新闻数据中的社会网络挖掘

郭丹琪

2018202067

摘　要：给定一批新闻数据，可以从中挖掘出新闻中实体的社交网络关系。本文基于中国政府网上的 29699 条重要新闻，完成了社会网络的构建及网络的基本分析，包括网络图的验证、基本信息的统计和结点的影响力计算。在此基础上，还对小世界理论进行了验证，进行了结点中介中心性和聚集系数的计算。通过计算和分析可以得知中国政府网的重要新闻中出现的人物可以构成一个小世界网络，任意两个人平均在三篇新闻中就可以产生联系。新闻中影响力大、出现次数多的重要人物在网络中处于中心聚簇和超大连通分量的中心位置，同时也是网络中影响力大、中介中心性高的结点。而新闻中出现次数少的人在网络中处于边缘位置，依靠热门结点与其他结点产生联系。

关键词：社交网络；连通分量；PageRank 算法；小世界理论；聚集系数；中介中心性

Social Network Mining of News Data

Danqi Guo

Abstract: Given a batch of news data, we can dig out the entity social network relationship in the news. Based on 29,699 important news on the Chinese Government Website, this paper completed the construction of social network and the basic analysis of the network, including the verification of network diagram, statistics of basic information and influence calculation of nodes. On this basis, the small-world theory is verified, and the centrality and clustering coefficient of nodes are calculated. Through calculation and analysis, it can be known that the characters appearing in important news on the Chinese Government Website can form a small-world network, and any two people can be connected in an average of three news articles. People with great influence and frequent occurrence in news are at the center of central cluster and super-connected component in the network, and they are also the nodes with great influence and high betweenness centrality in the network. Those who appear less frequently in news are on the edge of the network and rely on hot nodes to make connections with other nodes.

Key words: Social network; Connected components; PageRank; Small-world theory; Clustering coefficient; Betweenness centrality

1 数据预处理及基本信息统计

首先对新闻的标题和内容进行分词和词性提取处理，从中提取出人名、地名、机构名这几类实体，用实体及实体在新闻中体现出的关系建立社交网络。

使用结巴分词进行分词和词性提取。发现分词结果和词性标注不够精确，存在部分问题。第一个问题就是有些词的词性在不同的语境下是不同的，比如“许可”这个词，可以作为人名，但是在政府新闻中大多是以普通名词的形式出现，而分词和词性标注的结果将其统一看作人名。第二个问题是分词可能会把一个人的人名截断，而多个不同的人的名字在截取后可能是相同的，导致后续构建网络的时候多个结点的特点会被叠加到一个点上。例如分词结果中获得的“邹世”这个词，虽然看起来是一个人名，但是经过搜索发现只能搜索到“邹世春”、“邹世龙”等人。

所以在这种情况下，只能在后续处理过程中进行人工筛选和判断，并且在分析时考虑分词结果不精确的因素。

1.1 热门人物和机构统计

对新闻中出现的热门实体进行统计。

因为同一条新闻讲的是同一件事，涉及到的实体可能多次出现，而这不能反映实体的热门程度。所以将热门的判断标准设定为实体在不同新闻中的出现次数。在越多的新闻中出现，就说明该实体越热门。得到的结果如表一所示。

表 1　热门人物和机构统计（前十名）

热门实体	频率
新华社	17581
国务院	7439
习近平	5817
李克强	4352
党中央	2943
王毅	1645
中国政府	1479
财政部	1477
联合国	1172
中共中央	1121

从热门实体的统计结果就可以看出数据的正确性。因为新闻数据的来源是中国政府网，这些新闻自然与国家、政府相关，涉及到的人物和机构多为国家领导人和政府机构，而热门实体的统计结果就体现了这一点。其中新华社是国家通讯社，是中国政府网重要的新闻来源， 因此出现次数最多，最为热门。剩下的热门实体里的国务院、党中央等均为政府新闻里的热门实体，而作为国家领导人的习近平、李克强和外交部发言人王毅自然也是政府新闻里的热门人物。

