基于Gephi的航运招聘信息可视化分析

王扬[1]，田野[1]，李铁山[1]，陈俊龙[1,2]，彭东成[1]，周义华[3]
1. 大连海事大学航海学院，辽宁 大连 116026； 2. 澳门大学科技学院，澳门 999078；
3. 大连湾辽渔集团港务分公司，辽宁 大连 116026

摘要
将基于Gephi的社会网络可视化分析方法引入船员招聘数据的关联性分析领域中，通过构造船员招聘数据的属性共现网络，分析了船员招聘信息中各属性之间的关系。通过交互式分析，得到了职位与航线、航线与证书等属性之间的关系，探索了船员数据中各个属性之间的关联程度以及航运招聘信息网络的核心属性，以期为相关部门以及广大船员了解船员市场的变化提供有益的参考。

关键词
航运招聘信息；可视化分析；Gephi；关联性

中图分类号：TP393 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018032

Visualization analysis of shipping recruitment information based on Gephi

WANG Yang[1], TIAN Ye[1], LI Tieshan[1], CHEN C.L.Philip[1, 2], PENG Dongcheng[1], ZHOU Yihua[3]
1. Navigation College, Dalian Maritime University, Dalian 116026, China
2. Faculty of Science and Technology, University of Macau, Macau 999078, China
3. Dalian Bay Liao Fishery Group Port Branch, Dalian 116026, China

Abstract
The Gephi-based social network visualization analysis method was introduced into the field of association analysis of crew recruitment data. By constructing the attribute co-occurrence network of crew recruitment data, the relationship between various attributes in the crew recruitment information was analyzed. Through interactive analysis, the correlations such as positions and routes, routes and certificates, can be obtained. What's more, the degree of correlation between various attributes in the crew data and the core attributes of the shipping recruitment information network were explored, which in order to provide reference and basis for the relevant enterprises and crews to deal with changes in the crew market.

Key words
shipping recruitment information, visualization analysis, Gephi, correlation

1 引言

随着大数据的提出，各行各业对于数据的重视程度达到了前所未有的高度，航运业也不例外。航运招聘信息作为航运市场信息的重要组成部分，对于海事部门的资源调度以及应聘者的求职都有着非常重要的参考价值。数据可视化作为数据分析以及数据呈现的重要手段，受到各界越来越多的重视。传统的数据可视化仍然多以条形图、折线图、饼图为主，形式相对来说过于单一，而且呈现的信息仍然较为抽象，较难反映各因素各水平之间的关联性。如何将海量的船员招聘信息进行统计处理并以合理的方式呈现出来成为一个亟待解决的问题。

作为专业的航运类网站，航运在线网基本涵盖了航运产业链的各个环节，包括散杂货租船、集装箱订舱、船舶买卖、备件物料供应、船员招聘、陆上管理人才招聘、船舶管理、船舶保险等。

本文以航运在线网站上的船员招聘信息为例，借助复杂网络可视化分析软件Gephi对航运招聘信息中的主要因素进行分析以及可视化，以发现各因素之间的联系。通过直观地展现各因素各水平之间关联的强弱，揭示航运招聘信息的特点，以期为海事部门的管理以及应聘者提供直观的分析参考。

2 数据可视化准备

在数据可视化之前，首先需要获取所需数据，其次进行数据预处理，为数据可视化做好充分的准备。

2.1 数据获取

本文数据的主要来源是航运在线网上的船员招聘信息，利用Python编程工具和网络爬虫技术对其进行批量获取，并存储到本地Excel文件中，以便进行后续的可视化分析。其中，获取到的部分数据见**表1**，获取到的数据主要有编号、职位、证书、月薪、总吨、航线区域、招聘船型、招聘公司、日期共9列信息（这里只选取其中一个月的数据，以2017年4月为例进行分析，经初步的数据清洗之后一共有1 864条数据）。

2.2 数据预处理

本文使用的数据预处理主要包括数据去重、数据异常值剔除以及精简研究对象。本文使用Excel对数据进行去重处理。对于异常值的处理，由于其数量不是很多（占比不到5%），对于要分析的数据总体几乎没有影响，所以这里采用的方法是直接从总体中剔除。由于月薪一列的观测值多为“面议”以及各种非标准的数据，所以在这里不予考虑，全部删除。此外，船员编号和招聘公司信息以及招聘日期在本文中用处不大，予以删除。这样处理之后，可以将需要的信息进行集中整合，见**表2**。

