 B value 人名 A，人名 C，。。。。。。
key 人名 A value #人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。
输出：key value label links

第二阶段：
Map：**输入**：key 第一阶段的输出文本偏移量 value 人名 A label_A 人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。
输出： key 人名 B value label_A
key 人名 C value label_A
。。。。。。
key 人名 A value #人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

Reduce：输入：key 人名 B　　value label1，label2，。。。。。。。

　　key 人名 A　　value #人名 B，sum（人名 A、B 共现次数）/SUM|人名 C，sum（人名 A、C 共现次数）/SUM|。。。。。。

　　输出：key　　value　label　links

具体思路：

第一阶段：map 阶段：输入为 value:

A B,0.222|C,0.333|...|，对每一个 value，输出 key 是 B，value 是 A（作为 B 的初始标记候选），等等，最后再像 pagerank 那样输出 key 是 A，value 是#B,0.222|....|。

Reduce 阶段定义 links 和 label，对 map 的每一个输入，先判断是否以#开头，如果是，就把 links 赋值为 B,0.222|...|，如果不是，（就是形如 B　A，C，D...）比如是 A，判断一下以 B 为 key，A 有没有出现过，如果没有，就记 A,1，否则就把 A 的值加一（用到了 map.put(),map.get()）,这样遍历完一个 key 之后，只要找到其中的值最大的就行了，最大的一样大随机取一个，最后输出 key，value：label links。

第二阶段：同上，不过这个时候输入就比第一阶段多一个 label 了，比如为 value：A C B,0.222|C,0.333|...|。map 阶段：输入为 value: A C B,0.222|C,0.333|...|，对每一个 value，输出 key 是 B，value 是 C（这

里就不是初始标记，而是输入中的 label，即所有的 value 都输出这个 C label）等等，最后再像 pagerank 那样输出 key 是 A，value 是 #B,0.222|....|。，0.333|...|。

reduce 阶段同上，最后输出 key label links。

3.6 分析结果整理（选做）

这个结果整理对应的是 rankSort.java 和 LPASort.java，由于 MapReduce 的机制问题（根据主键值自动排序），输出结果自动排序，所以分析结果整理就自动完成，很简单，不过默认的是从小到大排序，这里稍微修改一下，变成从大到小排序就行了。

（具体实现：重载了 map 的 key 输出排序函数 compare 函数：

```
public static class DoubleWritableDecreasingComparator extends Comparator {
        public int compare(DoubleWritable a,DoubleWritable b){
            return -super.compare(a, b);

        }
        @Override
        public int compare(byte[] b1, int s1, int l1, byte[] b2, int s2, int l2) {
            // TODO Auto-generated method stub
            return -super.compare(b1, s1, l1, b2, s2, l2);
        }
    }
```

最后结果输出为负的）

四、实验优化

1、人名分词、人物共现分析使用了 combiner，这样就导致它们的 reduce 可以几乎不用做什么别的事了，简化了代码。

2、我们的 pagerank 和标签传播部分分了两个部分，有 pre_pagerank 和 preLPA，作用是为了方便迭代，在原来的 key 和 value 基础上加上了 pagerank 初始值以及 label 初始值。

3、标签传播过程中，当两个名字的标签一样时，一开始是随机选一个，后来经过优化，改成了选两个里面 pagerank 值更大的一个，这样就会导致最后的标签更容易使主角了，使得最后结果更加精确。

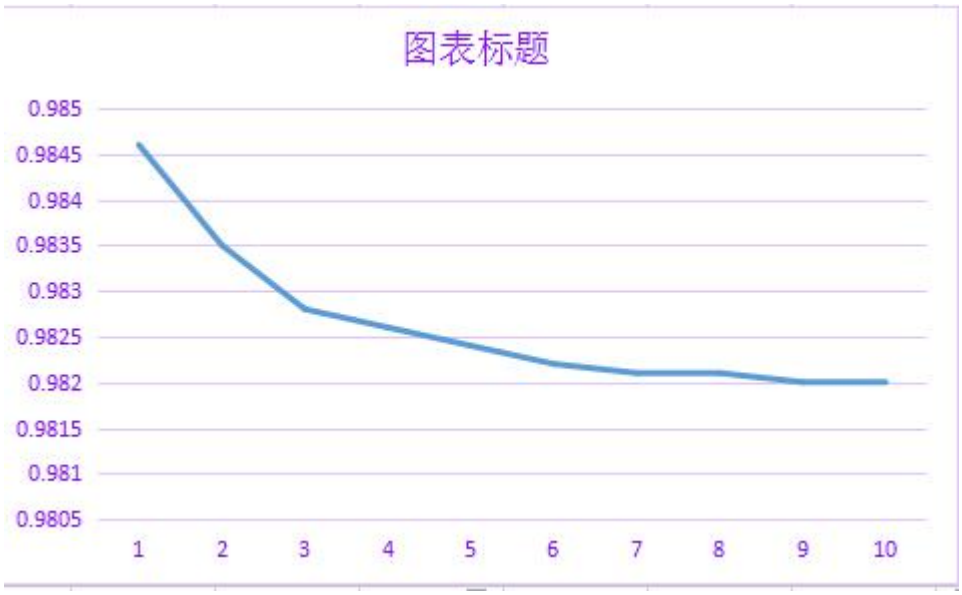
4、关于确定迭代是否终止，另外写了一个小程序，判断每一次迭代得到的所有人的 pagerank 值的平均值之间的差，最后发现当迭代到一定程度之后，差值稳定在一个值上下，就判断迭代终止，这样大大减少了运算量。

