政府新闻社交网络挖掘

2018202177 官佳薇

摘 要 社会是一个由多种多样的关系构成的巨大网络，而人际关系是人类社会中最复杂且有趣的关系之一。对社会网络中人际关系的研究，可以揭示关系的结构，解释一定的社会现象，发掘人际交往间的静态和动态性变化。本文利用从政府新闻网 gov.cn 获取的新闻数据，对其中的人物关系进行了多角度的挖掘。首先使用结巴分词提取人物、地区、机构三种实体，在此基础上进行了数据清洗等预处理，按共现关系构建了新闻人物关系网络，其中认为两个人共现篇数越多，关系越强。本文构建的网络共 25725 个节点，414226 条无向边，110 个连通分量，其中最大连通分量包含 25424 个节点，是一个较为紧密的关系网络。使用 PageRank 算法计算了节点影响力大小，得到以主席为核心的中央人物影响力最大，大多数人的影响力较弱，PageRank 分值服从长尾分布。为了进一步分析网络结构，本文求解了网络各节点的局部聚集系数、中介中心性，并利用 Louvain 算法对网络整体进行社区挖掘。实验结果表明，该新闻网络有超过 60%的节点聚集系数为 1，且网络整体可划分为有限的几个较大社区，且最大社区包含 7015 个节点，说明该网络具有极高的聚集性，人物关系比较紧凑。

关键词 **连通分量；PageRank；聚集系数；社区挖掘；中介中心性；Louvain 算法；图；Neo4j**

Government News Social Network Mining

Jiawei Guan

Abstract Society is a huge network of diverse relationships, and interpersonal relationships are one of the most complex and interesting relationships in human society. The study of interpersonal relationships can reveal the structure of relationships, explain certain social phenomena, and discover static and dynamic changes in interpersonal communication. This article uses the news data obtained from gov.cn to explore the relationship among the characters from multiple angles. First, this paper use jieba to extract three entities: person, region, and organization. On this basis, data cleaning and other preprocessing are performed, and a network of news figures is constructed according to the co-occurrence relationship. It is believed that the more the two people co-occur, the stronger the relationship is. The network constructed in this paper has 25,725 nodes, 414,226 undirected edges, and 110 connected components, of which the largest connected component contains 25,424 nodes, which is a relatively tightly-connected network. In this paper, the PageRank algorithm is used to calculate the influence of nodes, and the central figure with the chairman as the core has the greatest influence, and the PageRank score obeys the long-tail distribution. In order to further analyze the network structure, this paper solves the clustering coefficient and betweenness centrality of each node, and uses Louvain algorithm to conduct community mining on the entire network. The experimental results show that more than 60% of the nodes have a clustering coefficient of 1, and the network as a whole can be divided into a limited number of larger communities, and the largest community contains 7015 nodes, indicating that the network has extremely high aggregation and the relationship between the characters is relatively tight.

Key words Connected components; PageRank; Clustering coefficient; Community detection; Betweenness centrality; Louvain Algorithm; Graph; Neo4j

1 数据处理

数据文件 gov_news.txt 包含从 gov.cn 上获取的 29700 条重要新闻，各条新闻由网址、日期、来源、标题和新闻内容 5 个部分组成。为分析其中包含的社交网络关系，需对数据进行预处理，提取其中的人物、地区和机构三种实体，从而构建知识图谱。

1.1 命名实体识别

使用结巴分词，对新闻标题和正文内容进行分词处理，并抽取其中包含的人名、机构名和地名，并对各实体计数统计。为避免新闻内部实体名出现次数对实体权重的干扰影响，多次出现在同一篇新闻中的同一实体记为出现一次，即统计出现各实体的新闻篇数。得到人名 28013 个、机构名 2459 个、地名 10098 个。

然而，结巴分词的结果中包含较多“垃圾”数据，比如，人名中包含许多四字词语和成语、单字姓氏、专有名词等，如“双循环”、“智能化”、“平稳过渡”。地名中包含许多重复项，例如“山东”和“山东省”。故需要进一步对数据进行人工筛选，并简化问题，剔除单字人名。得到人名 25725 个、机构名 2452 个、地名 9909 个。