本节主要通过Python爬取航运在线网上的船员招聘信息，并使用Excel对爬取到的数据做简单的数据预处理工作，为后续的数据可视化做好相应的数据准备。

3 数据可视化分析

3.1 基于Gephi的初步可视化

首先将数据加工成Gephi需要的CSV

表1 获取的部分数据

编号	职位	证书	月薪	总吨/吨	航线区域	招聘船型	招聘公司	日期
1242677	三副	甲类	面议	4万~30万	内外贸兼营	油轮	永州市杰泰	4月1日
1244978	高级机工	甲类	面议	2万~3万	东南亚	油化船	永州市杰泰	4月1日
1243996	船长	甲类	面议	2553TEU	东南亚	集装箱船	永州市杰泰	4月1日
1243983	水手长	丙一	面议	6万~10万	国内南北线	散杂货船	永州市杰泰	4月1日
1243392	船长	甲类	面议	1万~2万	中日韩	集装箱船	永州市杰泰	4月1日
1245366	实习机工	甲类	面议	4万~30万	内外	油轮	永州市杰泰	4月1日
1243993	船长	甲类	面议	2553TEU	东南亚	集装箱船	永州市杰泰	4月1日
1245365	实习水手	甲类	面议	4万~30万	内外	油轮	永州市杰泰	4月1日
1244001	船长	甲类	面议	3万~6万	环球航线	木材船	永州市杰泰	4月1日
1243695	大副	甲类	面议	3万~6万	环球航线	油化船	永州市杰泰	4月1日
1245024	船长	丙一	面议	3万~6万	内贸	集装箱船	永州市杰泰	4月1日
1244000	船长	甲类	面议	3万~6万	环球航线	木材船	永州市杰泰	4月1日
1243991	水手长	丙一	面议	6万~10万	国内南北线	散杂货船	永州市杰泰	4月1日
1243523	新证水手	甲类	面议	3万~6万	看船	散杂货船	永州市杰泰	4月1日
1245425	机工	甲类	面议	1万~2万	东南亚	散杂货船	永州市杰泰	4月1日
1245421	机工长	甲类	面议	1万~2万	东南亚	散杂货船	永州市杰泰	4月1日
1244409	二副	甲类	面议	15万~30万	环球航线	散杂货船	永州市杰泰	4月1日
1242257	水手	丙一	面议	1万~2万	国内南北线	油轮	永州市杰泰	4月1日
1242220	水手长	甲类	面议	6万~10万	东南亚	散杂货船	永州市杰泰	4月1日
1242274	水手	丙一	面议	1万~2万	国内南北线	油轮	永州市杰泰	4月1日
1242222	高级水手	甲类	面议	6万~10万	东南亚	散杂货船	永州市杰泰	4月1日
1241941	三副	甲类	面议	3万~6万	环球航线	油化船	永州市杰泰	4月1日
1242311	水手长	甲类	面议	3万~6万	远洋	油轮	永州市杰泰	4月1日
1242012	三副	甲类	面议	3万~6万	远洋	油化船	永州市杰泰	4月1日
1244408	大副	甲类	面议	15万~30万	环球航线	散杂货船	永州市杰泰	4月1日
1241822	水手	丙一	面议	3万~6万	国内南北线	散杂货船	永州市杰泰	4月1日
1242437	大副	甲类	面议	3万~6万	环球航线	油化船	永州市杰泰	4月1日
1241965	高级水手	甲类	面议	6万~10万	东南亚	散杂货船	永州市杰泰	4月1日
1242676	二副	甲类	面议	4万~30万	内外贸兼营	油轮	永州市杰泰	4月1日

格式。Gephi的图形要素为节点和边，相应地，需要导入节点数据和边数据。导入数据之后，Gephi默认的布局算法是正方形轮廓的随机布局算法[1]，如**图1**所示。节点之间显得比较拥挤和混乱，部分节点被遮挡，因此效果不是很好。为了使可视化的效果更好一些，先使用Force Atlas算法[2]，再使用改进的弹簧模型（FruchtermanReingold）算法以优化节点的布局。由于节点较多，Force Atlas算法的运行速度和效果都不太理想，所以改用Force Atlas2和FruchtermanReingold结合的方式进行布局[3]，最后稳定下来的预览图如**图2**所示。

利用Gephi进行初步可视化有助于了解数据的整体情况和大致分布。如**图2**所示，不难看出数据在几个点比较集中。由于节点较多，初步可视化的效果并不是很好，也很难从繁杂的连线中进一步解读信息。针对这些问题，需要对数据进行进一步加工。

表 2 预处理之后的部分数据集

职位	证书	总吨/吨	航线区域	招聘船型
高级机工	甲类	2万~3万	东南亚	油化船
水手长	丙一	6万~10万	国内南北线	散杂货船
船长	甲类	1万~2万	中日韩	集装箱船
船长	甲类	3万~6万	环球航线	木材船
大副	甲类	3万~6万	环球航线	油化船
船长	甲类	3万~6万	环球航线	木材船
水手长	丙一	6万~10万	国内南北线	散杂货船
机工	甲类	1万~2万	东南亚	散杂货船
机工长	甲类	1万~2万	东南亚	散杂货船
二副	甲类	15万~30万	环球航线	散杂货船
水手	丙一	1万~2万	国内南北线	油轮
水手长	甲类	6万~10万	东南亚	散杂货船
水手	丙一	1万~2万	国内南北线	油化船
高级水手	甲类	6万~10万	东南亚	油轮
三副	甲类	3万~6万	环球航线	油化船
水手长	甲类	3万~6万	远洋	散杂货船
三副	甲类	3万~6万	远洋	散杂货船
大副	甲类	15万~30万	环球航线	油化船
水手	丙一	3万~6万	国内南北线	散杂货船
大副	甲类	3万~6万	环球航线	油化船
高级水手	甲类	6万~10万	东南亚	散杂货船
水手长	甲类	3万~6万	环球航线	油轮
水手	甲类	6万~10万	东南亚	散杂货船
大副	甲类	3万~6万	远洋	油轮
水手	甲类	6万~10万	东南亚	散杂货船
水手	甲类	15万~30万	远洋	散杂货船
水手	甲类	3万~6万	环球航线	散杂货船