下图就是每一个名字的 pagerank 在每一次迭代之后的值的变化情况：

陆乘风	0.460801	0.544788	0.572671	0.590275	0.601879	0.609605	0.615104	0.619125	0.622155	0.624482
黎生	0.455136	0.377157	0.379205	0.386725	0.392326	0.396325	0.399156	0.401226	0.402777	0.403968
曹云奇	1.591516	1.457972	1.468517	1.380469	1.336987	1.297019	1.269949	1.24948	1.234699	1.223814
姚伯当	0.422903	0.558632	0.487516	0.476463	0.454139	0.448278	0.442567	0.440576	0.439053	0.438378
殷无福	0.651035	0.492564	0.45312	0.437016	0.432054	0.430387	0.429965	0.429954	0.430076	0.430226
大汉	7.147112	6.086636	5.823401	5.702402	5.633383	5.591978	5.565217	5.54721	5.534666	5.525688
说不得	2.162955	2.617351	2.672084	2.719961	2.738751	2.750073	2.755882	2.759532	2.761806	2.763366
喀丝丽	0.551101	0.545103	0.52984	0.52975	0.532114	0.5355	0.538459	0.540896	0.542808	0.544293
黄面道人	0.337422	0.278226	0.240479	0.228626	0.223355	0.221223	0.220222	0.219728	0.219451	0.219279
刘鹤真	0.639865	0.920963	0.827036	0.817609	0.797265	0.786227	0.777688	0.771715	0.767303	0.764054
刚相	0.152099	0.156535	0.154519	0.154789	0.154785	0.154799	0.154803	0.154808	0.154812	0.154816
吴道通	1.035294	0.614002	0.586912	0.546882	0.537195	0.530056	0.52663	0.524301	0.522754	0.52162
马夫人	0.910447	1.403001	1.377982	1.368684	1.35451	1.343744	1.336887	1.332095	1.328938	1.326741
若克琳	0.308675	0.332998	0.30452	0.297574	0.293489	0.291532	0.290374	0.289671	0.289191	0.28885
彭三春	0.50211	0.439221	0.439518	0.442176	0.445411	0.448227	0.450576	0.452468	0.45396	0.455121
本观	0.647026	0.699764	0.601771	0.580627	0.564541	0.5578	0.554151	0.552345	0.551382	0.550851
胡青牛	1.496846	1.516545	1.550412	1.580461	1.594145	1.60288	1.60811	1.611587	1.613978	1.615692
风清扬	0.650494	0.810729	0.79904	0.805901	0.808045	0.810266	0.811851	0.813056	0.813948	0.814613
宝树	1.063138	1.029664	0.978251	0.944353	0.918876	0.900424	0.88684	0.876866	0.869518	0.864101
段正明	0.331099	0.43679	0.416413	0.427737	0.428714	0.430869	0.431476	0.431858	0.431931	0.431906
小龙女	3.288396	4.803066	4.89591	5.049012	5.117352	5.169815	5.207351	5.236837	5.260427	5.279713
阎世章	0.778282	0.673091	0.657335	0.654185	0.656733	0.66038	0.663776	0.66657	0.668771	0.670473
德鄂	0.160222	0.16974	0.163785	0.163588	0.163474	0.16354	0.163627	0.163707	0.163773	0.163824
孙克通	0.17856	0.19789	0.200294	0.202119	0.203186	0.20382	0.204262	0.204584	0.204827	0.205012
陆冠英	0.966279	0.982162	1.031186	1.065936	1.089913	1.106429	1.118219	1.126827	1.133279	1.138215
汉子	16.15226	14.65598	13.86023	13.68827	13.56692	13.51452	13.47873	13.45641	13.44031	13.42868
多尔衮	0.800528	0.801081	0.755809	0.737284	0.726062	0.719924	0.716032	0.713404	0.711498	0.710056

下图是每次迭代，所有人的 pagerank 值的平均值的变化情况，可以

看出，在迭代 9 次之后，平均值趋于稳定于 0.9820，迭代终止。



五、程序运行的截图

系集群上的运行截图

application_1467969296998_6642	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:39:30 +0800 2016	Tue Jul 19 13:40:02 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6641	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:38:52 +0800 2016	Tue Jul 19 13:39:27 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6639	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:38:16 +0800 2016	Tue Jul 19 13:38:45 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6638	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:37:50 +0800 2016	Tue Jul 19 13:38:13 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6637	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:37:19 +0800 2016	Tue Jul 19 13:37:47 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6636	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:36:48 +0800 2016	Tue Jul 19 13:37:12 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6635	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19 13:36:19 +0800 2016	Tue Jul 19 13:36:44 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6634	2016st18	pre_pagerank	MAPREDUCE	root.default	Tue Jul 19 13:35:42 +0800 2016	Tue Jul 19 13:36:11 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6633	2016st18	job3	MAPREDUCE	root.default	Tue Jul 19 13:35:00 +0800 2016	Tue Jul 19 13:35:35 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6632	2016st18	job2	MAPREDUCE	root.default	Tue Jul 19 13:34:28 +0800 2016	Tue Jul 19 13:34:57 +0800 2016	FINISHED	SUCCEEDED		History
application_1467969296998_6631	2016st18	job1	MAPREDUCE	root.default	Tue Jul 19 13:33:36	Tue Jul 19 13:34:24	FINISHED	SUCCEEDED		History

application_1467969296998_6659	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:49:12 +0800 2016	Tue Jul 19 13:49:48 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6658	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:48:35 +0800 2016	Tue Jul 19 13:49:05 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6657	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:47:55 +0800 2016	Tue Jul 19 13:48:27 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6656	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:47:09 +0800 2016	Tue Jul 19 13:47:47 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6655	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:46:20 +0800 2016	Tue Jul 19 13:47:02 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6654	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:45:36 +0800 2016	Tue Jul 19 13:46:12 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6653	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:44:58 +0800 2016	Tue Jul 19 13:45:33 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6652	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:44:17 +0800 2016	Tue Jul 19 13:44:51 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6651	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:43:30 +0800 2016	Tue Jul 19 13:44:10 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6650	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:43:02 +0800 2016	Tue Jul 19 13:43:26 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6649	2016st18	pre_LPA	MAPREDUCE	root.default	Tue Jul 19 13:42:32 +0800 2016	Tue Jul 19 13:42:55 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6648	2016st18	rankSort	MAPREDUCE	root.default	Tue Jul 19 13:41:52 +0800 2016	Tue Jul 19 13:42:25 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6647	2016st18	pageRank	MAPREDUCE	root.default	Tue Jul 19	Tue Jul 19	FINISHED	SUCCEEDED	History
application_1467969296998_6660	2016st18	LPA_sort	MAPREDUCE	root.default	Tue Jul 19 13:49:51 +0800 2016	Tue Jul 19 13:50:18 +0800 2016	FINISHED	SUCCEEDED	History
application_1467969296998_6659	2016st18	LPA	MAPREDUCE	root.default	Tue Jul 19 13:49:12	Tue Jul 19 13:49:48	FINISHED	SUCCEEDED	History