图 1.1 人名词云图

统计各实体出现新闻篇目数，作为实体重要程度的一种度量。

(1) 人名词频统计

在所有人物中，习近平出现次数最多，高达 5817 次，其次李克强 4382 次，王毅 1645 次，其余人均在 700 次以下，最少出现 1 次。下图为词频 Top 12 的统计柱状图。

图 1.2 人物词频统计

(2) 机构名词频统计

在所有机构中，新华社出现次数最多，高达 17581 次，遥遥领先于国家政府部门机关，与政府新闻事实相悖。分析其原因，在 gov_news.txt 数据中，多条新闻来源为新华社，故在数据处理中删除来源信息，使新华社频度降低至真实数值。在最终统计数据中，国务院占主导地位，出现共 7349 次，其次为财政部 1477 次，联合国 1172 次。下图为 Top 10 机构图。

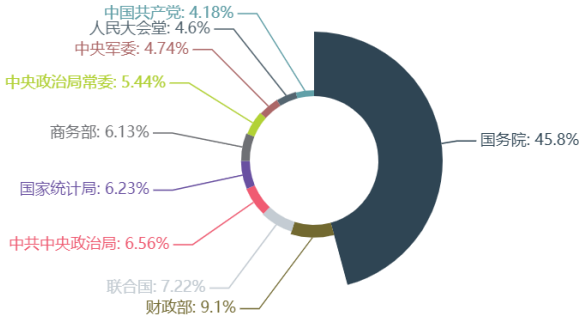

图 1.3 机构词频统计

(3) 地名词频统计

在所有地点中，中国出现的频度最高，高达 13573 次，其次为北京 10404 次，河北 2487 次，上海 2287 次。下图为地名 Top 12 统计图

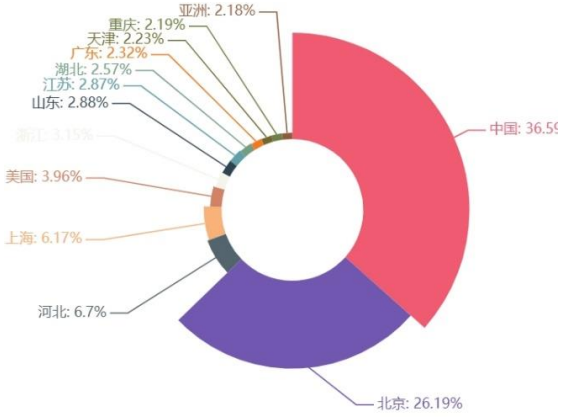

图 1.4 地名词频统计

综上可见，人名集中在以习近平为核心的党中央，机构名集中在政府主要执政部门，地名中“中国”一词占据主导地位。与来自政府重要新闻的新闻数据背景相符合。

1.2 构建社交网络图

利用提取出的实体名进行社交网络的构建。若两个人出现在同一篇新闻中，则假设这两个人有联系，其联系的强弱通过共同出现的文章数目来表示。下图为网络中部分节点的关系图，其中节点大小暗示节点出度。