3.2 数据二次加工

对于初步可视化的结果，为了有一个定量的判断，这里使用统计分析的利器——R语言进行数据的进一步加工。首先将数据表导入R软件中进行必要的描述统计，结果如**图3**所示，可以看出每一个因子的主要水平以及没有列全的许多次要水平。为了消除次要水平的干扰，这里只对主要水平进行统计分析。经过第二次删减之后，数据精简到1 417个观测值。

3.3 基于Gephi的可视化分析

除了使用进一步加工的数据源，Gephi的基本操作和第4.1节中的描述几乎一致，这里不再赘述。接下来依据实际收集到的数据的特点对Gephi进行可视化方案的设计。

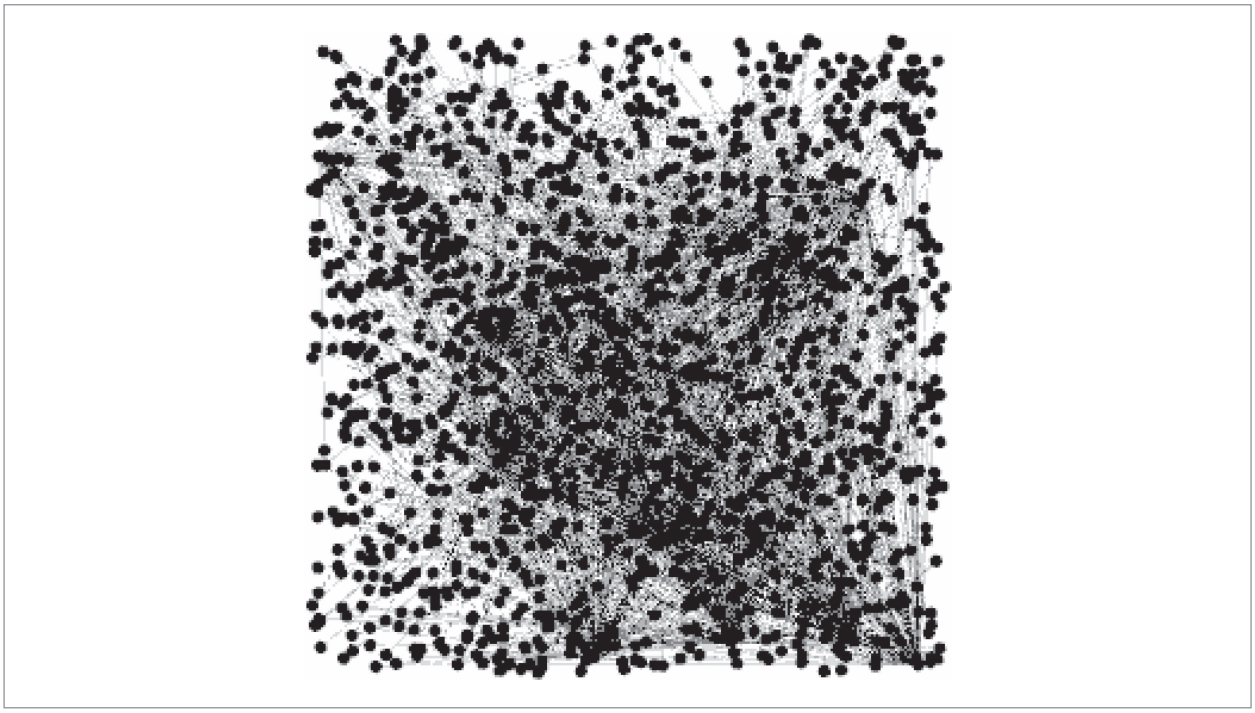

图1 正方形轮廓的随机布局概览

在权重设计阶段，因为每两个不同因素水平之间的频数最小为1，最大为534，相差较大，所以这里将频数处理为边的权重，以体现两个水平之间的紧密程度是不合理的。但是，为了在一定程度上反映它们之间联系的紧密程度，在权重的设置上采取参考频数的做法进行权重划分。将频数在1~100的权重设置为1，其比例占到了所有统计量的92.77%；将频数在101~200的权重设置为2，其比例为4.34%；将频数在201~300的权重设置为3，其比例为1.2%；将频数在301~400的权重设置为4，其比例为0.96%；将频数在401~500的权重设置为5，其比例为0.48%；将频数在501~600的权重设置为6，其比例为0.24%。当然也可根据实际的分析需求对权重的划分采取更精确更细致的算法，在这里只是做一个相对简单的划分。因为本文重点考察各水平之间的相互关系，而各因素之间的联系是相互的，如果采取有向边的

