

6.1.1	新媒体数据可视化的定义	152
6.1.2	新媒体数据可视化的作用	153
6.2	新媒体数据可视化的设计原则	155
6.3	新媒体数据可视化的常用工具	158
6.4	新媒体数据可视化的常见方法	159
6.4.1	关于关系的可视化	159
6.4.2	关于差异的可视化	164
6.4.3	关于比例的可视化	171
6.4.4	关于分布的可视化	174
6.4.5	关于时间的可视化	176
6.4.6	关于地理空间的可视化	178
6.5	案例分析——拿破仑远征图	180
	实践训练	180
	课后习题	181
第7章	新媒体数据分析与应用实例	184
	【学习目标】	184
	引导案例	184
7.1	淘宝网购物数据分析	185
7.1.1	淘宝网购物数据分析内容	185
7.1.2	淘宝网购物数据分析过程	186
7.2	今日头条数据分析	200
7.2.1	今日头条数据分析内容	201
7.2.2	今日头条数据分析过程	203
7.3	微博用户社交数据分析	210
7.3.1	微博用户社交数据分析内容	210
7.3.2	微博用户社交数据分析过程	211
	实践训练	217
	课后习题	218
	参考文献	219

7.3 微博用户社交数据分析

2009年8月,新浪微博上线,将我国微博发展推向一个新的阶段;之后,各大门户网站相继开通微博服务。微博成为机构和个人宣传及交流信息的平台,其用户群庞大。对其数据内容进行分析和算法讲解,可以为读者在该平台上采集和分析相关数据提供参考。

本节主要讲解微博用户聚类、微博突发事件复杂网络分析及聚类算法、复杂网络可视化、复杂网络中心性分析方法,旨在通过实践加深读者对复杂网络数据处理的方法与手段的理解。需要注意的是,SPSS系列软件虽然提供了复杂网络相关的算法,但其缺乏实践性和相关可视化手段,因此本节中复杂网络相关的处理均使用Gephi实现。

7.3.1 微博用户社交数据分析内容

1. 微博用户聚类

聚类是一种基于目标对象各维度属性值的差异,实现无监督地对目标对象进行分组的算法。聚类算法能够实现对目标对象的描述和分类,具有广泛的应用范围。4.3节、7.1节的内容都是聚类算法的延伸应用。

微博用户聚类是根据用户在微博中的信息行为数据对其进行聚类,是后续用户信息推广、用户画像构造、用户识别等内容的基础步骤。

2. 微博突发事件复杂网络分析

微博突发事件复杂网络分析能够对信息的传播节点、传播形式进行把控,对舆情信息传播、谣言管控、信息透明具有非常重要的意义。

新媒体复杂网络分析方法根据其目的和方法不同可以分为以下几种:链路预测,其预测复杂网络中的两个节点是否会相互连接;网络社区划分,其将网络按照某种属性分布划分为不同区域;网络相似性,其度量两个网络之间的相似性。限于篇幅,本书不对这些内容展开说明,读者可以自行查阅相关资料进行了解。

除了上述分析方法,另一种新媒体复杂网络分析方法则是对网络基础属性的度量,如度中心性分析。度中心性也叫点度中心性、连接中心度,顾名思义是一种度量一个节点与其他节点直接连接的总和的指标,是一种衡量节点在网络之中的中心性的基础指标。在在线社交网络中,与节点连接的边的数量称为该节点的

度，这就是节点的度的概念。这部分内容读者可以对照第5章进行学习。

7.3.2 微博用户社交数据分析过程

1. 微博用户聚类过程

以新浪微博中“高考”话题为例展示微博用户聚类过程。首先通过“高考”话题进行事件数据的采集，字段包括 rankid、发布时间、点赞数、转发数、评论数、图片数、视频数，数据预览如表 7-9 所示。

表 7-9 数据预览

rankid	发布时间	图片数	视频数	转发数	评论数	点赞数
1	2020/7/25	3	0	57	10	106
2	2020/7/26	6	0	23	1	39
3	2020/7/26	0	1	8	6	63
4	2020/7/26	0	1	14	51	216
5	2020/7/26	0	0	0	0	122
6	2020/7/26	1	0	0	20	14
7	2020/7/26	9	0	17	42	113
8	2020/7/26	1	0	2	1	12
9	2020/7/26	0	0	22	36	112
10	2020/7/26	1	0	0	6	9
11	2020/7/26	0	0	0	9	37
12	2020/7/26	0	0	3	1	5
13	2020/7/26	1	0	0	6	15
14	2020/7/26	0	0	2	0	3
15	2020/7/26	0	1	0	0	0

(1) 导入数据。在 SPSS 中，依次选择“源”“EXCEL”选项卡，参考前文步骤导入数据。本数据集保存在本书配套数字资源的“案例数据\第7章\微博数据集.xlsx”中，如图 7-34 所示。

