大作业实验报告

张配天[1)]

[1)](中国人民大学 信息学院, 北京中国 100872)

摘 要 随着大数据日渐渗透人们的日常生活, 伴其产生的复杂网络诸如社交网络, 知识图谱的发展也越来越迅速, 这些网络中的节点表示现实世界中的各个事物或者概念, 而事物间的联系用网络中节点间的连接表示。如何分析这些网络, 如何从这些复杂网络的庞大数据中抽取得到有价值的信息成为一项重要的研究；本次课程设计聚焦于从中华人民共和国中央政府新闻网爬取的从 2016 到 2020 年间发布的新闻, 从中提取人物和机构作为图的节点, 将在不同节点出现在同一篇新闻视为共现关系, 在此基础上构建共现网络, 并利用 Neo4j 作为工具对该网络进行一系列分析: 除了对节点, 边, 联通分量的统计信息；查询某一节点的邻居, 计算 PageRank 的基础功能外, 实现了**最短路径计算, 社区挖掘, 节点中介中心性计算和节点聚集系数计算的四个任务**, 我将在实验部分详细呈现整个实验内容和结果。

关键词 网络分析

中图法分类号 TP391 DOI 号 10.11897/SP.J.1016.01.2020.00001

Final Task Experimental Report

Peitian Zhang[1)]

[1)](Department of Information,Renmin University, 100086, China)

Abstract With the increasing penetration of big data into people’s daily life, the complex networks that it produces, such as social networks, are developing more and more rapidly. The nodes in these networks represent various things or concepts in the real world, and the connections between them are fomulated as edges in the network. How to analyze these networks and how to extract valuable information from the huge data from these complex networks has become an important research for about these networks; This course project focuses on news crawled from *http://www.gov.cn/xinwen/* between 2016 and 2020, from which people and institutions entities are exxtracted as nodes in the graph, while treating different nodes appearing in the same news as a co-occurrence relationship.Therefore, a co-occurrence network are established. I use Neo4j as toolkit to perform a series of analysis on the network, basic analysis of which includes statistical information on nodes, edges, and connectivity components , query the neighbors of a certain node and calculation of PageRank. Besides these basic functions, I achieve following superior analysis: **shortest path calculation, community detection, node betweenness centrality calculation and node clustering coefficient calculation four tasks) are realized. I will present the whole experiment in detail in the third section.**

Key words network analysis

1 引言

随着科技的发展, 越来越多的数据渗透在人们的生活中；这些数据有各种各样的类型和结构, 其中图数据占据了很重要的位置。人类的诸多行为都自然地形成了图的形式, 比如以人作为节点, 关注作为关系的社交图；又比如以实体作为节点, 实体之间的联系作为关系的知识图谱... 进一步地, 在这些图的边上赋以权重, 这些权重通常情况下代表了关系的强度, 则会得到一个个“网络”。由于图和网络在很多方面天然地建模了人的诸多行为, 且图上节点之间的关系蕴含了很多信息尚待挖掘, 因此建立在这些图和网络上的研究正在受到越来越多的关注。

本实验的目标在于从一个现实生活中的图上出发, 进行一系列分析, 验证课堂上涉及的部分理论, 并挖掘出图中有价值的信息。在数据方面, 采用从政府新闻网爬取的两万条新闻, 从中抽取人物和机构作为图的节点, 将出现在同一篇新闻中的节点之间连边, 视作共现关系, 共现的次数即为边的权重, 构建了包含有 22306 个节点, 109344 条边的实体共现网络。由于政府新闻代表了国家的声音, 越

收稿日期：2020-12-31；修改日期：2020-12-31 张配天，男，2000 年生，本科，否，主要研究领域为新闻推荐 E-mail: zpt@ruc.edu.cn.
第 1 作者手机号码：18801200703，E-mail: zpt@ruc.edu.cn

重要的事物会得到越多的报道, 且共现的两个实体可以视作出现在同一事件之中, 即发生过主动或者被动的互动, 因此本网络在一定程度上可以解释为对国家而言重要的人物和机构, 以及他们之间互动的频率。