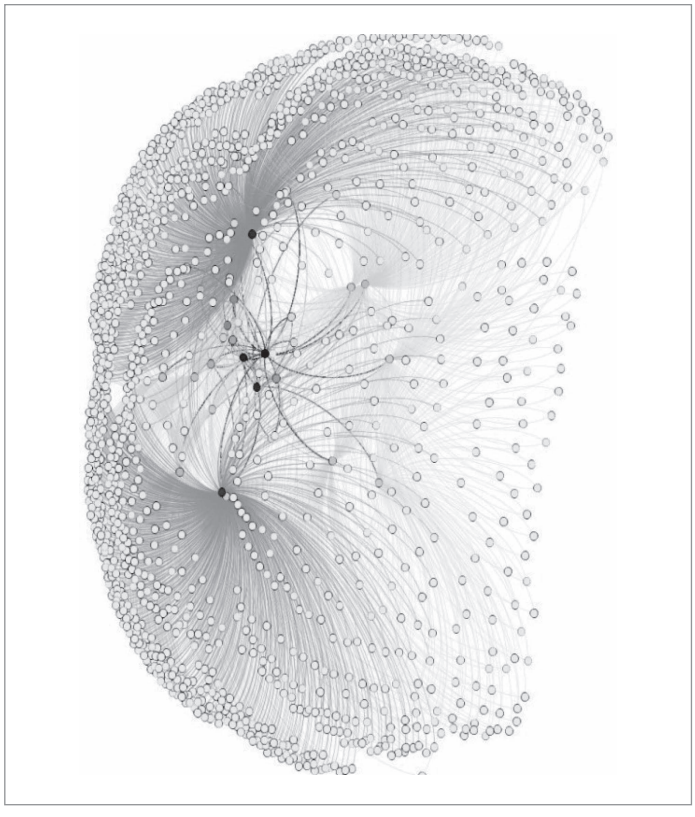

图2 数据的初步可视化预览

```
> summary (mydata,maxsum = 50)
      编号                职位              证书              总吨
 Min.   : 840820   船长：150        丙二：106        10万~15万吨：17
 1st Qu. :1245409  大副：134        丙一：610        15万~30万吨：64
 Median:1251042    大管轮：98       甲类：631        1万~2万吨：644
 Mean   :1247753   二别：115        乙一：70         2万~3万吨：207
 3rd Qu.:1255223   二管轮：130                       3万~6万吨：337
 Max.   :1259672   机工：138                         6万~10万吨：148
                   机工长：50
                   轮机长：114
                   三副：163
                   三管轮：110
                   水手：159
                   水手长：56
     航线区域      招聘船型
 东南亚：228       多用途船：26
 国内南北线：670   工程船：16
 环球航线：325     集装箱船：258
 江海航线：60      木材船：26
 其他：61          散杂货船：977
 远洋：20          油化船：36
 中日韩：53        油轮：78
```

图 3 R 软件进一步处理之后的数据统计信息

设置，就会导致不必要的重复计数，所以本文采取无向边的设置。这样做不仅可以少计算一半边的数量，大大简化构建的网络图的复杂度，同时还能保证分析的精度。为了充分显示两个水平之间具体的关联程度，将统计的两个水平同时出现的频数以边标签的形式映射到图中，而将每个水平出现的总频数以点标签的形式映射到图中。此外，结合专业知识对职位因素进一步细分为甲板部船员和轮机部船员并区别显示。

经过进一步的数据可视化准备，通过Gephi进行可视化的整体效果如**图4**所示，一共得到了36个节点以及415条无向边（数据仅限航运在线网2017年4月的招聘信息）。经过简单的统计计算，可以得到以下参数值。

- 平均度：23.056。
- 平均加权度：26.111。
- 网络直径：2。
- 图密度：0.659。
- 模块化：0.075。
- 平均聚类系数：0.608。
- 平均路径长度：1.341。

通过无向图的滤波功能，可以很方便地通过人机交互的方式查看需要考察的各因素之间的关系。

在本文的研究中，将同一变量下的不同水平（即同一变量的不同取值，如船型变量下有油船、散货船、集装箱船等取值）作为节点，节点的大小代表该水平在招聘信息中出现的次数。本研究中的边代表的是不同变量的不同水平在招聘信息中成对出现的次数，即边反映了两端节点同时出现的次数。

3.3.1 船员职位与航线因素之间的关系

接下来按照类似的方法可以考察职位因素和航线因素之间各水平的联系。首先由**图5**可以看出，节点数为19（占比52.78%），边的个数为83（占比20%），平均度为8.737，平均加权度为8.842，图密度达到0.485。通过边的占比不难发现，航线与职位因素的联系较为紧密。通过计算平均度或平均加权度也能较清晰地看到这一点，它们之间是比较一致的。职位因素的各水平在航线上的分布相对均匀。轮机部船员各水平与航线因素各水平同时出现的频数均不超过100。由**图5**可以看到，三副在众多职位水平中对航线因素中的国内南北线水平贡献较大，贡献比为105/670=15.7%，反过来说，在各航线水平中，国内南北线对三副的贡献也最大，贡献比为105/163=64.4%。

上述计算结果表明国内南北线对三副的需求巨大。

3.3.2 航线与船员证书因素之间的关系

图6是对航线因素和船员证书因素各水平之间关系的可视化呈现。**图6**中共有11个节点（占比30.56%），其中，航线因素

有7个水平（占比19.44%），证书因素有4个水平（占比11.11%），边的个数是17（占比4.1%）。若以航线为考察对象，则对国内南北线贡献最大的是丙一证书，贡献比约为79.7%，接近80%，而环球航线、东南亚航线等国际航线对于证书的需求主要是甲类证书，甲类证书对以上两条航线的贡献比分别约为89.85%和90.35%。

上述计算结果表明国内南北线需要的证书类型主要是丙一证书，这与丙一证书适用于国内沿海航线的服务范围是高度一致的。对于国际航线来说，证书绝大多数都是甲类证书。这也与甲类证书适用于全球航线的规定相一致。