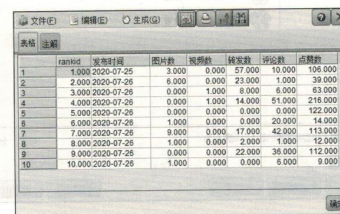


图 7-34 导入数据

(2) 设置变量类型。在“字段”选项中选择“类型”节点，双击节点设置变量，选择“读取值”自动设置变量类型，将数据中表征用户的 rankid 作为目标变量，其余作为输入变量，如图 7-35 所示。

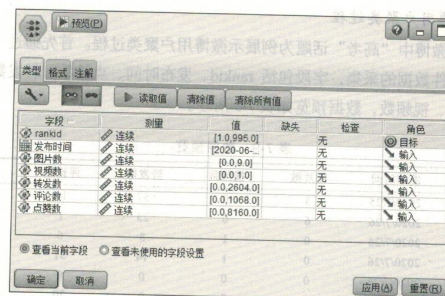


图 7-35 设置变量类型

(3) 模型构建。单击“建模”选项中的“自动聚类”节点，在 SPSS Modeler 中，该节点提供了多种聚类模型，并根据模型表现的优劣自动筛选最终呈现的聚类模型，操作界面如图 7-36 所示。



图 7-36 操作界面

(4) 运行模型。双击新生成的节点打开模型并运行，模型结果如图 7-37 所示。

是否使用?	图形	模型	构建时间 (分钟)	轮廓	聚类数	最小聚类 (N)	最大聚类 (%)	最大聚类 (N)	最小聚类 (%)	重要性
☒			< 1	0.685	5	1	0	518	52	0.002
☐			< 1	0.578	4	41	4	495	49	0.083
☐			< 1	0.572	7	52	5	315	31	0.165

图 7-37 模型结果

由图 7-37 可知，在本次聚类中以轮廓作为指标来看，K-means 算法表现最为优越，因此 SPSS 选择使用该聚类模型，舍弃其余聚类模型。

(5) 解读模型。双击模型图标打开模型详情，聚类模型详细信息如图 7-38 所示。

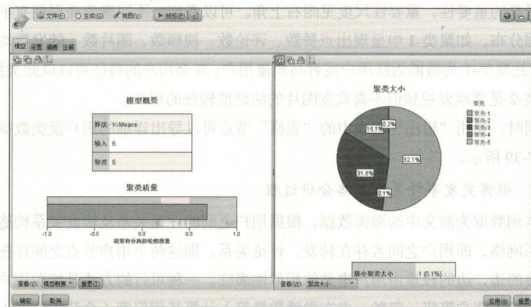


图 7-38 聚类模型详细信息

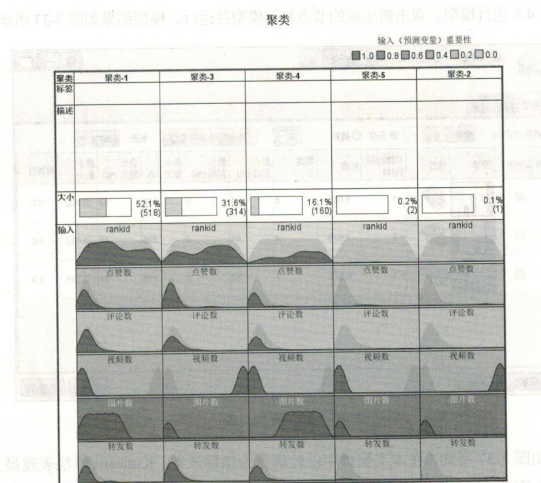


图 7-38 聚类模型详细信息 (续)

由图 7-38 可知, 在本数据集中对于用户的聚类, 其发布内容中的“图片数”具有最高的重要性, 重要性尺度见图右上角。可以看出, 不同聚类类型拥有不同的数据分布, 如聚类-1 中呈现出点赞数、评论数、视频数、图片数、转发数均较高。因此聚类-1 为微博活跃用户或者高质量用户。其余用户的特征可以以此类推, 如聚类-2 是喜欢发视频但不喜欢发图片的活跃度较低的用户。

同时, 使用“输出”选项中的“表格”节点可以导出详细的用户聚类数据, 如图 7-39 所示。

2. 微博突发事件复杂网络分析过程

本例数据为前文中的微博数据, 根据用户之间的评论关系及转发关系构造用户关系网络, 即用户之间若存在转发、评论关系, 则这两个用户节点之间存在互相连接的边, 边的权重由用户事件的相似性表征。一种可行的方法是按照用户节点的向量 (账户等级、年龄、发布微博数量等) 计算其相似度 (余弦相似度)。

