中国人民大学社会网络挖掘——基于人民大学新闻网数据

李梓童[1)]

[1)](中国人民大学信息学院，北京市海淀区 100872)

摘 要 本文基于中国人民大学新闻网上的数据，建立了以人与人之间关系为主的社会网络图，其中，对于每一对在同一篇新闻中出现过的人物，我们认为他们之间在图中有边相连。在验证了社会网络图的准确性基础上，我们对该图进行了基础分析，如求与某个人物关系最近的 10 个人物、统计图的相关信息等。以及如下自选分析：一，影响力计算，即使用 PageRank 算法计算每个人的影响力大小，并给出影响力最大的前 20 个人；同时，为了挖掘强影响力人物和强影响力机构之间的关联，我们计算了影响力最大的前 20 个机构，并计算强影响力人物和机构同时出现的概率。二，三元闭包验证，即验证该数据上的社交网络关系是否符合三元闭包理论；三，中心性计算，即计算每个节点的中介中心性，并输出中介中心性最大的 10 个人；四，节点的聚集系数计算，即计算每个节点的聚集系数，并输出聚集系数最大的 10 个人。通过对图进行如上分析，我们对人民大学的社会网络有了更细致的洞察。

关键词 社会网络；图；数据挖掘；Pagerank

social network mining in Renmin University of China -- based on the university news website

Zitong Li[1)]

[1)](Department of Information, Renmin University of China, Haidian district, Beijing)

Abstract In this paper, we build a social network graph based on the data from the news website of Renmin University of China. Among them, for each pair of characters appearing in the same news story, we believe that they are connected by an edge in the graph. On the basis of verifying the accuracy of the social network graph, we analyze the graph as follows: First, influence calculation. the PageRank algorithm is used to calculate the size of each person's influence, and give the top 20 people with the greatest influence; At the same time, in order to explore the correlation between strong influential people and organizations, we calculated the top 20 most influential organizations and calculated the probability of the co-appearence of influential people and organizations. Second, three-way closure verification, which means to verify whether the social network relationship on the data conforms to the three-way closure theory; Three, centrality calculation, that is, to calculate the betweenness centrality of each node, and output the betweenness centrality of the maximum 10 people; Fourth, the clustering coefficient of nodes is calculated and the 10 persons with the maximum aggregation coefficient are output. Through the above analysis of the graph, we have a more detailed insight into the social network of the university.

Key words Social network; graph; data mining; Pagerank

李梓童，女，1999年生，人民大学2017级本科生，E-mail: lizitong.cc@foxmail.com.

1 引言

随着互联网的发展，个体之间的联系也从物理层面展开，出现了新的联系。两个人之间的关联，不仅可以从一次交谈、一封电子邮件建立，也可以从同时出现在一条新闻中建立。本次课程设计利用来自人民大学新闻网的新闻数据，将人物抽象为网络图。在图的层面，利用数据分析对人物之间的社会关系进行分析，会有什么样的发现呢？

2 数据预处理

本次课程设计所用新闻数据来自中国人民大学新闻网，网址为 http://news.ruc.edu.cn/archives/category/important_news。原始数据中包括新闻编号、发表时间、标题、浏览次数、作者、以及正文内容。每一行为一条新闻，其中各个字段用\t 隔开。从中挖掘出这些新闻中的社交网络关系。本次课程设计的目标为挖掘这些新闻中的社交网络关系。

2.1 抽取人名、机构名和地名

（1）这一步骤用到的工具为 jieba 中文分词工具，在程序中对应的函数为 Proc_File。

（2）基本思路如下

获取正文信息。首先读取整个 txt 文本文件作为一个字符串；然后通过 split('\n')，把数据按行分割；接着通过 split('\t')，把数据按\t 分割，获取第 6 个字段、即正文内容。

对正文内容进行 jieba 分词（jieba.posseg.cut）。此处根据分出词语的词性分别将词语归入不同的字典中，并统计词语在新闻中出现的次数。若词语在一篇新闻中出现，即计一次，在同一篇新闻中多次出现则次数不累加。