3.3.3 船员职位与船型因素之间的关系

图7显示的是船型因素与船员职位之间的关系。若以船型为考察对象，可以很明显地看出三副和水手对散杂货船的贡献相对较大，占比分别达到10.64%和11.96%。反过来说，在所有统计的船型中，散杂货船对三副和水手贡献也最大，分别达到了63.8%和73.58%，占比都超过了2/3。以油轮来说，高级船员（二副、二管轮及以上）占比约为53.85%（42/78），低级船员占比约为25.64%（20/78）。

上述计算结果从侧面反映了对于机械化和信息化水平都不是很高的散杂货船来说，它对三副和水手的需求量都是很大的。这也在一定程度上解释了坊间常说的“有经验的水手比刚刚毕业的高级船员要受欢迎”的现象。对于自动化水平较高的船舶来说，例如集装箱船和油轮，这种现象则很不明显，甚至不存在，高级船员的需求仍是主流。

3.3.4 船员职位与船员证书因素之间的关系

图8展示了职位因素与证书因素各水

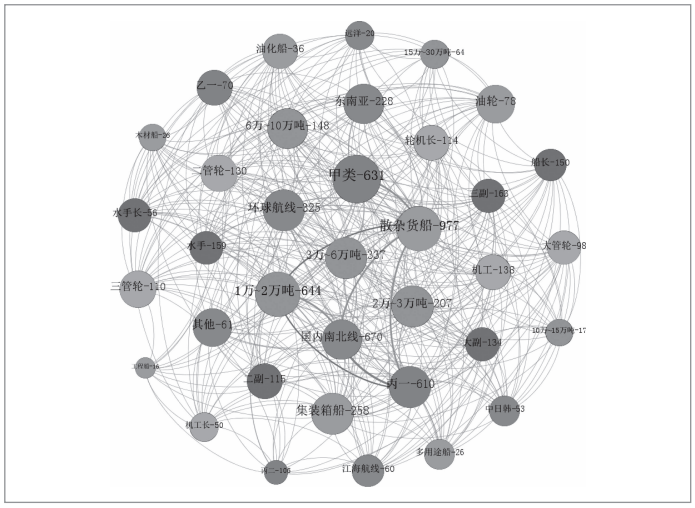

图 4 整体效果

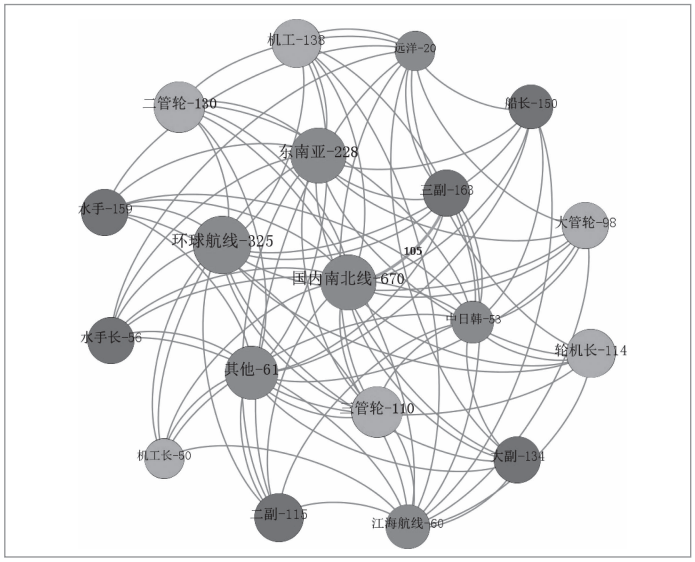

图 5 航线—职位网络关系

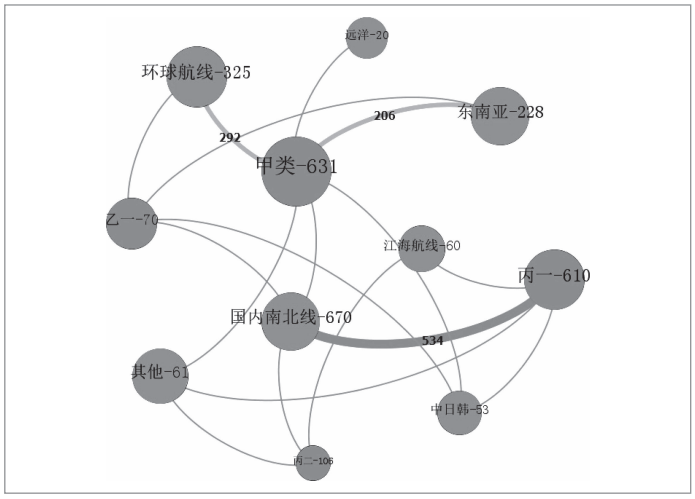

图 6 航线—船员证书网络关系

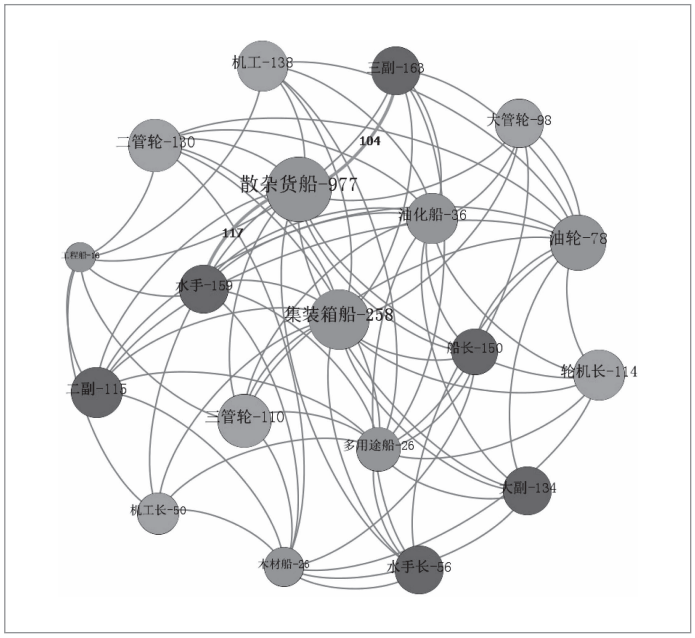

图 7 船型—职位网络关系

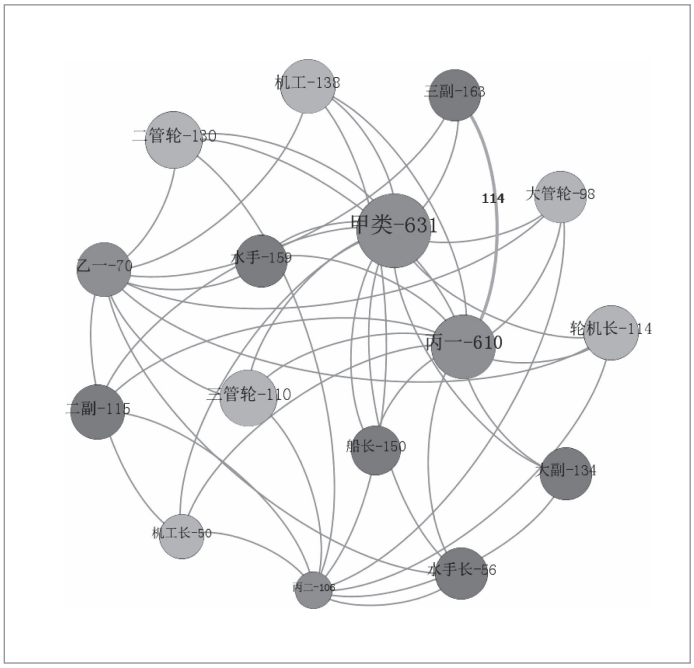

图 8 船员职位—船员证书网络关系

平之间的关系。其中，只有三副和丙一共同出现的频数超过100。若以前者为考察对象，则占比约为69.94%（114/163）；若以后者为考察对象，则占比约为18.69%（114/610）。

上述结果表明在所考察的数据中，三副的证书以丙一居多。换句话说，这里统计的资料表明，低水平的三副占了大多数，超过2/3，对于高水平的三副来说，数量还是不多。

3.3.5 船员职位与船舶吨位因素之间的关系