于是, 对这个网络的分析可以解释为对这些重要人物和机构的分析, 从中通过 PageRank 可以发掘出影响力大的实体；通过最短路径计算可以验证任意两个实体之间的路径, 即实体的能量通过事件传输的可能性, 从而在一定程度上验证小世界理论；通过社区检测算法可以计算出弱连接的社区, 观察同属于同一个社区的实体的性质；通过中心度计算可以从某个方面得到一个实体支配资源的能力；通过聚集系数计算可以反映实体之间的相互影响并在一定程度上验证三元闭包理论的正确性。基于这些目标, 我完成了上述各个指标的计算, 并得到了相应的结论, 具体的细节将会在实验部分详细阐述。

2 相关工作和实验意义

本实验使用的各个算法在 Neo4j 中都已有高效实现, 这里将之稍作总结, 同时给出我完成的每一项任务在此实体共现网络上的意义的详细分析。

2.1 PageRank

PageRank 算法是用来给互联网网页评分的重要算法之一, 其能基于入边的个数和入边的重要性给当前节点评分, 使重要的节点获得更高的分数。根据课上所学内容以及 Google 原论文, 简要给出 PageRank 的公式, 记图 G 的邻接矩阵为 M , 有 $M_{i,j} = \frac{1}{out_i}(e(i,j) \in E)$ 或者 $M_{i,j} = 0(e(i,j) \notin E)$, 其中 out_i 为第 i 个节点的出度, 有

$$[p_1, p_2, \cdots, p_n]^{(T+1)} = [p_1, p_2, \cdots, p_n]^{(T)} * M$$

记 M 的转置为 M', 有

$$[p_1, p_2, \cdots, p_n]^{(T+1)'} = M' * [p_1, p_2, \cdots, p_n]^{(T)'}$$

, 在此基础上实现同比缩减和补偿, 有

$$[p_1, \cdots, p_n]^{(T+1)'} = \alpha * M' * [p_1, \cdots, p_n]^{(T)'} + (1-\alpha) * [\frac{1}{n}]'_n$$

其中 $[\frac{1}{n}]_n$ 为各个元素均为 $\frac{1}{n}$ 的行向量；迭代计算 k 次, 直到最终收敛, 即得到每个节点的 PageRank 值。

在本实验使用的网络中, 每个节点也由于不同的共现关系拥有不同的重要性 (e.g. 和“习近平”共现的节点的重要性要高于和“叶菲莫娃”共现的节点), 因此符合 PageRank 的特性, 即可以使用 PageRank 来量化节点的重要性。

2.2 单源最短路径

单源最短路径旨在寻找从给定起始节点到图上剩余所有节点的最短路径, 常使用 Djikstra 算法, 本实验中最短路径计算也采用该算法。在实体共现网络上, 由于共现关系反映了实体和实体之间的互动, 即两者主动或被动地参与了同一个事件, 因此可以将最短路径的意义描述为: 一个实体 (人物或机构) 的权威性的影响范围, 这个影响范围通过实体参与的不同事件得以扩张。

由此我们可以验证小世界理论——图上存在丰富的短路径, 即图上的实体有着隐含的庞大的影响范围, 该范围通过事件得以传递。

2.3 社区检测

社区检测旨在发掘图中聚类的节点和被划分开的节点, 以及这些节点分开的倾向。Louvain 算法是社区挖掘最为经典的算法, 该算法最大化各个社区的 modularity, 该 modularity 衡量了一个社区中的节点之间彼此连接有多“紧密”。具体的细节这里不再赘述。

在本实验中进行社区挖掘的目的有二:

- 统计整个图的聚集程度
- 观察一个社区内节点的相似性质, 以此从一种角度验证图的同质性

2.4 中介中心性

中介中心性反映了一个节点对于整个图上的资源流动的控制能力, 其值越高, 控制力越强。计算公式为

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{1}$$

其中 $C_B(v)$ 即为中介中心性。

在本实验中由于共现关系反映了实体之间的互动, 这种互动可以视作是该网络上能量/资源的传递, 因此计算节点的中介中心性是有意义的, 中介中心性越高的节点将对整个网络上各个事件有着更强的掌控力。