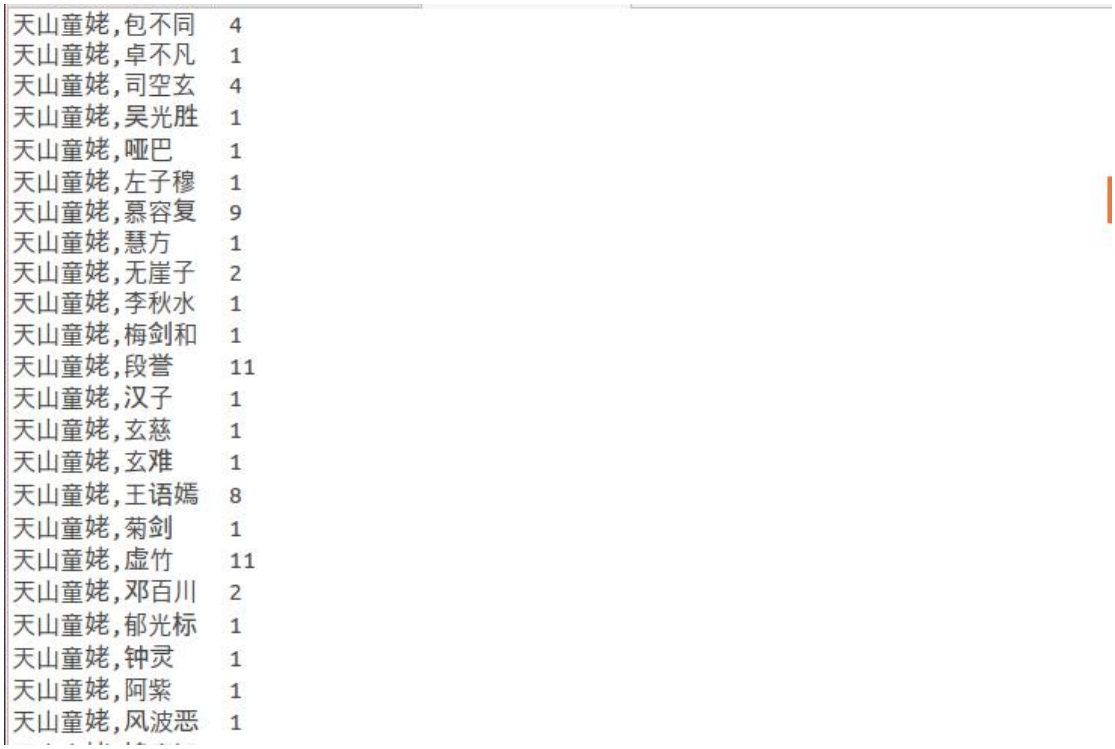
人名分词的截图（总共运行了 48s）

Counter Group	Name	Map	Reduce	Total
		Counters		
File System Counters	FILE: Number of bytes read	0	1,729,662	1,729,662
	FILE: Number of bytes written	3,477,746	1,846,140	5,323,886
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	25,730,180	0	25,730,180
	HDFS: Number of bytes written	0	1,332,611	1,332,611
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	45	3	48
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Data-local map tasks	0	0	12
	Launched map tasks	0	0	15
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	3
	Total megabyte-seconds taken by all map tasks	0	0	191,968,256
	Total megabyte-seconds taken by all reduce tasks	0	0	9,307,136
	Total time spent by all map tasks (ms)	0	0	187,469
	Total time spent by all maps in occupied slots (ms)	0	0	187,469
	Total time spent by all reduce tasks (ms)	0	0	9,089
	Total time spent by all reduces in occupied slots (ms)	0	0	9,089
	Total vcore-seconds taken by all map tasks	0	0	187,469
	Total vcore-seconds taken by all reduce tasks	0	0	9,089
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	579,280	3,280	582,560
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	69,753	41	69,794
	Input split bytes	1,914	0	1,914
	Map input records	47,486	0	47,486
	Map output bytes	1,632,108	0	1,632,108
	Map output materialized bytes	1,729,746	0	1,729,746
	Map output records	47,486	0	47,486
	Merged Map outputs	0	15	15
	Physical memory (bytes) snapshot	4,582,080,512	176,459,776	4,758,540,288
	Reduce input groups	0	20,996	20,996
	Reduce input records	0	47,486	47,486
	Reduce output records	0	20,996	20,996
	Reduce shuffle bytes	0	1,729,746	1,729,746
	Shuffled Maps	0	15	15
	Spilled Records	47,486	47,486	94,972
	Total committed heap usage (bytes)	2,980,052,992	201,326,592	3,181,379,584
	Virtual memory (bytes) snapshot	24,740,155,392	1,657,151,488	26,397,306,880
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	25,728,266	0	25,728,266
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	1,332,611	1,332,611

一灯大师,武三通,农夫,李莫愁,渔人,朱子柳,
一灯大师,裘千尺,黄蓉,武三通,
一灯大师,郭靖,一灯大师,一灯大师,一灯大师,一灯大师,黄蓉,郭靖,黄蓉,郭靖,渔人,郭靖,樵子,农夫,郭靖,一灯大师,黄蓉,郭靖,郭靖,黄蓉,黄蓉,郭靖,黄蓉,欧阳克,裘千仞,
一灯大师,郭靖,一灯大师,黄蓉,郭靖,一灯大师,黄蓉,
一灯大师,郭靖,周伯通,黄蓉,黄药师,郭靖,
一灯大师,郭靖,郭靖,琴儿,郭靖,郭靖,郭靖,郭靖,郭靖,鲁有脚,郭靖,
一灯大师,郭靖,郭靖,郭靖,渔人,樵子,一灯大师,
一灯大师,郭靖,黄蓉,
一灯大师,黄蓉,
一灯大师,黄蓉,一灯大师,裘千尺,一灯大师,杨过,黄蓉,
一灯大师,黄蓉,农夫,黄蓉,黄蓉,黄蓉,农夫,黄蓉,黄蓉,黄蓉,黄蓉,黄蓉,
一灯大师,黄蓉,杨过,小龙女,杨过,黄药师,
一灯大师,黄蓉,武三通,耶律齐,裘千尺,黄蓉,黄蓉,裘千尺,
一灯大师,黄蓉,郭靖,郭靖,黄蓉,郭靖,黄蓉,郭靖,郭靖,郭靖,黄蓉,哑巴,郭靖,黄蓉,郭靖,汉子,哑巴,大汉,乔寨主,乔寨主,乔寨主,乔寨主,乔寨主,乔寨主,乔寨主,黄蓉,黄蓉,一灯大师,
一灯大师,黄蓉,郭靖,黄蓉,郭靖,黄蓉,一灯大师,一灯大师,一灯大师,天竺僧人,郭靖,黄蓉,黄蓉,黄蓉,农夫,农夫,一灯大师,一灯大师,农夫,一灯大师,一灯大师,黄蓉,郭靖,郭靖,
一灯大师,黄蓉,黄蓉,
一灯大师,黄蓉,黄蓉,农夫,一灯大师,
一灯大师,黄蓉,黄蓉,渔人,农夫,黄蓉,黄蓉,黄蓉,黄蓉,