图9展示了船员的职位因素和船舶吨位因素各水平之间的联系。从**图9**可以看出，所有的水平并没有明显的集中趋势，至少频数并没有超过100的。为了更进一步探索船员职位因素和船舶吨位因素各水平之间的联系，需要重新设置权重。本文为进一步探究船员职位和船舶吨位各水平之间的关系，采用频数与10的比值进行权重设置。为了使频数小于10的边也能在网络图中有所显示，本文在Excel中采用表达式weight=INT(n/10+1)对边进行权重设置。其中n代表边两端的节点同时出现的次数，加1是为了对计算的结果进行平滑处理。通过重新设置权重，可以将船员职位水平与船舶吨位水平之间关系的差异放大，以便进行对比分析。**图10**是权重调整之后的网络关系。其中，节点数为18（占比为50%），边的个数为68（占比为16.4%），平均度为7.556，平均加权度为19.889，网络直径达到3，图密度达到0.444。从**图10**中不难看出，在船舶吨位的分布中，1万~2万吨级的船舶招聘船员的职位占比为644/(644+207+337+148+17+64)=45.4%，对船员职位招聘岗位的贡献接近一半。其中，对三副岗位的贡献达到12.4%（80/644），在其提供的所有岗位中占比最大。对于考察的三副岗位总体来说，

该吨级提供的80个岗位也接近三副总体岗位的一半，占比达49.1%（80/163），可以说1万~2万吨级的船舶对三副的需求量最大。因为权重的阈值设小之后，权重的取值也随之增多，所以为了进一步看清楚，可以通过Gephi的滤波功能，逐步筛选出权重大于3的边以及权重大于6的边，分别如**图11**和**图12**所示。通过**图12**可以很明显地看到，1万~2万吨级的船舶提供的岗位主要是大副、二副、三副、二管轮和三管轮，这些岗位占了1万~2万吨级船舶提供岗位的52.8%（(64+62+80+72+62)/644），占据了所有招聘岗位的一半以上。虽然提供的三副岗位是最多的，但是提供占比较大的岗位却是二副、二管轮、三管轮，提供的岗位占比分别达到了53.9%（62/115）、55.4%（72/130）、56.4%（62/110），提供占比最大的岗位是三管轮。