建立边权字典。对同一篇新闻中出现的所有人名进行组合，对应的二元组作为字典的键，字典的值为共同出现在新闻中的次数，可作为后来建图的边权。

（3）初次使用 jieba 分词时，发现它对人名的判断很不准确，很多如“爱尔兰”“女硕士”“满怀希望”“努力提高”这样的词也被归为人名。

由于分词准确性直接影响后面的网络分析，笔者对出现频数大于 30 的人名进行手动筛选。对于出现频数大于 30 的、既不属于人名也不属于机构名或地名的词语，笔者选择将其放入停用词列表中，即分词工具遇到这些词语时会自动忽略。另外，对于被误判为人名的机构名或地名，笔者将其在用户自定义字典中对其词性进行重新定义。

对于机构名的分类准确率较高。其中，“中国人民大学”和“人民大学”是出现次数最多的两个机构。由于语义相同，笔者选择将这两个名词的统计信息累加作为同一词语。

（4）通过对停用词和用户自定义字典进行修正后，笔者令程序输出出现频数最多的前十个人名和机构名进行准确性验证。

前 10 个人名列表如下：

表 1 前 10 位热门人物（按出现频数排序）

人名	出现频数
纪宝成	1906
靳诺	1531
陈雨露	1529
程天权	1302
习近平	1275
王利明	1142
刘伟	904
张建明	847
牛维麟	753
刘向兵	743

前 10 个机构名列表如下：

表 2 前 10 位热门机构（按出现频数排序）

人名	出现频数
中国人民大学	16414
社会科学	2403
法学院	1601
商学院	1204
北京大学	1175
宣传部	1094
清华大学	1014
中国共产党	981
国际关系学院	865
国际化	844

2.2 建立社交网络图

（1）这一步骤用到的工具为 networkx，在程序中对应的函数为 createGraph。

（2）基本思路如下：

读取 1.1 中建立的边权字典。其中，字典的键为两个人名（节点），字典的值为人名同时出现的新闻数（边权）。根据字典的每一项，将带权边加入社交网络图中。

3 基础内容

基础内容包含两方面：一是图的验证，即提供界面，输入一个人 A 进行查询，可以输出和 A 关系最强的前 10 个人（邻居）；二是图的统计，即计算图的节点个数，边数，连通分量的个数，最大连通分量的大小。

3.1 图的验证

（1）这一部分内容用的主要工具是 networkx，在程序中对应的函数为 Findtop10neighbors。

（2）基本思路如下：

通过 networkx 中的 neighbors 方法，获取想要查询节点的所有邻居。对所有与邻居相连的边按照权值进行排序，排序后输出前 10 个权值最大的邻居。

（3）尝试查询与“靳诺”相连的前 10 个边权值最大对应的节点，输出结果如下：

表 3 与“靳诺”相连的前 10 个权值最大边对应的节点

人名	关联边权值
习近平	518
刘伟	442
张建明	419
王利明	357
陈雨露	265
贺耀敏	219
杜鹏	215
伊志宏	182
吴晓球	180
顾涛	174

（4）尝试查询与“杜鹏”相连的前 10 个边权值最大对应的节点，输出结果列表如下：

表 4 与“杜鹏”相连的前 10 个权值最大边对应的节点

人名	关联边权值
靳诺	215
刘伟	160
习近平	148
张建明	126
王利明	116
贺耀敏	78
吴晓球	77
周荣	62
顾涛	56
冯惠玲	52

3.2 图的统计

（1）这一步骤用到的工具为 networkx，在程序中对应的函数为 createGraph。

（2）基本思路如下：

通过调用 networkx 中的 number_of_nodes、number_of_edges、number_connected_components、connected_components 方法，获取所建立图的节点个数，边数，连通分量的个数，最大连通分量的大小。

（3）输出结果中，人物图节点数为 29865，边数为 716921，联通子图数为 15，最大联通子图大小为 29832。据此可以推测图的结构，即一个巨大的连通分支周边有若干微型连通分支。

另外，我们也根据机构名建立了图，机构的建图规则与人物图相同，即若两个机构在同一篇新闻中出现，则两机构在图中相关联。机构图的节点数为 2263，边数为 82104，连通子图数为 1，最大连通子图大小为 2263。