1.2 构建社交网络图

通过获取到的人物实体及之间的关系构建社交网络图。

对两个实体间是否存在关系的判定是通过判断两个实体是否出现在了同一新闻中，而该关系并无方向性，因此将得到的数据构建为无向图。将提取到的人物实体作为图中的结点，边的权重为两个实体出现在同一篇新闻中的次数。

2 图的基础分析

2.1 信息统计

对图中的结点信息和连通分量信息进行统计。统计数据如表 2 所示。

表 2　图的统计

结点数	27946
边数	892795
连通分量数	21
最大连通分量	27895

由统计数据可知，所有新闻中一共出现了 27946 个不同的人，而图中最大连通分量的大小达到了 27895，说明图中几乎所有结点都在这个最大连通分量中。即在构建的网络中，这些人物之间基本都是互相可达的，都存在直接或间接的关系。这主要是因为数据来源是重要的政府新闻，其中涉及的人物都是所处领域中的重要人物，彼此之间很可能因某些重要事件而产生联系，因此出现在同一篇新闻或者同一类新闻中，所以就会有直接或间接的联系。

画出构建的社交网络图，如图 1。可以很明显

的看出，大部分结点都聚集于中心区域，在中心区域形成了一个超大连通分量。同时也存在部分没有在中心区域的结点，这些点数量较少，较分散，之间的联系也较少。这可能是因为有些人只在某个特定新闻中出现，相关的联系较少，也有可能是因分词错误而出现的结果。

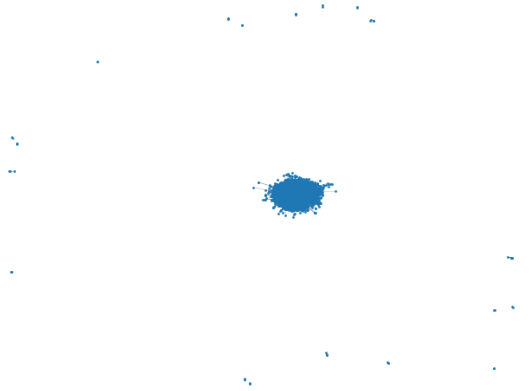

图 1 完整的网络结构

画出图中一个极大连通分量，如图 2。可以看出其中大部分结点都联系紧密，之间存在大量的边。而很多外围的结点只存在一条边，只是与某个热门区域的结点联系紧密，可以通过几个热门的结点到达大部分的结点。猜测是因为有很多人其实不常在新闻中出现，但是出现过的那篇新闻中有几个重要人物或者热门人物。

对该猜测进行验证，统计只出现过一次的人的数目，发现共有 16509 人，占到总人数的 59%，说明有超过一半的人实际只在新闻中出现过一次。而由图 1 可以看出整个网络结构中的离群点只占少数。这也就侧面证实了上述猜测。

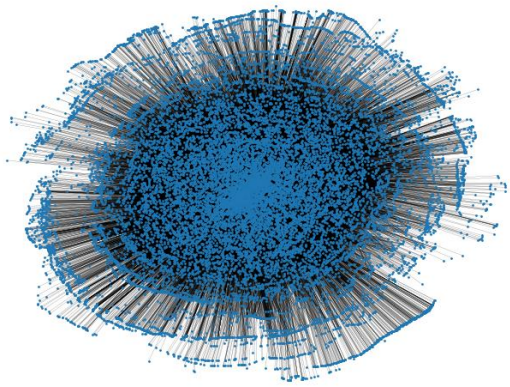

图 2：极大连通分量的网络结构

2.2 图的验证

在该部分需要提供某个人在网络图中关系最强的十个人，即找到与这个人在新闻中共同出现次数最多的十个人。

在热门实体统计部分已经得到习近平、李克强、王毅是所有新闻中最为热门的三个人物，在这里就找到与他们关系最强的十个人物来进行验证。得到的结果在表 3、表 4、表 5 中体现。