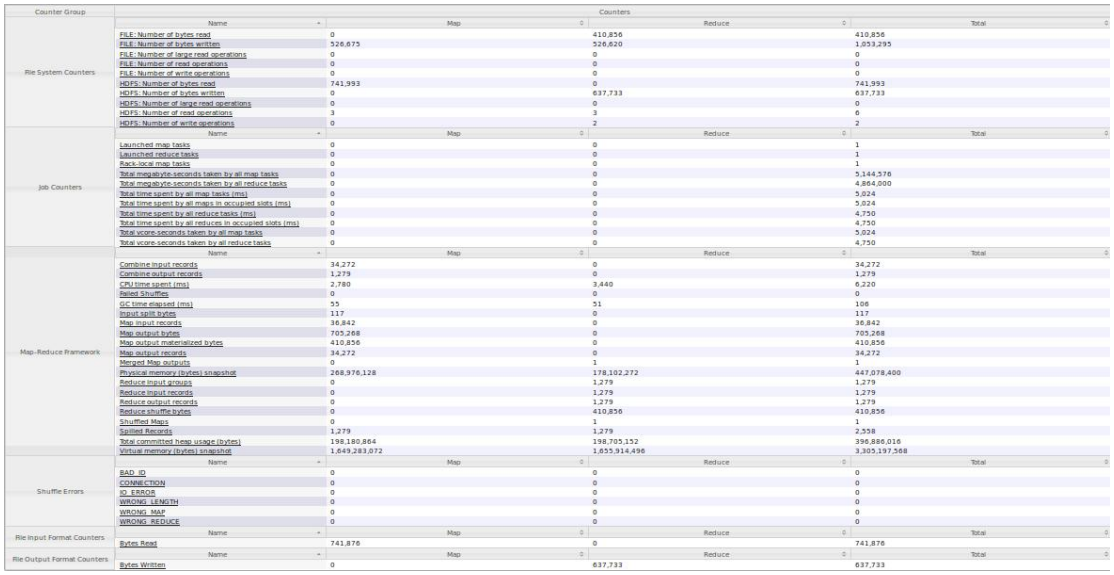
人物同现统计的截图（总共运行了 29s）

Counter Group	Counters			
File System Counters	Name	Map	Reduce	Total
	FILE: Number of bytes read	0	882,330	882,330
	FILE: Number of bytes written	998,314	998,259	1,996,573
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	1,332,728	0	1,332,728
	HDFS: Number of bytes written	0	741,876	741,876
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	5,131,264
	Total megabyte-seconds taken by all reduce tasks	0	0	10,422,272
	Total time spent by all map tasks (ms)	0	0	5,011
	Total time spent by all maps in occupied slots (ms)	0	0	5,011
	Total time spent by all reduce tasks (ms)	0	0	10,178
	Total time spent by all reduces in occupied slots (ms)	0	0	10,178
	Total vcore-seconds taken by all map tasks	0	0	5,011
	Total vcore-seconds taken by all reduce tasks	0	0	10,178
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	337,796	0	337,796
	Combine output records	36,842	0	36,842
	CPU time spent (ms)	3,560	3,190	6,750
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	57	63	120
	Input split bytes	117	0	117
	Map input records	20,996	0	20,996
	Map output bytes	6,598,052	0	6,598,052
	Map output materialized bytes	882,330	0	882,330
	Map output records	337,796	0	337,796
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	269,721,600	177,991,680	447,713,280
	Reduce input groups	0	36,842	36,842
	Reduce input records	0	36,842	36,842
	Reduce output records	0	36,842	36,842
	Reduce shuffle bytes	0	882,330	882,330
	Shuffled Maps	0	1	1
	Spilled Records	36,842	36,842	73,684
	Total committed heap usage (bytes)	199,753,728	201,326,592	401,080,320
	Virtual memory (bytes) snapshot	1,649,283,072	1,655,930,880	3,305,213,952
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	1,332,611	0	1,332,611
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	741,876	741,876



天山童姥,包不同 4
天山童姥,卓不凡 1
天山童姥,司空玄 4
天山童姥,吴光胜 1
天山童姥,哑巴 1
天山童姥,左子穆 1
天山童姥,慕容复 9
天山童姥,慧方 1
天山童姥,无崖子 2
天山童姥,李秋水 1
天山童姥,梅剑和 1
天山童姥,段誉 11
天山童姥,汉子 1
天山童姥,玄慈 1
天山童姥,玄难 1
天山童姥,王语嫣 8
天山童姥,菊剑 1
天山童姥,虚竹 11
天山童姥,邓百川 2
天山童姥,郁光标 1
天山童姥,钟灵 1
天山童姥,阿紫 1
天山童姥,风波恶 1

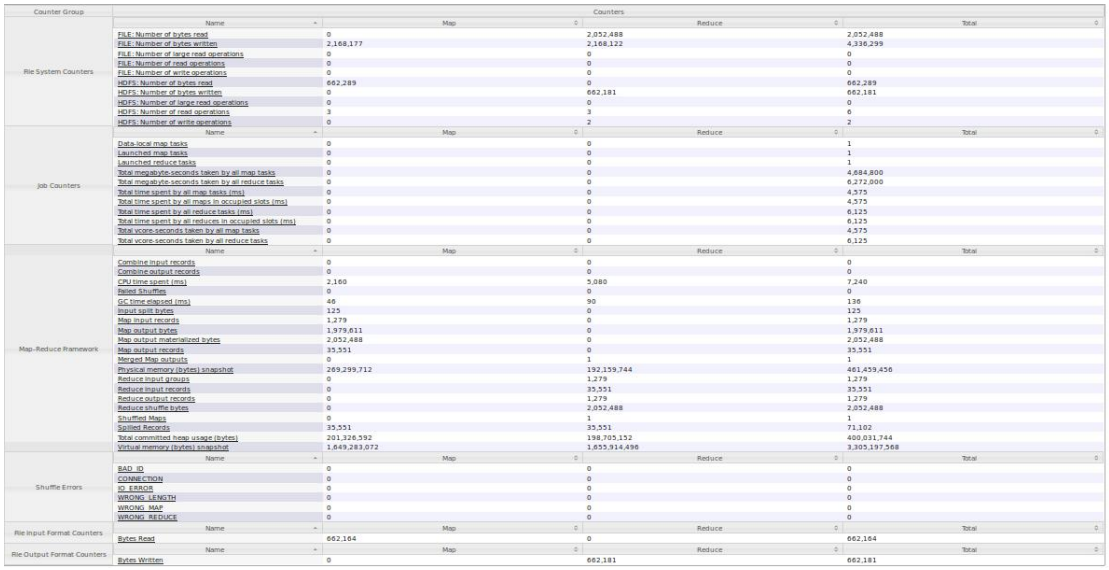
人物关系图构建与特征归一化的截图（总共运行了 35s）



Counter Group	Counters			
File System Counters	Name	Map	Reduce	Total
	FILE: Number of bytes read	0	410,856	410,856
	FILE: Number of bytes written	526,675	526,620	1,053,295
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	741,993	0	741,993
	HDFS: Number of bytes written	0	637,733	637,733
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	5,144,576
	Total megabyte-seconds taken by all reduce tasks	0	0	4,864,000
	Total time spent by all map tasks (ms)	0	0	5,024
	Total time spent by all maps in occupied slots (ms)	0	0	5,024
	Total time spent by all reduce tasks (ms)	0	0	4,750
	Total time spent by all reduces in occupied slots (ms)	0	0	4,750
	Total vcore-seconds taken by all map tasks	0	0	5,024
	Total vcore-seconds taken by all reduce tasks	0	0	4,750
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	34,272	0	34,272
	Combine output records	1,279	0	1,279
	CPU time spent (ms)	2,780	3,440	6,220
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	55	51	106
	Input split bytes	117	0	117
	Map input records	36,842	0	36,842
	Map output bytes	705,268	0	705,268
	Map output materialized bytes	410,856	0	410,856
	Map output records	34,272	0	34,272
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	268,976,128	178,102,272	447,078,400
	Reduce input groups	0	1,279	1,279
	Reduce input records	0	1,279	1,279
	Reduce output records	0	1,279	1,279
	Reduce shuffle bytes	0	410,856	410,856
	Shuffled Maps	0	1	1
	Spilled Records	1,279	1,279	2,558
	Total committed heap usage (bytes)	198,180,864	198,705,152	396,886,016
	Virtual memory (bytes) snapshot	1,649,283,072	1,655,914,496	3,305,197,568
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	741,876	0	741,876
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	637,733	637,733