2.5 聚集系数

节点的聚集系数表达了其邻居之间存在关系的可能性, 有公式如下

$$C(v) = \frac{2T_v}{d_v * (d_v - 1)} \tag{2}$$

在本实验中的聚集系数体现了事件发生在邻近节点之间的可能性, 即如果节点 A 分别和节点 B,C 互动 (存在边 $e(A,B), e(A,C)$), 那么大概率有 $e(B,C) \in E$ 即 B 和 C 之间也有互动；因此, 计算聚集系数可以很好地反映实体之间的相互影响。

3 实验

3.1 数据

为了构建实体共现网络, 需要将文本内容处理为邻接矩阵, 该过程需要经过以下三步:

3.1.1 预处理

官方提供的数据中包含空值, 本过程将空值、重复内容全部剔除；

3.1.2 分词与构建预料

新闻中是连贯的文本, 需要将其分词。由于 jieba 分词的词性标注效果很差, 本过程使用 thulac 包对文本进行分词, 其准确度显著高于 jieba。分词结束后, 遍历分词结果, 统计各个实体出现的次数, 形成 word_count 字典和 part_of_speech 字典；之后抽取需要的两种实体, 即人名和机构名, 直接查表即可得出相应类型的实体的分布, 将出现次数排序, 即可得到热门人物和机构。

除此之外, 需要根据抽取出来的实体节点构建字典, 即将实体对应给唯一的 id, 并据此生成反字典, 即将每个 id 映射回实体。

3.1.3 邻接矩阵构建

本步骤构建实体共现网络的邻接矩阵 $\boldsymbol{M}$。出现在同一篇新闻中的实体 i 和实体 j 之间具有可累计的共现关系, 在邻接矩阵中对应元素即 $\boldsymbol{M}_{i,j}$ 上加一；由于共现关系可能在一篇新闻中由 i 指向 j, 在另一篇新闻中由 j 指向 i 且其实共现关系没有指向性, 因此需要将得到的邻接矩阵 $\boldsymbol{M}$ “对称化”, 即

$$\boldsymbol{M}_{i,j} = \frac{\boldsymbol{M}_{i,j} + \boldsymbol{M}_{j,i}}{2} \tag{3}$$

由此, 我们将所有的共现关系均匀地分配到头结点和尾结点上, 即将无向的共现关系以对称的有向形式表示出来, 在不破坏边的性质的同时方便之后 PageRank 等算法的计算。

3.1.4 构建网络

得到邻接矩阵后, 本步骤使用 *Neo4j admin-import* 工具将其导入 Neo4j 中, 构建最终的实体共现网络。

3.2 实现技术

在总和考量了可视化、算法运行效率等因素后, 我使用 Neo4j+python 的工具组合完成本实验；由于 Neo4j 优秀的可视化能力, 整个实体共现网络一目了然, 井井有序；同时, 由于其出色的并行能力和成熟的算法封装, 本实验的所有功能均可以在 **3s** 之内计算完成并输出结果；

本实验涉及的功能只是 Neo4j 的一小部分, 这次对于该工具的接触和学习将为我之后的一些工作打下良好的基础。

3.3 结果

本部分将按照课程设计中给出的顺序展示每一项功能的运行结果。

3.3.1 热门人物和机构

排行前十的热门人物如下表所示:

表 1 热门人物表

热门人物	出现次数
习近平	27297
李克强	19330
王毅	4660
韩正	1947
汪洋	1601
胡春华	1308
刘延东	1154
张高丽	1087
孙春兰	1078
杨洁篪	1046

可以看出在热门人物中出现的均为国家政要, 符合预期分布。

排行前十的热门机构如下表所示:

表 2 热门机构表

热门机构	出现次数
新华社	48845
国务院	20420
中共中央	6294
党中央	6098
联合国	3959
财政部	3184
中央军委	2133
公安部	1643
教育部	1458
人社部	857

热门的机构均为国家重要机关部处, 同时由于大多数新闻中都会提到“新华社报道”, 因此新华社的出现次数最高, 符合预期分布。

3.3.2 建立社交网络图

该社交网络图即实体共现图, 由于图片大小限制这里仅匹配 20 个节点, 将其可视化后呈现如下:

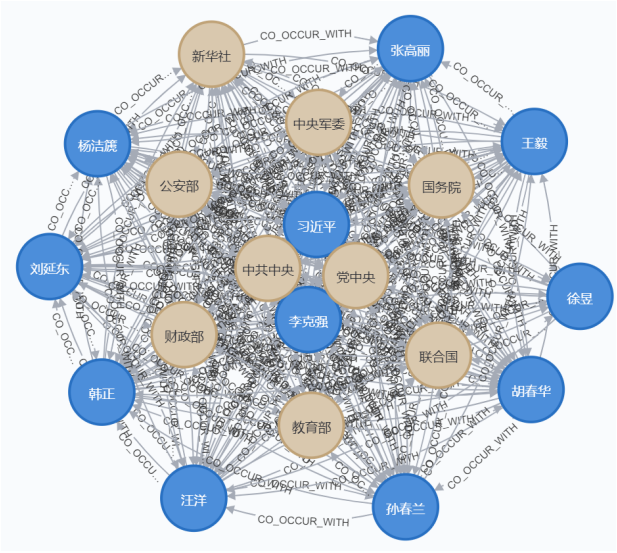

图 1 浅色节点为机构, 深色节点为人物, 边为共现关系

3.3.3 基础功能

1. **图的验证:** 这里以查询“习近平”为例, 结果如下图:

```
Node:习近平
新华社, CO_OCCURANCE_COUNTS: 4000
国务院, CO_OCCURANCE_COUNTS: 2295
中共中央, CO_OCCURANCE_COUNTS: 1997
党中央, CO_OCCURANCE_COUNTS: 1898
李克强, CO_OCCURANCE_COUNTS: 1157
王毅, CO_OCCURANCE_COUNTS: 881
中央军委, CO_OCCURANCE_COUNTS: 725
联合国, CO_OCCURANCE_COUNTS: 634
何立峰, CO_OCCURANCE_COUNTS: 557
杨洁篪, CO_OCCURANCE_COUNTS: 520
```

图 2 图的验证 (查询邻居)

2. **图的统计:** 经过计算, 该实体共现网络上一共有 22306 个节点, 218688 条边, 包含 15616 个强联通分量, 其中最大的强联通分量中有 6656 个节点;

3. **影响力计算:** 使用 damping factor 为 0.85 的 PageRank 计算结果中前二十名如下表所示:

```
No.1: Node:新华社, PageRank:515.1097936034203
No.2: Node:国务院, PageRank:254.92586806416512
No.3: Node:习近平, PageRank:160.3539272427559
No.4: Node:李克强, PageRank:80.73677828907967
No.5: Node:党中央, PageRank:67.97083262205125
No.6: Node:中共中央, PageRank:58.81735602468253
No.7: Node:联合国, PageRank:58.78994832336902
No.8: Node:王毅, PageRank:43.638569520413874
No.9: Node:财政部, PageRank:39.93896519839763
No.10: Node:何立峰, PageRank:22.990929841995243
No.11: Node:中央军委, PageRank:20.797505190968515
No.12: Node:杨洁篪, PageRank:19.634477253258222
No.13: Node:公安部, PageRank:19.023527302593
No.14: Node:丁薛祥, PageRank:17.196882846951485
No.15: Node:教育部, PageRank:16.712425094097853
No.16: Node:韩正, PageRank:14.674566375464197
No.17: Node:汪洋, PageRank:13.02538147941232
No.18: Node:肖捷, PageRank:12.626122868806123
No.19: Node:清华大学, PageRank:12.453439006954433
No.20: Node:刘鹤, PageRank:12.025691653043031
```

图 3 PageRank 前二十名

3.3.4 自选功能

1. **小世界理论验证:** 经过计算, 所有点对之间的有限长度的最短路径的平均值为 **2.33474**, 这显著地体现了图上存在着丰富的短路径。根据之前的分析, 由于图的边即共现关系反映了两个实体在同一事件中主动或被动地发生互动, 这些短路径的意义就可以解释为远距离的实体之间可以通过很少的事件产生互动和联系, 这种性质反映了实体潜在的影响范围很大, 仅通过