表 3 与习近平联系最紧密的 10 个邻居

人物	频率
李克强	1174
王毅	888
杨洁篪	521
何立峰	504
丁薛祥	481
韩正	396
谢环驰	386
鞠鹏	355
胡春华	346
孙春兰	311

表 4 与李克强联系最紧密的 10 个邻居

人物	频率
习近平	1174
肖捷	274
王毅	274
韩正	229
王沪宁	222
何立峰	200
汪洋	194
王勇	136
赵乐际	133
胡春华	122

表 5 与王毅联系最紧密的 10 个邻居

人物	频率
习近平	888
何立峰	418
杨洁篪	396
丁薛祥	362
李克强	274

彭丽媛	100
谢环驰	100
肖捷	87
赵克志	83
李学仁	73

不难看出，这些关系最强的邻居其实也都是相关人物，并且由于习近平、李克强、王毅三人本身联系紧密，有些人在三者联系最紧密的邻居中均有出现。例如何立峰，因为是国家发展和改革委员会主任、党组书记，所以自然与习近平、李克强联系紧密，常在同一篇新闻中出现。

这里关系最强其实就是在两个人同时出现在一条新闻里的次数最多，而这其实也意味着两个人的重要程度相当，且所处的领域、相关的事件相似。

2.3 影响力计算

利用网络图中各结点的 PageRank 分数作为每个人的影响力大小。获得的影响力最大的 20 个人如表 6 所示。

表 6　PageRank 排名前 20 的人

人名	分数
习近平	0.01586952674702383
李克强	0.008042386411843063
王毅	0.0036934457925281916
何立峰	0.00207679617005792
施策	0.0020529745228062917
杨洁篪	0.0019495926793227337
丁薛祥	0.001946910088071923
韩正	0.0018066272326057401
许可	0.0017390948835948362
谢环驰	0.0017013451977385976
汪洋	0.001609755483493026
王沪宁	0.0014972703525777849
鞠鹏	0.0014907292674138537
孙春兰	0.001125173297740148
刘鹤	0.001072517725724767
肖捷	0.001037106271786618
赵乐际	0.0009966866063858824
胡春华	0.0009824694055506125
王晔	0.0009605794974752775
李学仁	0.000896523374681558

不难发现，影响力最大的前 20 人基本也是出现次数多的人。

这是因为 PageRank 体现的是影响力的大小，而这里是根据新闻构建的网络图，影响力大意味着相关的事件和联系到的人多，在新闻中出现的次数自然就多。

并且根据无向带权图中 PageRank 计算的迭代方程

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

其中$weight(p_j)$是边(p_i, p_j)的权重，$degree(p_j)$是结点p_j的度数。

可以看出，如果一个结点的 PageRank 值高，则其边的权重和邻居的 PageRank 值也较高。在该网络中就反映出这个人与相关人物在新闻中出现的次数都较多，因此也正是重要的、影响力大的人。

3 自选分析

本部分内容为自选分析内容，验证了小世界理论，实现了对结点的中心性计算和聚集系数计算，并由这些数据对社交网络图进行了进一步的分析。

3.1 小世界理论验证

由 2.1 中对图结构的分析，可以得知该网络中有很多结点之间联系紧密，用这些结点间的路径不足以验证小世界理论。需要找到新闻中最不热门的人，也就是只在一篇新闻中出现过的两个人来进行路径查找。两个分别只在一篇新闻中出现过的人存在直接联系，也就是出现在同一篇新闻中的概率很低，对其进行路径搜索更能体现小世界理论。