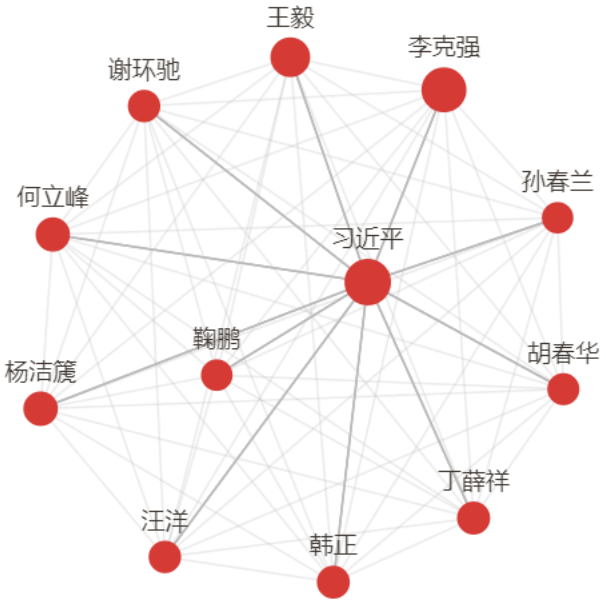

图 2.1 部分节点的社交网络图

1.3 图的统计

借助 networkx 统计图中基本信息：

表 1.1 图的统计

节点数	边数	连通分支数	最大连通分量大小
25725	414226	110	25424

2 图的验证：强弱关系

在新闻网络关系的构建中，节点间的关系强度是由其二者共现的文章篇数决定的，共现篇数越多，代表二者关系越强。统计各节点对间共现频率，验证新闻网络与真实网络间的差异及正确性。

观察以外交部长王毅为核心的关系网络，其显著的强关系邻居包括习近平、何立峰、杨洁篪等国家重要人物，且其中与习近平关系最强。

观察以美国总统特朗普为核心的关系网络，其显著的强关系邻居为主席习近平，且除与总理李克强、外交部长王毅有较强关系以外，与其具有强关系的节点包括德国总理默克尔等国外领导人。

由于数据来源为政府新闻，故出现上述结果并不意外。数据新闻以国内领导人和事件为主，因此在以特朗普为例的外国领导人的网络关系中，强关系节点以国内人物主导。虽与真实情况有所偏差，但不妨碍我们在中国大背景下研究社交网络人物关系。

下图展示了王毅和特朗普的强关系 Top 10 节点。

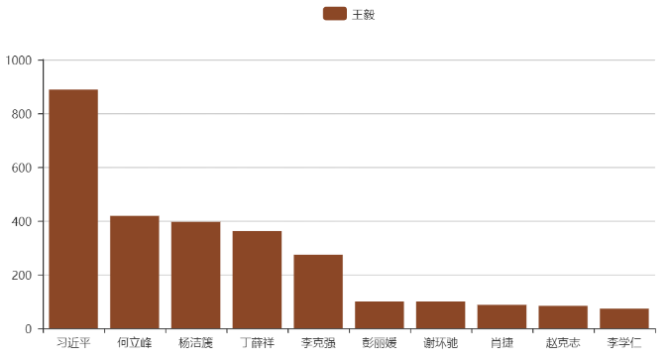

图 2.2 外交部长王毅的 Top 10 强关系

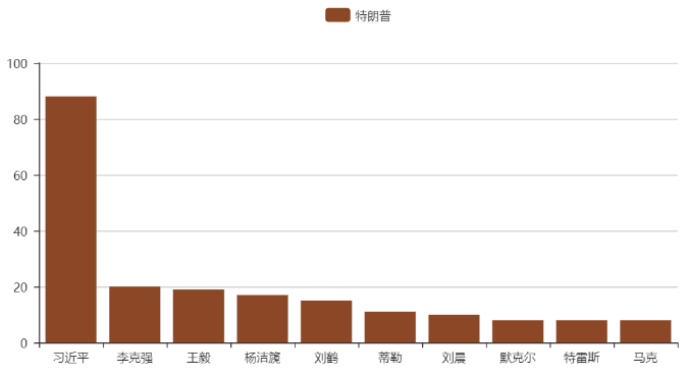

图 2.3 美国总统特朗普的 Top 10 强关系

3 影响力计算

政治家利用影响力赢得选举，商人利用社会网络中节点的影响力将商品推销到整个社会，社会舆论的引导和创新理论的传播都可以借助社会网络上具有高影响力的个体用户。社会网络的出现为定义和研究节点影响力提供了定量基础。影响力可以表

达为一个个体的特性，也可以表达为个体之间的作用形式，因此影响力具有全局和局部范围。节点的全局影响力越大，节点的信息、行为在整个社会网络中的传播控制能力越强，社会网络中一小部分最具影响力的节点能够控制整个社会网络中大部分的传播，而一个节点对另一个节点的影响力则属于局部影响力，节点对另一个节点的影响力越大，后者在社会网络中就越会追随和模仿前者的行为。