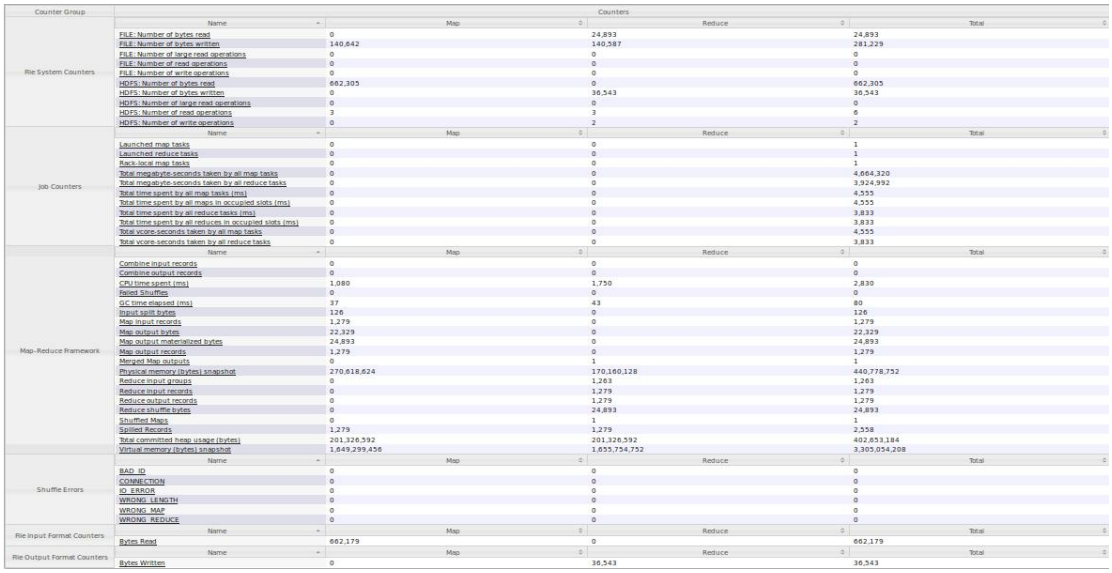
张君宝　卫天望,0.007576|周伯通,0.015152|小龙女,0.007576|尹克西,0.196970|方天劳,0.007576|无相禅师,0.015152|无色,0.075758|杨过,0.151515|武修文,0.007576|汉子,0.007576|潇湘子,0.037879|潘天耕,0.015152|觉远,0.204545|觉远大师,0.007576|郭芙,0.007576|何足道,0.068182|郭襄,0.143939|郭靖,0.015152|农妇,0.007576|
张夫人　空闻,0.009804|上官云,0.009804|不戒和尚,0.009804|东方不败,0.009804|严三星,0.049020|于嫂,0.009804|仇松年,0.127451|令狐冲,0.078431|何太冲,0.009804|何足道,0.009804|余沧海,0.039216|冲虚道长,0.009804|史火龙,0.009804|司马大,0.019608|哑婆婆,0.009804|姚清泉,0.009804|岳不群,0.039216|岳灵珊,0.019608|张三丰,0.009804|张无忌,0.009804|张翠山,0.029412|朱九真,0.009804|朱长龄,0.009804|林平之,0.019608|桃叶仙,0.009804|桃干仙,0.009804|桃枝仙,0.009804|桃根仙,0.009804|殷素素,0.029412|汉子,0.019608|游迅,0.098039|玉灵道人,0.088235|田伯光,0.009804|祖千秋,0.009804|空智,0.009804|老头子,0.009804|聋哑婆婆,0.009804|蒋涛,0.009804|西华子,0.009804|西宝,0.009804|计无施,0.009804|谢逊,0.029412|贺老三,0.009804|静玄师太,0.009804|高则成,0.009804|黄伯流,0.009804|
张妈　吴三桂,0.062500|双儿,0.062500|大汉,0.062500|守备,0.062500|少妇,0.062500|店伴,0.062500|徐天川,0.125000|杨夫人,0.062500|汉子,0.062500|汤和,0.062500|老不死,0.062500|韦小宝,0.250000|
张安官　周仲英,0.200000|周绮,0.100000|常伯志,0.100000|常赫志,0.100000|张召重,0.100000|徐天宏,0.200000|蒋四根,0.100000|陆菲青,0.100000|