4 自选分析

自选分析内容包含四部分：一，影响力计算，即使用 PageRank 算法计算每个人的影响力大小，并给出影响力最大的前 20 个人；二，三元闭包验证，即验证该数据上的社交网络关系是否符合三元闭包理论；三，中心性计算，即计算每个节点的中介中心性，并输出中介中心性最大的 10 个人；四，节点的聚集系数计算，即计算每个节点的聚集系数，并输出聚集系数最大的 10 个人。

4.1 影响力计算

（1）这一步骤用到的工具为 networkx，在程序中对应的函数为 PageRank。

（2）基本思路如下：

已知有向图的 pagerank 计算公式如下：

$$PR(p_i) = \frac{1-d}{n} + d\sum_{p_j \epsilon M(p_i)} \frac{PR(p_j)}{L(p_j)} \quad \text{(a)}$$

其中$PR(p_i)$表示p_i节点的 pagerank 值，d 表示

阻尼系数，M(p_i)表示p_i节点的邻居节点集合，$L(p_j)$表示p_j节点的出度。

首先，通过 networkx 的 to_directed 方法将无向图转化为有向图。然后，调用 pagerank 方法，计算各个节点的 pagerank 值。

最后，每个节点值的 pagerank 值计算完毕后，使用 sorted 函数对形式为{节点：pagerank 值}的字典进行降序排序，输出前 20 个节点，即 pagerank 值最大的前 20 个节点。

（3）输出结果列表如下：

表 5 人物图 Pagerank 值最大的前 20 个节点

人名	pagerank 值
靳诺	0.008714362
纪宝成	0.008421764
陈雨露	0.007739211
习近平	0.006957247
程天权	0.006342436
王利明	0.006263157
刘伟	0.005410301
张建明	0.004783634
冯惠玲	0.00414248
杜鹏	0.003852915
伊志宏	0.003745976
刘向兵	0.003557508
袁卫	0.003533537
杨慧林	0.003394729
牛维麟	0.003380261
孔子	0.003021197
贺耀敏	0.002788875
林岗	0.002261841
吴晓球	0.002240923
吴付来	0.00221171

（4）同时我们对机构图进行 pagerank 计算，输出结果列表如下：

表 6 机构图 Pagerank 值最大的前 20 个节点

机构名	pagerank 值
中国人民大学	0.034678417
社会科学	0.02174657
北京大学	0.015759132
清华大学	0.0143462
法学院	0.012881575
中国共产党	0.011060383
国务院	0.009512513
商学院	0.009101732
中国社会科学院	0.00862366
北京师范大学	0.008605785
国际化	0.008268438
宣传部	0.008201817
国际关系学院	0.007113299
新华社	0.006192196
全国政协	0.006082259
党中央	0.005766178
复旦大学	0.005408691
南京大学	0.004824436
武汉大学	0.004759219
文学院	0.004643384

（5）和 1.1.（4）中的热门人物、机构列表比较后，我们发现，热门人物、机构和影响力最大的人物、机构有很高的重合度。也就是说，一个人或机构在新闻中出现的频率越高，其影响力也往往就越大。

（6）在此基础上，我们计算了在一篇新闻中，若出现了一个影响力排在前 20 为的机构名，则出现影响力排名前 20 位人物名的概率。具体计算方式为：统计新闻中出现了强影响力机构和既出现强影响力机构、又出现强影响力人物的次数，次数不重复，即多次在新闻中出现也只计一次。计算后者和前者的比值，最终得到重合比例为 93.57%。即若一篇新闻中出现强影响力机构，则出现强影响力人物的概率为 93.57%。

与之相对称的，我们计算了在一篇新闻中，若出现了一个影响力排在前 20 为的人名，则出现影响力排名前 20 位机构名的概率，其值为 72.85%。

据此我们分析，强影响力机构与强影响力人物之间有着很强的关联性。尽管强影响力机构和强影响力人物同时出现的概率很高（超过了 50%），但是，若一个机构是强影响力机构、则与其同时出现的人物是强影响力人物的可能性大于若一个人是强影响力人物、则与其同时出现的机构是强影响力机构的可能性。例如，若“习近平”出现在一篇新闻中，则这篇新闻很可能出现强影响力机构如“中国人民大学”“中国共产党”“国务院”等，但出现的可能性小于若“中国人民大学”“中国共产党”“国务院”出现在一篇新闻中，则这篇新闻出现“习近平”的概率。