社会网络的拓扑结构、用户交互行为、用户内容构成了社会网络的 3 个要素[1]。拓扑结构能够从宏观层面上刻画节点的影响力，用拓扑结构来度量节点的影响力是一种常见的做法，然而网络拓扑结构中的连边无法描述节点间的复杂交互关系。

3.1 基于全局属性的度量：中介中心性

基于节点全局属性的节点影响力度量指标主要考察节点所在网络的全局网络信息，这些指标能够较好地反映节点的拓扑特性，但时间复杂度较高。

中介中心性（betweenness centrality）定义为网络中两个节点之间的最短路径经过当前节点的次数，结束中心性描述的是社会网络中传播时经过该节点的频率。该指标值越大，表示在网络拓扑中该节点越繁忙。若移除介数大的节点，则会造成网络拥堵，不利于信息传播。

实验中使用 Neo4j 实现中介中心性的计算，计算过程主要包括两部分：

(1) 使用宽度优先搜索（BFS）获取各对节点之间的最短路径。

(2) 统计各节点出现在最短路径中的次数。

由于数据中关键人物之间联系紧密，存在边数较多，故中介中心性取值较大，且最大值和最小值差距甚远，故对中介中心性取对数处理。总体人物中介中心性变化趋势如下图所示：

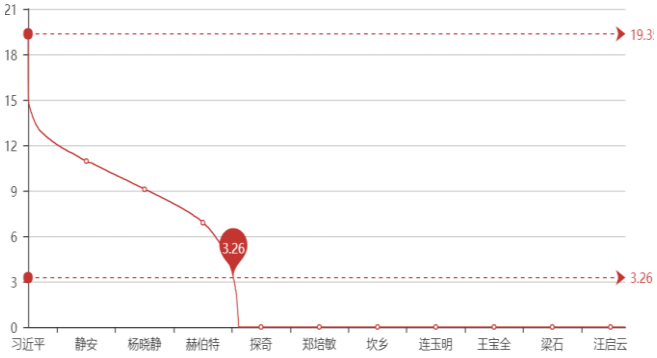

图 3.1 人物中介中心性

由上图可以观察到，绝大多数节点的中介中心性分值为 0，它们不处于任何节点的最短路径上。另外值得注意的是，因中介中心性的计算依赖于各节点处于其他节点对最短路径上的次数，节点间差异可能非常明显，故而趋势图中出现明显的断崖式下降，

展示部分关键节点中介中心性分值：

表 3.1 部分人物中介中心性分值

人物	中介中心性 (Betweenness Centrality)
习近平	253852779.60
李克强	66727447.19
王毅	10142901.86
徐昱	7694372.43
牟宇	6015568.02
杨世尧	4153955.22
赵文君	3918640.25
汪洋	3843054.51

3.2 基于随机游走的度量：PageRank

3.2.1 PageRank 算法基本思想

PageRank 算法是 Google 搜索引擎用来对搜索结果进行排序的计算方法，其基本思想来自于分析文献引文的重要性，一篇文献的重要性取决于它被引用的数量以及引文本身的质量和重要性。

从网络的观点看，PageRank 算法用于分析网络中各个节点的重要性。节点 A 的重要性来自于存在边指向 A 的其他节点的数量和这些节点本身的重要性大小。如果存在 B 指向 A 的边，那么 A 的重要性有一部分来自于 B，同样 B 将自身重要性分配给其所指向的各个顶点，说明 B 认为 A 有连接价值。

PageRank 算法前期大多用于分析有向网络图中节点的重要性，如果将无向图中存在的边看作两个节点进行互相“投票”，那么 PageRank 算法同样适用于分析无向网络图中节点的重要性。