基于人物关系图的 pagerank 计算的截图（总共运行了 29+366 = 395s）

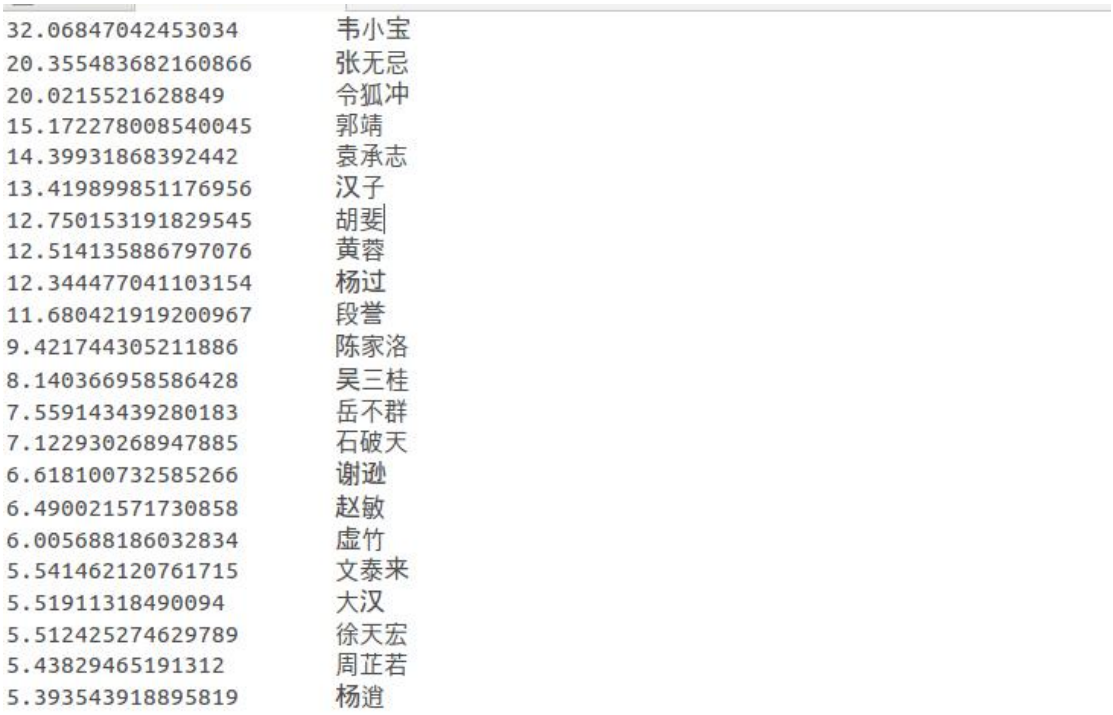


Counter Group	Counters			
File System Counters	Name	Map	Reduce	Total
	FILE: Number of bytes read	0	2,052,488	2,052,488
	FILE: Number of bytes written	2,168,177	2,168,122	4,336,299
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	662,289	0	662,289
	HDFS: Number of bytes written	0	662,181	662,181
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Data-local map tasks	0	0	1
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	4,684,800
	Total megabyte-seconds taken by all reduce tasks	0	0	6,272,000
	Total time spent by all map tasks (ms)	0	0	4,575
	Total time spent by all maps in occupied slots (ms)	0	0	4,575
	Total time spent by all reduce tasks (ms)	0	0	6,125
	Total time spent by all reduces in occupied slots (ms)	0	0	6,125
	Total vcore-seconds taken by all map tasks	0	0	4,575
	Total vcore-seconds taken by all reduce tasks	0	0	6,125
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	2,160	5,080	7,240
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	46	90	136
	Input split bytes	125	0	125
	Map input records	1,279	0	1,279
	Map output bytes	1,979,611	0	1,979,611
	Map output materialized bytes	2,052,488	0	2,052,488
	Map output records	35,551	0	35,551
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	269,299,712	192,159,744	461,459,456
	Reduce input groups	0	1,279	1,279
	Reduce input records	0	35,551	35,551
	Reduce output records	0	1,279	1,279
	Reduce shuffle bytes	0	2,052,488	2,052,488
	Shuffled Maps	0	1	1
	Spilled Records	35,551	35,551	71,102
	Total committed heap usage (bytes)	203,126,592	198,705,152	400,031,744
	Virtual memory (bytes) snapshot	1,649,283,072	1,655,914,496	3,305,197,568
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	662,164	0	662,164
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	662,181	662,181

pagerank 的排序截图



Counter Group	Counters			
	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	24,893	24,893
	FILE: Number of bytes written	140,642	140,587	281,229
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	662,305	0	662,305
	HDFS: Number of bytes written	0	36,543	36,543
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	4,664,320
	Total megabyte-seconds taken by all reduce tasks	0	0	3,924,992
	Total time spent by all map tasks (ms)	0	0	4,555
	Total time spent by all maps in occupied slots (ms)	0	0	4,555
	Total time spent by all reduce tasks (ms)	0	0	3,833
	Total time spent by all reduces in occupied slots (ms)	0	0	3,833
	Total vcore-seconds taken by all map tasks	0	0	4,555
	Total vcore-seconds taken by all reduce tasks	0	0	3,833
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	1,080	1,750	2,830
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	37	43	80
	Input split bytes	126	0	126
	Map input records	1,279	0	1,279
	Map output bytes	22,329	0	22,329
	Map output materialized bytes	24,893	0	24,893
	Map output records	1,279	0	1,279
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	270,618,624	170,160,128	440,778,752
	Reduce input groups	0	1,263	1,263
	Reduce input records	0	1,279	1,279
	Reduce output records	0	1,279	1,279
	Reduce shuffle bytes	0	24,893	24,893
	Shuffled Maps	0	1	1
	Spilled Records	1,279	1,279	2,558
	Total committed heap usage (bytes)	201,326,592	201,326,592	402,653,184
	Virtual memory (bytes) snapshot	1,649,299,456	1,655,754,752	3,305,054,208
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	662,179	0	662,179
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	36,543	36,543



32.06847042453034	韦小宝
20.355483682160866	张无忌
20.0215521628849	令狐冲
15.172278008540045	郭靖
14.39931868392442	袁承志
13.419899851176956	汉子
12.750153191829545	胡斐
12.514135886797076	黄蓉
12.344477041103154	杨过
11.680421919200967	段誉
9.421744305211886	陈家洛
8.140366958586428	吴三桂
7.559143439280183	岳不群
7.122930268947885	石破天
6.618100732585266	谢逊
6.490021571730858	赵敏
6.005688186032834	虚竹
5.541462120761715	文泰来
5.51911318490094	大汉
5.512425274629789	徐天宏
5.43829465191312	周芷若
5.393543918895819	杨逍

在人物关系图上的标签传播的截图（总共运行了 23+376 = 399s）

Counter Group	Counters			
	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	1,338,853	1,338,853
	FILE: Number of bytes written	1,455,397	1,455,342	2,910,739
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	649,293	0	649,293
	HDFS: Number of bytes written	0	649,143	649,143
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
	Name	Map	Reduce	Total
Job Counters	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	4,736,000
	Total megabyte-seconds taken by all reduce tasks	0	0	5,072,896
	Total time spent by all map tasks (ms)	0	0	4,625
	Total time spent by all maps in occupied slots (ms)	0	0	4,625
	Total time spent by all reduce tasks (ms)	0	0	4,954
	Total time spent by all reduces in occupied slots (ms)	0	0	4,954
	Total vcore-seconds taken by all map tasks	0	0	4,625
	Total vcore-seconds taken by all reduce tasks	0	0	4,954
	Name	Map	Reduce	Total
Map-Reduce Framework	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	1,720	3,650	5,370
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	51	39	90
	Input split bytes	132	0	132
	Map input records	1,279	0	1,279
	Map output bytes	1,265,976	0	1,265,976
	Map output materialized bytes	1,338,853	0	1,338,853
	Map output records	35,551	0	35,551
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	268,759,040	188,772,352	457,531,392
	Reduce input groups	0	1,279	1,279
	Reduce input records	0	35,551	35,551
	Reduce output records	0	1,279	1,279
	Reduce shuffle bytes	0	1,338,853	1,338,853
	Shuffled Maps	0	1	1
	Spilled Records	35,551	35,551	71,102
	Total committed heap usage (bytes)	201,326,592	201,326,592	402,653,184
	Virtual memory (bytes) snapshot	1,649,283,072	1,655,922,688	3,305,205,760
	Name	Map	Reduce	Total
Shuffle Errors	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
	Name	Map	Reduce	Total
File Input Format Counters	Bytes Read	649,161	0	649,161
	Name	Map	Reduce	Total
File Output Format Counters	Bytes Written	0	649,143	649,143

丁典 梅念笙
丁典 万圭
丁典 马大鸣
丁典 言达平
丁典 空心菜
丁典 鱼贩头子
丁典 陆天抒
丁典 宝象和尚
丁典 血刀老祖
丁典 吴坎
丁典 花铁干
丁典 丁典
丁典 鲁坤
丁典 卜垣
丁典 沈城
丁典 狄云
丁典 水笙
丁典 刘乘风
丁典 孙均
丁典 万震山
丁典 汪啸风
丁典 菊友
丁典 戚长发