（7）实验开始，我们尝试用网上搜索得到无向图的 pagerank 公式计算 pagerank 值：

$$PR(p_i)=\frac{1-d}{n}+d\sum_{p_j\epsilon M(p_i)}\frac{weight(pj)\times PR(p_j)}{degree(p_j)}$$

其中$PR(p_i)$表示p_i节点的 pagerank 值，d 表示阻尼系数，M(p_i)表示p_i节点的邻居节点集合，$degree(p_j)$表示p_j节点的度数，$weight(pj)$表示关联边的权重。

但经测试，该公式得到 pagerank 值并不收敛，在本数据集中所有节点的 pagerank 都随迭代次数增加而递增，故选择了现在的思路，即先将无向图转化为有向图再计算 pagerank 的方法。

4.2 三元闭包验证

（1）这一步骤用到的工具为 networkx，在程序中对应的函数为 ProveClosure。

（2）课堂讲授的示例中，三元闭包通过检测共同好友数和成为好友之间的概率关系来验证[2]。参考这一示例，在本次课程设计中对三元闭包验证的基本思路如下：

对社交网络图 G 中的节点进行遍历，记每一个遍历到的节点为 a，a 的邻居节点集合为 M(a)。对 M(a)中的节点进行遍历，记遍历到的节点为 b，则 a、b 之间有一条相连的边。记 b 的邻居节点集合为 M(b)，对 M(b)中的节点进行遍历，记遍历到的节点为 c。若 a、c 之间存在边，则 a、b 之间存在一个共同好友 c。

据此，通过对 G 中的好友对(a,b)进行计算，可得(a,b)的共同好友 c 的数目。用一个以 c 的数目为键、好友对的数目为值的字典来存储所得结果，则每一个共同好友数量对应的好友对数量除以 G 中总共的好友对数量（即边数），即可得成为好友的概率。

（3）在程序实际运行过程中，我们对节点进行了选择。例如，只对由出现频数大于 5、大于 10、大于 50 的节点构成的节点对进行统计。另一方面，出现频数过小的点，在图中出于边缘地带，我们认为舍去能提高图分析的准确性。

另外，对于共同好友数大于等于 10 的节点对，我们统一将其归为一类。即最终所得共同好友数的统计标量中，其范围是 0,1,2,…,8,9,>=10。

（4）输出结果列表如下：

表 7 节点对共同好友数统计（节点出现频数不小于 50）

共同好友数	节点对数
>=10	136
0	73

共同好友数≥10 的节点对数: 无共同好友的节点对数=1.863:1

表 8 节点对共同好友数统计（节点出现频数不小于 10）

共同好友数	节点对数
>=10	842
0	523
9	10
7	4
6	4
3	2
8	1
4	1

共同好友数≥10 的节点对数: 无共同好友的节点对数=1.610:1

表 9 节点对共同好友数统计（节点出现频数不小于 5）

共同好友数	节点对数
>=10	1732
0	1450
9	53
8	38
6	28
7	27
5	26
4	15
3	13
2	3
1	3

共同好友数≥10 的节点对数: 无共同好友的节点对数=1.194:1

总的来看，如果一对节点对是好友，那么它们之间存在共同好友的概率要大于没有共同好友的概率，即符合三元闭包规则。

另外，我们还发现这些有趣的现象：在此社交网络图中，如果一对节点对是好友，那么绝大多数情况下，它们之间的共同好友数要么很多（大于等于 10 个），要么没有。

同时，随着节点出现频数的增加，共同好友数很多的节点对所占比例也在增大。

4.3 中心性计算

（1）这一步骤用到的工具为 networkx，在程序中对应的函数为 Betweenness_nx。

（2）基本思路如下：

采用 Networkx 中提供的计算中介中心性（一个节点担任其它两个节点之间最短路的桥梁的次数）的函数 betweenness_centrality。该函数根据节点与节点间的最短路径计算中介中心性[3]。

中介中心性计算公式如下：

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

其中 V 表示节点集，$\sigma(s,t)$ 表示 s 到 t 的最短路径的数量，$\sigma(s,t|v)$ 表示 s 到 t 的最短路径中通过 v 的路径数量。如果 $s=t$，则 $\sigma(s,t)=1$。如果 $v \in \{s,t\}$，$\sigma(s,t)=0$。