本实验构建的网络为新闻人物关系网络，其中节点之间的联系及影响与引文、网页的联系相似，故可以使用 PageRank 算法计算节点影响力。

3.2.2 PageRank 计算结果

实验中借助 Neo4j 计算各节点 PageRank 分值，下图展示了全部节点的分值变化趋势，其中同样对分值进行对数处理：

图 3.2 人物 PageRank 分值

由于 PageRank 的计算方式考虑网络中各边之间的影响，迭代多轮直至收敛，故不会出现如前中介中心性的断崖式下降趋势，整体变化较为平滑。另外可以观察到，网络中影响力较大的重要节点 PageRank 分值相差较大，而绝大多数节点影响力不高，相差较小，使图像出现长尾现象。

下表列出了部分高 PageRank 值节点，可见 PageRank 计算得到的结果与人们的认知更为相符，高分值的节点均为影响力较大的核心人物。

表 3.2 部分人物 PageRank 分值

人物	PageRank
习近平	251.83
李克强	124.97
王毅	49.97
谢环驰	33.23
何立峰	31.20
杨洁篪	29.04

4 聚集系数

社会网络中，联系紧密的多个好友形成社团的现象在社会网络中很常见。局部聚集系数（local clustering coefficient）用于衡量节点的邻居结点之间联系的紧密程度。聚集系数等于节点邻居节点之间连边的数量与邻居节点之间可以连边的最大数量之比。无向图聚集系数计算公式如下所示：

$$C(v_i)=\frac{2|\{e_{jk}:v_j,v_k\in N_{v_i},e_{jk}\in E\}|}{k_i(k_i-1)},$$

其中 k_i 为节点 v_i 指向其他节点的连边数量与其他节点指向 vi 的连边数量的和。

借助 Networkx 计算新闻网络各节点聚集系数，并可视化全部节点聚集系数趋势如下图。可见绝大多数节点的聚集系数为 1，即大多数节点的朋友之间彼此也为朋友，网络聚集程度较高。

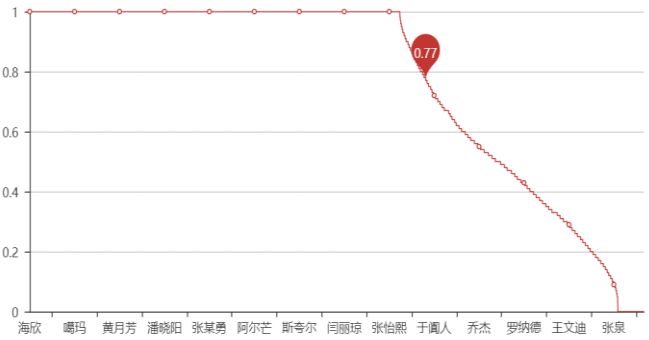

图 4.1 人物聚集系数

进一步分析各阶段分数占比，统计得到聚集系数为 1 的节点数占总体的 60%，占主导地位，可知该新闻网络具有极强的聚集性，节点之间联系紧密。

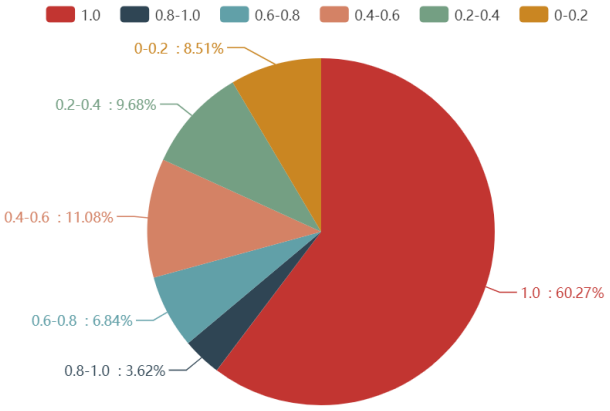

图 4.2 聚集系数比例

5 社区挖掘

网络中的社区结构是网络的重要特性之一，社区结构反映了网络中各个节点间的关系及网络的整体结构。同一个社区中的节点，他们之间有较大的相似性。社区内部节点间联系非常紧密，而社区之间的联系相对比较稀疏。