不到 3 个事件就可以将能量传递给图上的其他节点。

2. **社区挖掘:** 经过多轮计算, 实体共现网络中平均有 **40** 个社区, 其平均 modularity 为 **0.486943**; 其个数显著低于强连通分量, 体现了图上存在着很多相连但是不够紧密的关系, 这些共现关系可能表现为一些报道中偏向平民化的节点和重要的节点产生了互动。作为案例分析, 我发现一个社区中仅包含五个节点, 分别是“李某”,“张某”,“王某”,“赵某”,“刘某”, 可以猜测是一个案情的报道, 这些节点一般不会与别的节点产生联系, 因此处于同一个社区; 同时由于 Louvain 算法迭代进行并且种子随机, 其结果可能每次都不相同, 导致节点在不同的运行中属于不同的社区; 因此社区数和平均 modularity 是较有意义的统计量。

3. **中心性计算:** 经过计算, 中介中心性排名前 10 的实体节点如下图

```
calculating Betweenness Centrality...
No.1: Node:习近平, Betweenness Centrality:7418146.666844221
No.2: Node:李克强, Betweenness Centrality:2315457.122560973
No.3: Node:王毅, Betweenness Centrality:564679.3142469502
No.4: Node:郝群英, Betweenness Centrality:195027.0001320839
No.5: Node:王晓, Betweenness Centrality:187786.66477903462
No.6: Node:张立群, Betweenness Centrality:163215.70851331882
No.7: Node:耿玉, Betweenness Centrality:157819.11586230958
No.8: Node:曾金华, Betweenness Centrality:142879.12735758704
No.9: Node:何立峰, Betweenness Centrality:116848.4169411674
No.10: Node:刘满仓, Betweenness Centrality:103403.77970907082
```

图 4　中介中心性最大的十个人物

可以看出, 其节点排序和 PageRank 中相同节点的排序相同, 根据上述的分析, 这是因为中介中心性反映了实体对于整个网络上资源的支配能力, 这是节点重要性的一个重要体现, 而 PageRank 也反映了各个节点的重要性, 因此两者中重合节点的排序大致相同是符合预期的。

同时, 可以计算中介中心性最大的十个机构以及排名前十的实体, 结果如下图

```
calculating Betweenness Centrality...
No.1: Node:新华社, Betweenness Centrality:1989.4655899655897
No.2: Node:国务院, Betweenness Centrality:667.41558996559
No.3: Node:党中央, Betweenness Centrality:151.5639665889665
No.4: Node:联合国, Betweenness Centrality:144.32046287046282
No.5: Node:财政部, Betweenness Centrality:141.52035187035187
No.6: Node:中共中央, Betweenness Centrality:129.26396658896658
No.7: Node:塔斯社, Betweenness Centrality:87.06666666666668
No.8: Node:教育部, Betweenness Centrality:73.09812964812966
No.9: Node:中央军委, Betweenness Centrality:37.21111666111668
No.10: Node:公安部, Betweenness Centrality:20.527891552891553
```

图 5　中介中心性最大的十个机构

```
No.1: Node:新华社, Betweenness Centrality:13304922.285703188
No.2: Node:国务院, Betweenness Centrality:5423333.645671296
No.3: Node:习近平, Betweenness Centrality:2434832.6349466345
No.4: Node:李克强, Betweenness Centrality:943569.9637506976
No.5: Node:党中央, Betweenness Centrality:561872.3772741603
No.6: Node:联合国, Betweenness Centrality:512165.8041273183
No.7: Node:财政部, Betweenness Centrality:377466.08903910924
No.8: Node:中共中央, Betweenness Centrality:324328.91936996963
No.9: Node:王毅, Betweenness Centrality:208666.9396686756
No.10: Node:公安部, Betweenness Centrality:145696.89856921084
```

图 6　中介中心性最大的十个实体

由于我分别构建了只包含人物实体的网络, 只包含机构实体的网络和包含所有实体的网络, 因此三个截图中相同节点具有不同的分数, 但可以发现将人物实体和机构实体汇入同一张图后其人物之间的相对顺序和机构之间的相对顺序是没有改变的, 这是由于涉及到这些人物/机构的新闻有更大的可能性同时涉及别的实体;