选取两个的分别只出现过一次的人：曾庆军、李陆勇，查找这两个人之间的最短路径，得到的结果如表 7。

表 7　前 10 条最优路径

路径	长度
['曾庆军', '黄鑫', '李', '梁旭', '李陆勇']	4
['曾庆军', '陈肇雄', '李', '梁旭', '李陆勇']	4
['曾庆军', '黄鑫', '智慧', '梁旭', '李陆勇']	4
['曾庆军', '陈肇雄', '智慧', '梁旭', '李陆勇']	4
['曾庆军', '黄鑫', '智能化', '梁旭', '李陆勇']	4
['曾庆军', '陈肇雄', '智能化', '梁旭', '李陆勇']	4

['曾庆军', '黄鑫', '习近平', '梁旭', '李陆勇']	4
['曾庆军', '陈肇雄', '习近平', '梁旭', '李陆勇']	4
['曾庆军', '黄鑫', '明显增强', '梁旭', '李陆勇']	4
['曾庆军', '黄鑫', '惠民', '梁旭', '李陆勇']	4

根据这个结果我们不难看出，即使是只在一篇新闻中出现过的不热门人物，他们之间也存在大量短路径。这很好地验证了小世界理论。

而根据上述在 2.1 中的分析，这主要是因为他们出现过的新闻中可能有同样的重要人物或者热门人物，他们就通过这样的人物产生联系。

同时计算出整个图的平均路径长度，得到的结果约为 2.413，这说明这些新闻数据中的任意两个人平均在三篇新闻内就可以建立联系。这进一步验证了小世界理论。

3.2 中介中心性计算

计算网络中各结点的中介中心性，排名前十的结点如表 8 所示。

表 8　中介中心性排名前十的人

人物	分数
习近平	0.12206774414153812
李克强	0.03712548995058477
高峰	0.011769393704056563
王毅	0.007976954789840132
许可	0.007653265709069514
施策	0.007614878045370535
徐昱	0.005369938191865533
牟宇	0.004532334649142658
李晓果	0.003530454774536484
杨世尧	0.003005556735117341

不难发现在中介中心性较高的人在影响力分数中排名也靠前。中介中心性指的是一个结点担任其它两个结点之间最短路的桥梁的次数。一个结点充当“中介”的次数越高，它的中介中心度就越大。所以这意味着这个网络是一个具有中心聚簇的网络，网络中的结点需要通过这个中心聚簇与其他结点形成联系，这个中心聚簇也就是网络中最大的连通分量。这一点与 2.1 中所作的网络图的特点相符。

这也进一步验证了在 2.1 和 3.1 中的分析，大量结点通过热门人物产生联系，热门人物担任了大量中介的人任务。

3.3 聚集系数计算

计算网络图中各个的结点的聚集系数，得到聚集系数最高的十个人物如表 9 所示。

表 9　聚集系数排名前十的人物

人物	聚集系数
龚静毅	1.0
徐仰辉	1.0
陈星杰	1.0
邹世	1.0
闻效仪	1.0
巩某	1.0
冯建某	1.0
叶进国	1.0
施汉生	1.0
郭晓林	1.0

发现排名前十的人的聚集系数均为 1，猜测还存在大量聚集系数为 1 的结点，因此对聚集系数为 1 的结点数目进行了计算。得到该网络图中共有 16252 个结点的聚集系数为 1，占到了总结点数的 58%。一个结点的聚集系数为 1 说明该结点的所有邻居相互之间也有直接联系。这侧面验证了三元闭包原理，也反映了这些结点间联系紧密，有很多人不仅在新闻中出现次数多，而且出现的新闻中还包含了很多其他出现次数多的人物。

4 结论

基于上述分析，我们可以得知中国政府网的重要新闻中出现的人物之间可以构成一个小世界网络，任意两个人间都存在短路径，即在新闻中存在联系。该网络中影响力大的结点与中介中心性高的结点出现大量重合，也正好对应了新闻中出现次数多、重要性高的热门人物。

该网络有一个超大连通分量，几乎连接了网络中的所有结点。该连通分量同时也是网络中的中心聚簇，重要结点在该中心聚簇中占据中心位置，新闻中出现次数少的、重要度较低的结点处在边缘位置，主要通过这些重要人物与其他人物产生联系。