通过 betweenness_centrality 函数获取 G 中各节点的中心性后，利用 sorted 对各节点的中心性进行排序，输出中介中心性最大的 10 个节点。其中函数的参数 k=20，表示该函数在计算中心性时会随机选择 20 个节点作为种子。k 越接近 G 的节点数，则准确性越高，但当 k 过大时运行时间过长，故经过测试，我们将 k 选在 20。同时，由于种子选取方式随机，每次中介中心性输出情况都会有所差异。

（3）输出结果列表如下：

表 10 中介中心性最大的 10 个节点

人名	中介中心性
纪宝成	0.061798211
张建明	0.058429356
陈雨露	0.044441673
靳诺	0.042285636
郑风田	0.039412832
王利明	0.036834111
张晓京	0.031927353
杨妮	0.028768565
牛维麟	0.0280175
冯惠玲	0.02570565

4.4 节点的聚集系数计算

（1）这一步骤用到的工具为 networkx，在程序中对应的函数为 Clusteringcoefficient。

（2）基本思路如下：

无向图中，聚集系数的计算公式为[4]：

$$C(v) = \frac{E(M(v))}{C^2_{degree(v)}}$$

其中，$E(M(v))$ 表示 v 的邻居节点中存在的边数，$degree(v)$ 表示 v 的度数。

根据此计算公式，我们对每个节点的所有邻居进行循环遍历，找出其中存在边的邻居数，最后除以邻居之间可以存在的最大边数，便可得到每个节点的聚集系数。

对聚集系数进行排序，最后输出聚集系数最大的前 10 个节点。

（3）输出结果列表如下：

表 11 聚集系数最大的前 10 个节点

人名	聚集系数
马小兰	0.333333333
张正	0.333333333
梁冉	0.333333333
于静冉	0.333333333
郑涵	0.333333333
刘灿河	0.333333333
郑卫	0.333333333
毕晓洋	0.333333333
郭希敏	0.333333333
顾涛	0.333333333

根据上表可见，许多节点有着相同的聚集系数。作为对比，我们查看了“刘伟”的聚集系数 9.3955×10^{-7}，“靳诺”的聚集系数 4.7358×10^{-7}，“杜鹏”的聚集系数 3.9156×10^{-7}，“习近平”的聚集系数 0.0001342。我们可以推测，度数较多、出现频数较高节点的聚集系数低的原因是，邻居众多而邻居之间存在的关联边较少。但对于聚集系数高的节点群中，节点聚集系数存在高一致性这一现象，我们推测原因是这些节点处在一个边缘的小连通分支中，而不是前面统计得到的有两万八千多个节点的大连通分支中。这些处在小连通分支里的节点邻居与邻居之间联系紧密，从而得到了较高的聚集系数。

5 结论

在本次课程设计中，我们对中国人民大学新闻网上的数据进行了处理与分析，建立了以人与人之间关系为联系的社会网络图和以机构与机构之间

关系为联系的社会网络图。通过对图的分析，我们可以得到与某个人物关系最近的 10 个人物、图的统计信息等，并作了如下拓展分析：一，使用 PageRank 算法计算每个人和机构的影响力大小，并给出影响力最大的前 20 个人和前 20 个机构，并计算强影响力人物和机构同时出现的概率。二，验证该数据上的社交网络关系是否符合三元闭包理论；三，计算中介中心性，并得到中介中心性最大的 10 个人；四，计算聚集系数，并得到聚集系数最大的 10 个人。通过对社会网络图的分析，我们对人民大学的社会网络有了更细致的洞察。

参 考 文 献

[1] Langville A , Meyer C . Deeper Inside PageRank[J]. Internet Mathematics, 2004, 1(3):335-380.

[2] Kossinets, G. Empirical Analysis of an Evolving Social Network[J]. Science (Washington D C), 2006, 311(5757):88-90.

[3] Brandes U . On variants of shortest-path betweenness centrality and their generic computation[J]. Social Networks, 2008, 30(2):136-145.

[4]Eggemann N , Noble S D . The Clustering Coefficient of a Scale-Free Random Graph[J]. Discrete Applied Mathematics, 2011, 159(10):953–965.