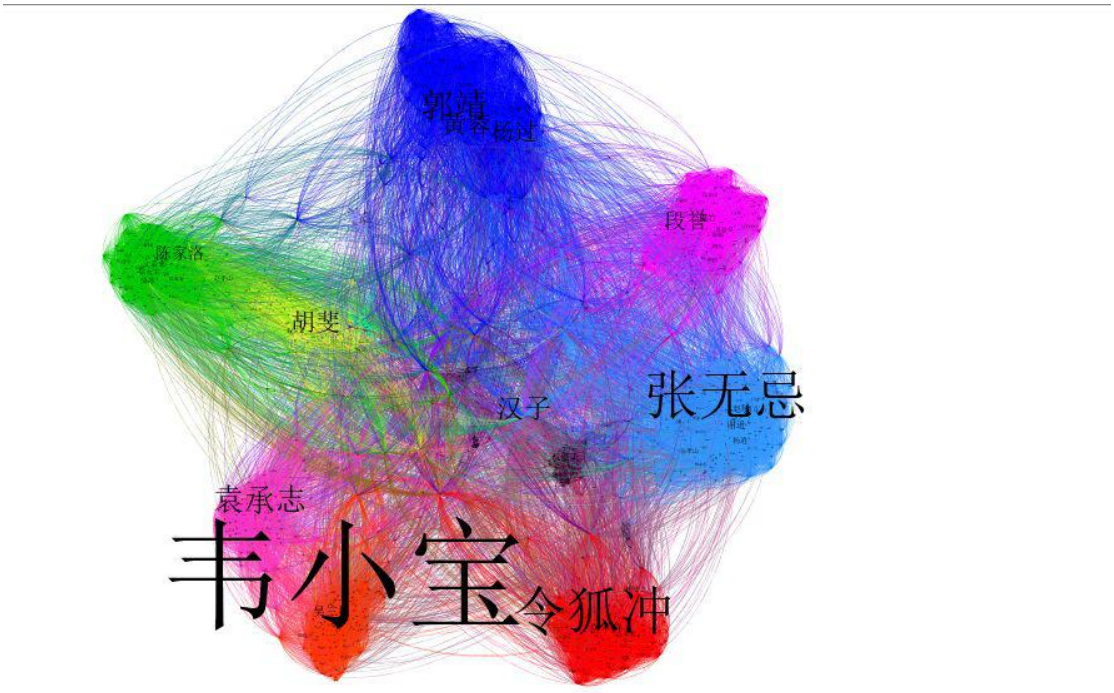
Gephi 中的点文件截图

1	id	label	class	pr
2	水岱	水岱	狄云	0.7977620831544401
3	水笙	水笙	狄云	2.0360371193686824
4	鲁坤	鲁坤	狄云	0.741468050716196
5	鱼贩头子	鱼贩头子	狄云	0.16479451464484302
6	血刀老祖	血刀老祖	狄云	0.8412192213885631
7	空心菜	空心菜	狄云	0.6404590482072932
8	凌退思	凌退思	狄云	0.5489774236117426
9	陆天抒	陆天抒	狄云	0.4843065652077062
10	水福	水福	狄云	0.1647945146484302
11	沈城	沈城	狄云	0.7477188526564287
12	菊友	菊友	狄云	0.179185414644786 24
13	冯坦	冯坦	狄云	0.5214939052898743
14	万震山	万震山	狄云	2.1710626630060728
15	孙均	孙均	狄云	0.6592343258655049
16	言达平	言达平	狄云	0.9032854327427144
17	吴坎	吴坎	狄云	1.1965741810785213
18	狄云	狄云	狄云	5.205637695683047
19	卜垣	卜垣	狄云	0.8367367975564676
20	马大鸣	马大鸣	狄云	0.4053509832462059
21	梅念笙	梅念笙	狄云	0.24948272697006946
22	戚芳	戚芳	狄云	1.968484273325338
23	花铁干	花铁干	狄云	1.603194857956522
24	戚长发	戚长发	狄云	0.9026716489510798
25	汪啸风	汪啸风	狄云	0.6087639825462954
26	丁典	丁典	狄云	1.2906709700595766
27	刘乘风	刘乘风	狄云	0.4810658408360975
28	万圭	万圭	狄云	2.02231965439258
29	周圻	周圻	狄云	0.6830207577864581
30	凌霜华	凌霜华	狄云	0.16455686546458825