3.3.6 基于度范围的网络拓扑图

通过过滤器中的拓扑选项卡可以查看本关系网络的度范围是14~31，即各水平至少与14个其他水平相关，至多与31个其他水平相关。调节度范围进行滤波，可以看出，将度范围的起始值设为25，可以保留大部分显著水平，同时精简网络结构。随着逐步提升度范围的起始值，可以逐步过滤度少于起始值的节点，而保留那些度大于或等于起始值的节点。**图13**展现的是度的起始值为27的结果。直到最后，可以发现，度范围在28~31的水平只有3个，如**图14**所示。**图14**中显示的分别为船型水平（散杂货船）、证书水平（甲类）、吨位水平（1万~2万吨级），这3个因素水平具有与其他因素水平联系多且频数大的特点，可以说是整个招聘信息网络的核心。

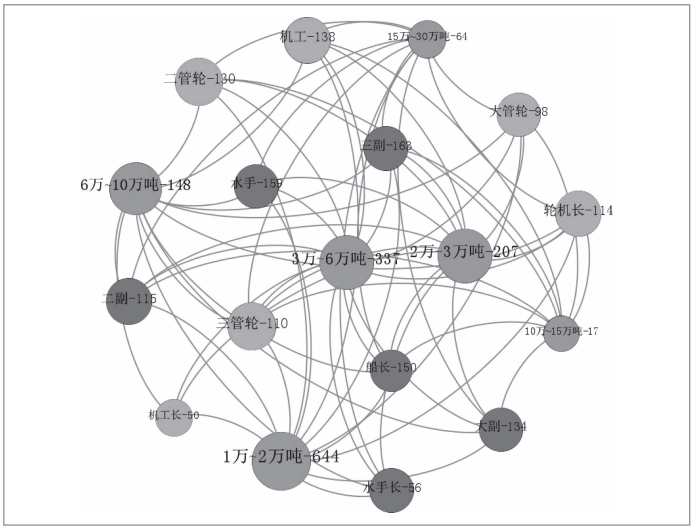

图 9 船员职位—船舶吨位网络关系

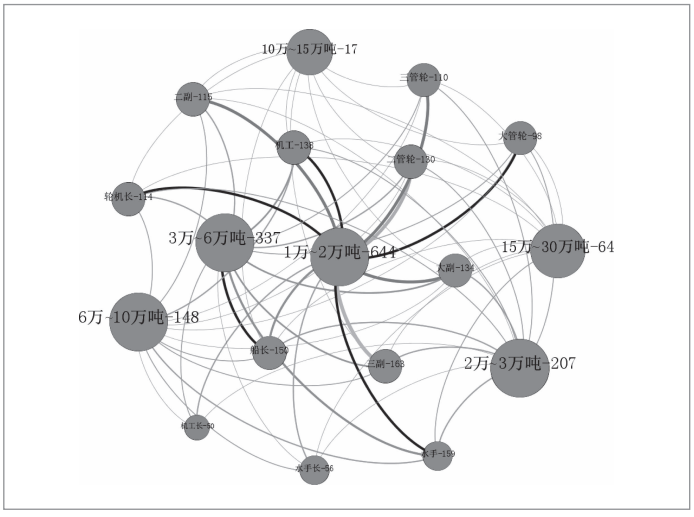

图 10 调整边权重后船员职位—船舶吨位网络关系

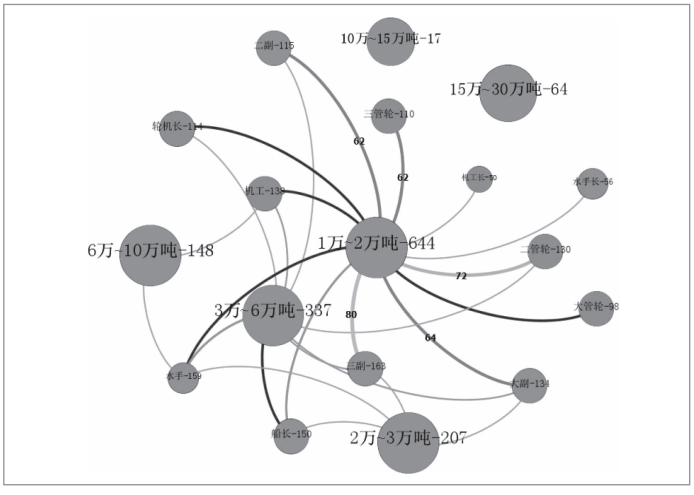

图 11 边权重 >3 的船员职位—船舶吨位网络关系

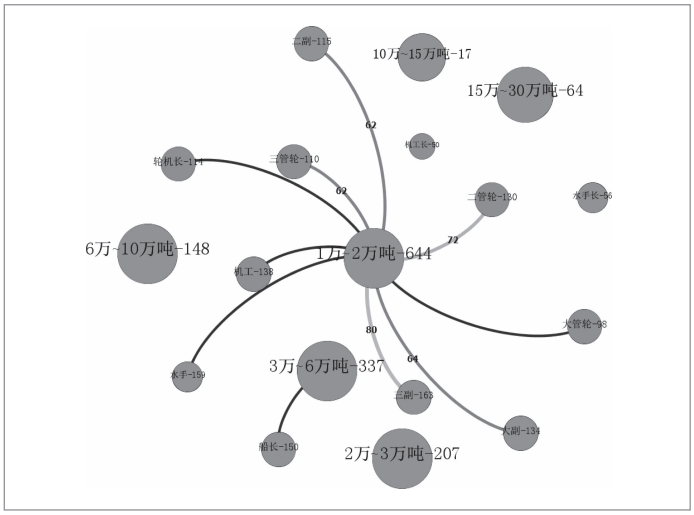

图 12 边权重大于 6 的船员职位—船舶吨位网络关系

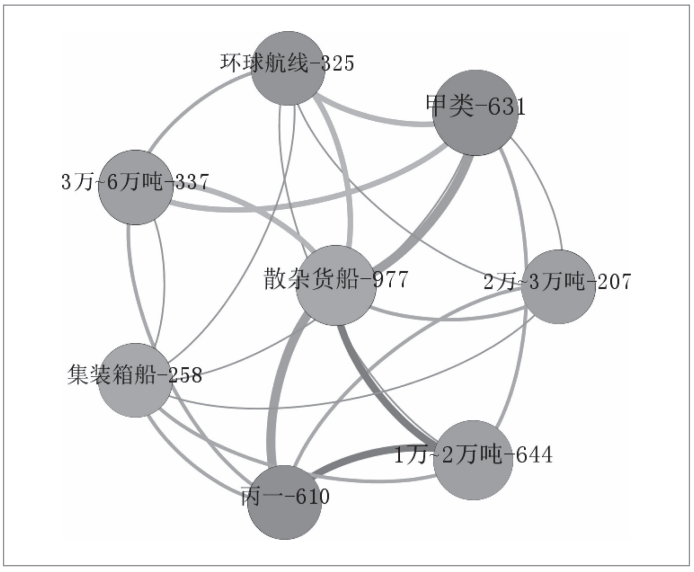

图 13 度范围为 27~31 的网络拓扑

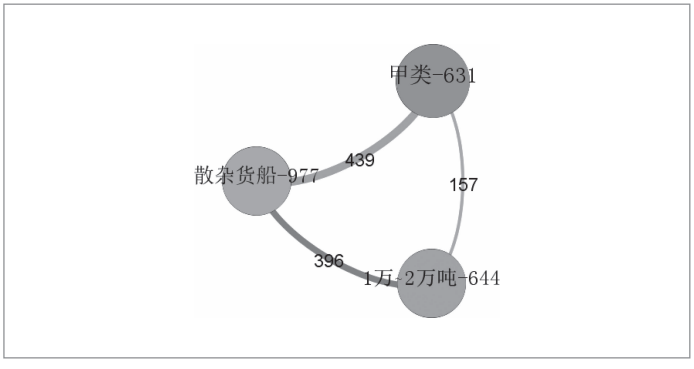

图 14 度范围为 28~31 的网络拓扑

4 结束语

本文基于数据可视化技术以及可视化工具Gephi，利用网络爬虫工具采集了航运在线网上的航运招聘数据，并对数据进行必要的清洗、过滤及加工等预处理操作，得到了相对粗糙的数据。基于Gephi对船员数据进行了初步的可视化分析。在得到数据大致分布之后，使用R语言对数据进行二次加工，将处理好的数据再次通过Gephi进行可视化呈现。从分析结果中可以很直观地看出航运招聘信息中各因素的主要水平值。此外，通过交互式分析，探索了船员数据中各个属性之间的关联程度以及航运招聘信息网络的核心属性，以期为海事部门探索船员市场需求和应聘者的求职提供一定的借鉴及参考。