社区结构揭示了网络隐藏的特性。复杂网络中的社区检测给人们提供信息，以便更好的了解网络内部成员及他们之间的关系。

5.1 Louvain 算法

Louvain 算法[2]是基于模块度的社区发现算法。该算法在效率和效果上都表现较好，并且能够发现

层次性的社区结构，其优化目标是最大化整个社区网络的模块度。模块度（也称 Q 值）用于度量网络中各社区内部联系的强度，在一个高 Q 值的网络中，各个社区内部的链路较为密集而社区间的链路则十分稀疏。

Louvain 算法步骤如下：

(1) 将图中的每个节点看成一个独立的社区，社区的数目与节点个数相同；

(2) 对每个节点 i，依次尝试把节点 i 分配到其每个邻居节点所在的社区，计算分配前与分配后的模块度变化ΔQ，并记录ΔQ 最大的那个邻居节点，如果 max ΔQ >0，则把节点 i 分配到ΔQ 最大的那个邻居节点所在的社区，否则保持不变；

(3) 重复步骤 2，直到 Q 值不再发生变化，即将一个节点转移到网络内的另一个相邻社区，将不能带来ΔQ 的提升，此时当前网络内所有节点都不再移动;

(4) 社区归并，这一步也可看做对原图的压缩，将前几步得到的各个社区作为新图的节点，同时将原社区内部所有节点对的边权重之和作为新的权重赋予新图的各条边。

Louvain 算法有如下优势：

(1) 算法得到的社区结构是分层的，每一轮计算完成后得到的新图都是对一个大社区内若干细分社区发现的结果，这样的分层结构是每个网络的自然属性，能深入了解某个社区内部结构和形成机制；

(2) 算法易于实现，并且计算过程全程无监督，即最终结果完全依赖于算法聚类，并不需要人为提前预设分类；

(3) 算法的性能较好，在进行一些经典社区分类算法的对比中，Louvain 算法对图的大小几乎没有上限要求，并且能在迭代几轮后快速收敛。这为处理拥有百万级别以上节点的移动通信网络甚至上亿节点的大型社交网络的社区发现提供了可能。

5.2 实验结果

Louvain 算法将网络划分为 132 个社区，其中 15 个社区成员数在 30~7015 之间，其余 117 个社区成员数均小于 20。出现极小社区较多的原因，是因为结巴分词结果有较多“垃圾”数据，人为对数据进行清理容易产生遗漏，使得出现较多相对孤立的节点。将小于 20 的社区合并，各社区占比展示如图：

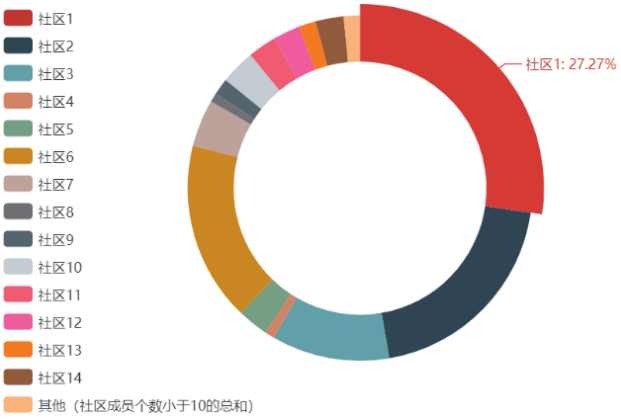

图 5.1 社区比例

若不考虑噪声数据造成的极小社区，该新闻网络可以划分成有限的几个主体社区，且各社区节点数较多，结构较清晰。

参考文献

[1]韩忠明,陈炎,刘雯,原碧鸿,李梦琪,段大高.社会网络节点影响力分析研究[J].软件学报,2017,28(01):84-104.

[2] Meo P D, Ferrara E, Fiumara G, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 2008(10):155-168.

[3]Barber M J, Clark J W. Detecting network communities by propagating labels under constraints[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(2 Pt 2):026129.