数据采用节点对的形式, 部分数据如表 7-10 所示。

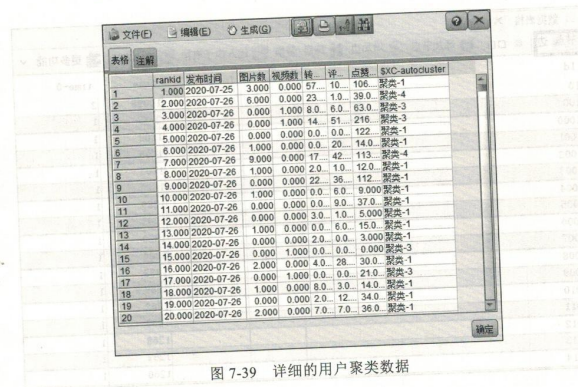


图 7-39 详细的用户聚类数据

表 7-10 部分节点数据

Source	Target	Weight
2	302	0.07
2	303	0.22
5	304	0.26
5	305	0.85
7	306	0.06
7	307	0.11
7	308	0.09
13	309	0.06
13	310	0.48
13	311	0.81
16	312	0.04
17	313	0.54
17	314	0.86
17	315	0.54
17	316	0.96
17	317	0.86
17	318	1.00
17	319	0.45

(1) 导入数据。将数据导入 Gephi 进行分析, Gephi 提供了节点与边的参数权重调整, 如图 7-40 所示。

数据表格					
节点 边					
添加节点 添加边 搜索/替换 输入电子表格 输出表格 更多功能					
Id	Label	Interval	time	time-0	
10			10098	2	
100			2738	1	
1000			1262	1	
1001			1259	1	
1002			1303	1	
1003			1420	1	
1004			1289	1	
1005			1293	1	
1006			1520	1	
1007			1284	1	
1008			1333	1	
1009			1259	1	
1010			1262	1	
1011			1252	1	
1012			1268	1	
1013			1289	1	
1014			1266	1	

图 7-40 导入数据

(2) 数据可视化设置。在 Gephi 左侧布局中，选择节点与边的颜色、大小、标签等属性，如图 7-41 所示。

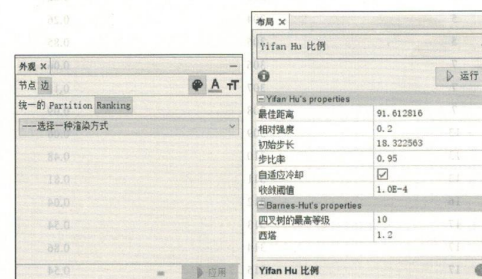


图 7-41 数据可视化设置

(3) 网络可视化。最终网络可视化如图 7-42 所示。

(4) 网络属性计算。在 Gephi 右侧使用“统计”选项卡，Gephi 能够对复杂网络的各类属性进行计算，详细内容读者可参考第 5 章。本例节点的度分布如图 7-43 所示。

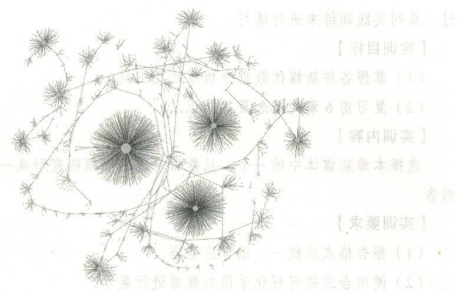


图 7-42 网络可视化

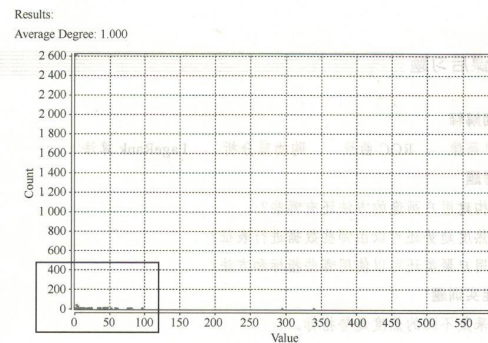


图 7-43 本例节点的度分布

我们可以发现,在实际微博中,节点的度分布是极其不均匀的。大量的节点具有较低的度,只有少数节点是超节点,这与我们正常的认知也是一致的。



实践训练

为了更好地理解新媒体数据分析与应用,并掌握相关的基础知识,下面将通

过一系列实践训练来进行练习。

【实训目标】

- (1) 掌握各种新媒体数据分析方法及工具。
- (2) 复习前6章的概念及其知识。

【实训内容】

选择本章新媒体中的一个，从数据采集到问题研究形成一个完整的数据分析报告。

【实训要求】

- (1) 报告格式应统一，语言应恰当。
- (2) 使用合适的可视化手段对数据进行展示。
- (3) 分享经验与不足，找出理论与实践的出入。



课后习题

一、名词解释

用户画像 ROC 曲线 购物篮分析 PageRank 算法

二、思考题

1. 构建用户画像的方法还有哪些？
2. 热度趋势还可以由哪些数据进行表征？
3. 用户聚类还可以使用哪些指标和方法？

三、技能实训题

1. 采集不同的热度趋势指标。
2. 找出上述指标的适用场景。