参考文献:

[1] 刘勇，杜一. 网络数据可视化与分析利器：Gephi中文教程[M]. 北京：电子工业出版社，2017: 163-164.
LIU Y, DU Y. Network data visualization and analysis tool: Gephi Chinese Tutorial[M]. Beijing: Publishing House of Electronics Industry, 2017: 163-164.

[2] 关迎晖，向勇，陈康. 基于Gephi的可视分析方法研究与应用[J]. 电信科学，2013(S1): 112-119.
GUAN Y H, XIANG R, CHEN K. Research and application of visual analysis method based on Gephi[J].Telecommunications Science, 2013(S1): 112-119.

[3] 李学兰. 基于Gephi的物流金融研究可视化[J]. 牡丹江大学学报，2017(1): 26-28, 46.
LI X L. Visualization of logistics finance research based on Gephi[J]. Journal of Mudanjiang University, 2017(1): 26-28,46.

作者简介

王扬(1991-),男,大连海事大学航海学院硕士生,主要研究方向为海事大数据挖掘与应用。

田野(1977-),男,大连海事大学航海学院博士生,主要研究方向为航运大数据挖掘与应用。

李铁山(1968-),男,博士,大连海事大学航海学院教授、博士生导师,主要研究方向为智能计算、智能控制理论及其应用,海洋航行器运动建模与控制,船舶通航安全。

陈俊龙(1959-),男,博士,澳门大学科技学院教授、博士生导师,国家“千人计划”特聘教授,主要研究方向为智能系统与控制、计算智能、混合智能、数据科学。

彭东成(1994-),男,大连海事大学航海学院硕士生,主要研究方向为海事大数据挖掘与应用。

收稿日期: 2018-01-29