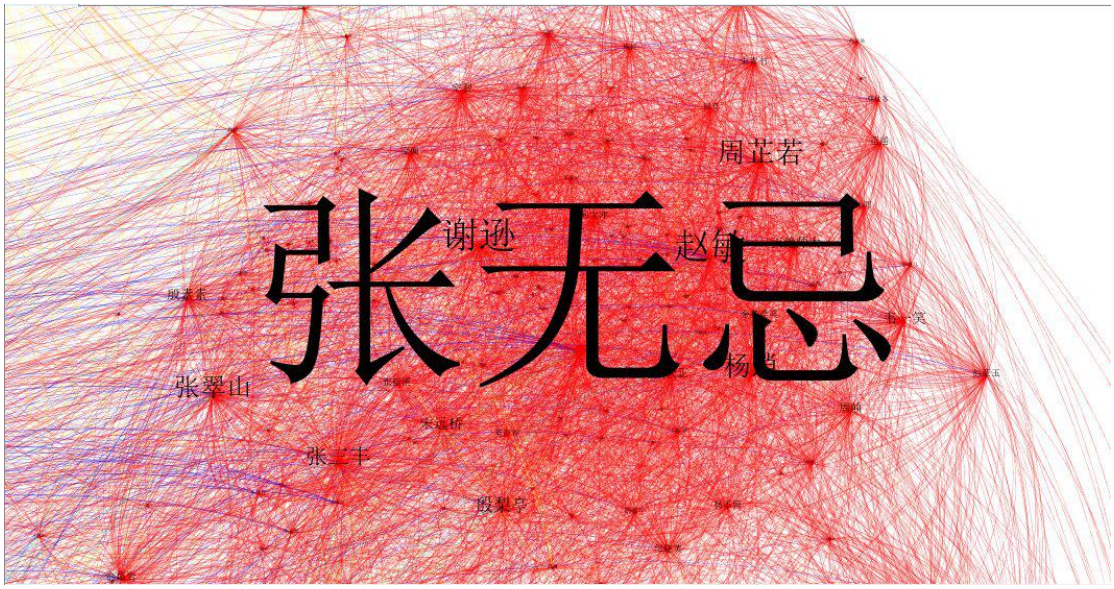
Gephi 中的边文件截图

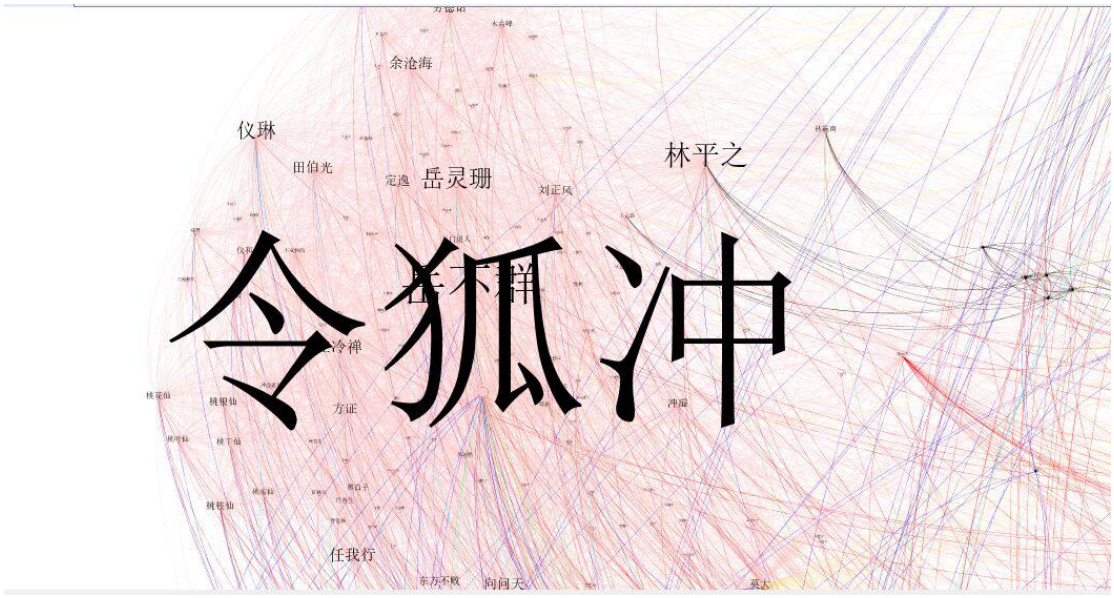
1896	彭莹玉	谢逊	0.005418
1897	徐达	谢逊	0.002322
1898	成昆	谢逊	0.034830
1899	执法长老	谢逊	0.003096
1900	掌棒龙头	谢逊	0.002322
1901	掌钵龙头	谢逊	0.002322
1902	无色	谢逊	0.000774
1903	易三娘	谢逊	0.002322
1904	本因	谢逊	0.000774
1905	朱九真	谢逊	0.002322
1906	朱元璋	谢逊	0.002322
1907	朱长龄	谢逊	0.003870
1908	杜百当	谢逊	0.000774
1909	杨不悔	谢逊	0.003096
1910	杨夫人	谢逊	0.000774
1911	杨过	谢逊	0.002322
1912	杨逍	谢逊	0.017802
1913	武烈	谢逊	0.003870
1914	武青婴	谢逊	0.002322
1915	殷天正	谢逊	0.011610
1916	殷梨亭	谢逊	0.008514
1917	殷离	谢逊	0.021672
1918	殷素素	谢逊	0.068111
1919	殷野王	谢逊	0.002322
1920	汉子	谢逊	0.013158
1921	汤和	谢逊	0.000774
1922	泉建男	谢逊	0.000774
1923	流云使	谢逊	0.008514
1924	渡劫	谢逊	0.007740
1925	渡厄	谢逊	0.012384

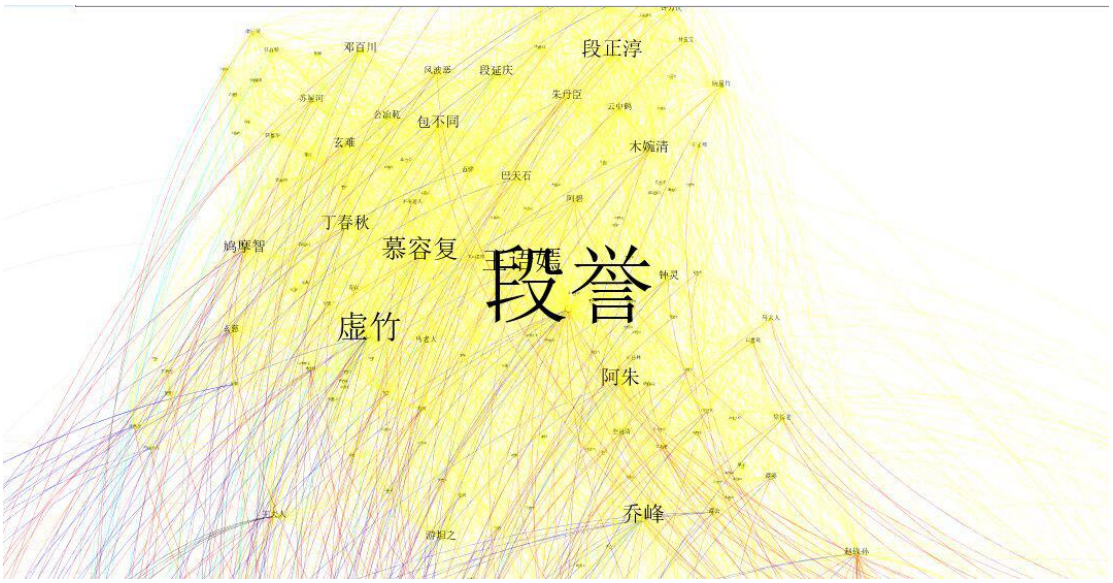
在人物关系图上的标签传播的总的结果截图

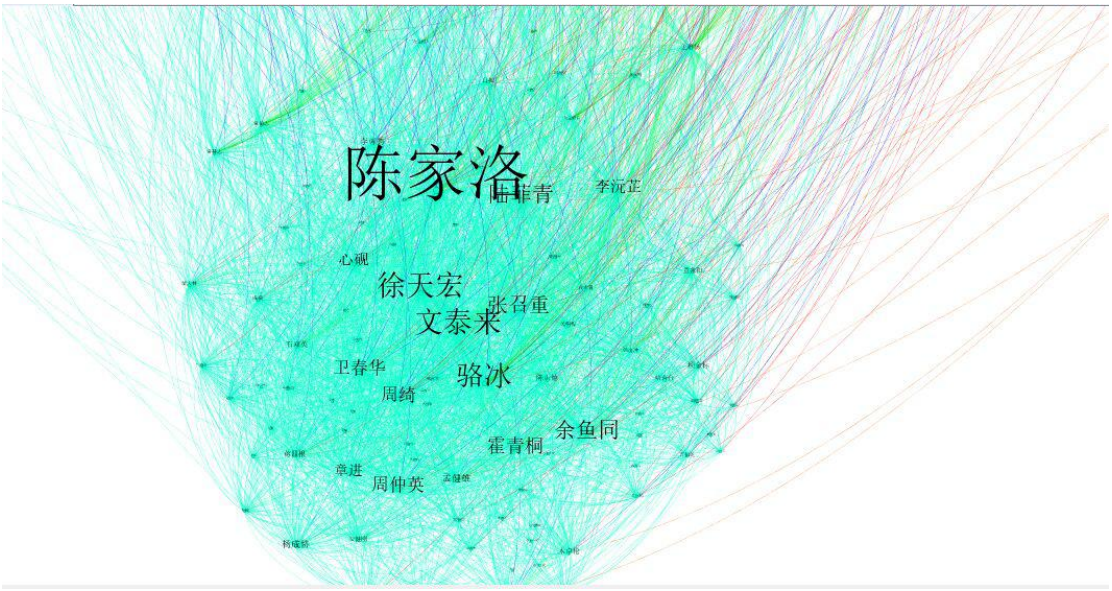


在人物关系图上的标签传播的结果细节截图









六、小组成员分工

查鹏（组长）	主要负责代码编写：jobOne，jobTwo，jobThree，pagerank 和标签传播代码的编写
许金强	主要负责开题报告的编写，gephi 绘图、实验结果数据分析和实验报告的编写

七、附录

在统计任务同现次数过程中我们曾经犯了一个错误，就是比如一行出

现的人名是 A、B、B、C、D，两两组合，只是判断不能自己和自己组合，并没有考虑重复出现这种情况，导致诸如（A,B）出现两次，这也直接影响了后面 PageRank 值的计算，导致最后结果不正确，后来我们用 java 特有机制 set，首先，把输入的一行人名直接存储在 set 里，这里 set 会自动去除重复的，然后对 set 里的人名进行两两配对，这样就解决了之前的问题。