进一步, 可以发现最终的实体共现网络中机构的中介中心性高于人物的, 这反映了机构对于资源的支配能力高于人物, 是健康的治国理政的模式。

4. **节点的聚集系数计算:** 经过计算, 整张图的节点聚集系数平均值为 **0.6381378**, 排名前十的人物实体如下图:

```
calculating Clustering Coefficient...
No.1: Node:黄守宏, Clustering Coefficient:1.0
No.2: Node:姚依林, Clustering Coefficient:1.0
No.3: Node:杨梅, Clustering Coefficient:1.0
No.4: Node:傅建斌, Clustering Coefficient:1.0
No.5: Node:陈泽国, Clustering Coefficient:1.0
No.6: Node:丁根厚, Clustering Coefficient:1.0
No.7: Node:郑万高铁, Clustering Coefficient:1.0
No.8: Node:李鹏, Clustering Coefficient:1.0
No.9: Node:刘东君, Clustering Coefficient:1.0
No.10: Node:陈晓东, Clustering Coefficient:1.0
```

图 7　聚集系数最大的十个人物

同样地可以得出聚集系数最大的前十个机构, 如图

```
calculating Clustering Coefficient...
No.1: Node:联合社, Clustering Coefficient:1.0
No.2: Node:三江源国, Clustering Coefficient:1.0
No.3: Node:法新社, Clustering Coefficient:1.0
No.4: Node:中山大学, Clustering Coefficient:1.0
No.5: Node:致公党, Clustering Coefficient:1.0
No.6: Node:九三学社, Clustering Coefficient:1.0
No.7: Node:中国银联, Clustering Coefficient:1.0
No.8: Node:微软, Clustering Coefficient:1.0
No.9: Node:北京大兴, Clustering Coefficient:1.0
No.10: Node:中国队, Clustering Coefficient:1.0
```

图 8　聚集系数最大的十个机构

以及最终整个实体共现网络中聚集系数最大的十个实体, 如图

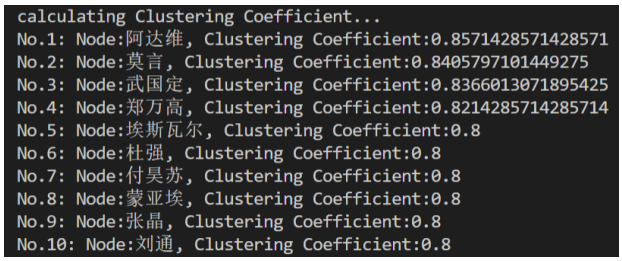

图 9 聚集系数最大的十个实体

注意到聚集系数为 1.0 的人物和机构并非之前 PageRank 或者中介中心性排名靠前的人物和机构, 这是因为那些重要性较高的实体节点往往拥有较多的入边和出边, 相应的其邻居个数的组合 $C_{d_v}^2$ 也会很大, 即使有一对之间不是朋友, 也会导致聚集系数小于 1, 而那些处于较小的强连通分量中的节点反而会由于其邻居很少, 取得 1.0 的聚集系数。

进一步, 整个实体共现网络中排行前 10 的实体也并不是出现在单独的人物网络和机构网络中的实体, 造成这个现象的原因和之前的类似, 即出现在单独的两个网络中的聚集系数为 1.0 的实体同时与一些机构 /人物相连, 然而这些额外的机构/人物在最终的实体共现网络中没有和节点之前的邻居相连, 导致聚集系数变小。

4 结语

通过这次实验, 我不仅从现实生活中的语料里提取了实体共现网络, 而且将课堂上学习的网络分析方法, 理论应用到了实践之中, 并验证了一系列理论的正确性, 同时根据各项计算结果在实体共现网络上阐明了对应的意义, 这让我将现实生活和数学建模协调在了一起, 并且能够通过实验来分析模型的各项特征, 我认为这是社会与科学结合的最佳方法；此外, 我也自学了 Neo4j 工具并最终能够较为熟练地使用, 这将对我之后的很多工作有帮助。**致谢** 感谢 Neo4j。

Peitian Zhang, Undergraduate. His/her research interests include nothing.

Background

This research is related to Graph Analysis, which includes shortest path calculation, community detection, node betweenness centrality calculation and node clustering coefficient calculation over an self-constructed entity-cooccurance network.

All of the above problems have been solved and the algorithm is mature. The main contribution of this work is personal. It helps me to understand the magic of social science and promising future about devising a model to express the phenomenon in real life.

This work belongs to no projects.