

引用格式:钟磊. 基于内容相似度的社区发现算法研究[J]. 微电子学与计算机, 2021, 38(4): 69-73. [ZHONG L. Research on community detection algorithm based on content similarity[J]. Microelectronics & Computer, 2021, 38(4): 69-73.]

基于内容相似度的社区发现算法研究

钟 磊^{1,2}

(1. 新疆财经大学 信息管理学院, 新疆 乌鲁木齐 830012;
2. 新疆财经大学 统计与数据科学学院, 新疆 乌鲁木齐 830012)

摘 要: 传统的社区发现算法中网络节点相似度多以空间距离度量, 这种度量往往不容易理解, 或者只能从距离的角度予以解释. 本文提出一种基于文本内容相似度的谱聚类方法, 它以网络社区用户的文本信息的相似性来度量网络节点的属性相似度, 考虑网络结构的同时使节点的属性联系更有意义. 在此基础上使用谱聚类的思想进行社区划分. 本文以实际数据进行实验发现, 发现划分结果的模块度和节点平均相似度较高, 聚类效果良好.

关键词: 社区发现; 内容联系; 相似度; 谱聚类; 图划分

中图分类号: TP301.6 **文献标识码:** A **文章编号:** 1000—7180(2021)04—0069—05

Research on community detection algorithm based on content similarity

ZHONG Lei^{1,2}

(1 Institute of information management, Xinjiang University of Finance and Economics,
Urumqi 830012, Xinjiang China;

2 Institute of Statistics and Data Science, Xinjiang University of Finance and Economics ,
Urumqi 830012, Xinjiang China)

Abstract: In traditional community detection algorithms, the similarity of network nodes is usually measured by spatial distance, which is not easy to understand or can only be explained from the perspective of distance. This paper proposes a spectral clustering method based on the similarity of text content, which measures the attribute similarity of network nodes by the similarity of text information of users in the network community, and makes the attribute connection of nodes more meaningful when considering the network structure. On this basis, the idea of spectral clustering is used to divide the community. The experimental results show that the modular degree and the average similarity of nodes are high, and the clustering effect is good.

Key words: community detection; content link; similarity ; spectral clustering ; graph division

1 引言

根据中国互联网络信息中心统计,截至 2019 年 6 月,中国网民数量达 8.54 亿,使用手机的网民达 8.47 亿,网民使用手机上网占 99.1%^[1]. 网络已经成为人们表达个人想法、交流和共享信息的不可或

缺的手段. 在网络空间中,用户几乎每天二十四小时不间断的发布海量信息,比如评论、帖子、图片、视频等等. 大量的网络用户由于知识结构,职业特点,兴趣爱好等因素的不同而具有不同的信息偏好. 具有共同偏好的用户往往成为一个社区,社区中的用户节点在网络空间中表达相似的观点和看法. 对网络

社区的研究可以帮助我们更有效的了解用户的信息需求,研究不同偏好网络用户的行为特征和规律,识别网络热点信息等.

2 社区发现的研究现状

社会网络分析的概念最早是由英国著名人类学家 R. 布朗提出的,布朗所探讨的网络概念聚焦于文化是如何规定群体内部成员的行为的. 从上世纪 30 年代到 60 年代,在心理学、社会学、人类学、统计学等研究领域,越来越多的学者开始构建“社会结构”概念,社会网络分析理论、方法和技术日益深入,成为一种重要的社会结构研究范式. 而在社会网络分析中对于网络节点的聚类分析,简称“图聚类”^[2],即我们所称的“社区发现”. 经典的社区发现方法主要包括基于分割的方法^[3];GN 方法^[4];基于模块度的方法^[5]等等. 2002 年,Newman 和 Girvan 两人提出了经典的社区发现算法——GN 算法,该算法利用社会网络中“边介数”进行社区发现. 随后一系列社区发现算法被提出. 基于 GN 算法的基本思想,Fortunato 等人将网络中节点间的拓扑关系抽象成节点的属性,提出一种新的“社区发现”算法^[6]. 针对 GN 算法的不足,Newman 本人在 2004 年提出一种改进的 GN 算法——快速 Newman 算法,该算法通过不断优化模块度的思想,自底向上不断凝聚节点,最后达到社区发现的目的. 2008 年,Blondel 等人提出了另一种基于模块度的 fastunfolding 社区发现算法,该算法基于“社区折叠”的概念,把已经发现的社区作为新的节点继续分析,从而形成分层结构的社区,且不用设置初始社区数目,对未知网络有很好的表现.

国内学者谢洺等人在 fastunfolding 算法基础上提出一种时间复杂度更低的社区划分算法^[7]. 张思源等人通过循环迭代不断移除社会网络中相异性最高的边,从而获得新的社区结构,同时计算出其模块度,以模块度的最大值为标准,得到网络的社区划分结果^[8]. 姜雅文、于剑等利用 Jccarrd 距离公式度量社会网络节点之间的相似度,然后运用多种聚类算法对模拟社会网络进行了社区发现,效果良好^[9]. 陈东明、王云开等人提出了一种单社区结构评价标准——社区密合度,在此基础上,设计了一种基于凝聚思想的社区发现算法,通过与其他社区发现算法的对比,获得了较好的划分效果和较低的时间复杂度^[10]. 孟彩霞等针对有向加权网络,提出一

种基于节点重要性和相似性的标签算法,实验证明该算法具有较好的鲁棒性^[11]. 李丛正等学者提出基于机器学习的多关系网络社区发现算法,使用机器学习得出网络节点的向量表示,然后在向量表示的基础上进行社区划分.

传统的社区发现算法从“距离”的角度解释划分结果,只考虑了网络节点间的拓扑结构关系,忽视了节点间属性,使得社区划分结果缺乏语义性,不便于理解和解释. 而基于节点属性的社区发现算法容易产生属性数据稀疏性问题^[12-13]. 所以很多学者考虑同时考虑网络节点拓扑结构和节点属性. 本文提出一种基于文本内容相似度的谱聚类方法,利用文本内容信息量大和维度丰富的特点,以文本相似性作为网络节点属性相似度的度量,将网络结构和节点属性相结合,同时还在一定程度上解决了属性数据的稀疏性问题,提高了社区划分效果.

3 基于内容相似度的谱聚类模型建立

3.1 内容相似度计算

先将用户信息文本进行分词处理,然后计算词的 $tf-idf$ 值. tf 值代表词频,其计算如公式(1)所示,其中分子代表一个词在一个文本中出现的频数,分母代表整个文本的总词数. 用来代表该词在文本中的重要程度.

$$tf_{i,j} = n_{i,j} / \sum_k n_{k,j} \quad (1)$$

idf 代表逆向文档频率,其计算如公式(2)所示,其中分子代表某个语料库中全部文本的个数,分母代表某个词在所有文本中的不重复次数,对商取对数即得到 idf 值.

$$idf_{i,j} = \log \frac{|D|}{\{j: t_i \in d_j\}} \quad (2)$$

将 tf 值与 idf 值相乘即得到 $tf-idf$ 值,如公式(3)所示:

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,j} \quad (3)$$

得到词的 $tf-idf$ 值后,用公式(4)和公式(5)计算用户信息内容的相似度. 其中 $tfidf_{v_1}$ 代表网络用户节点 V_1 的信息文本特征向量, $tfidf_{v_2}$ 代表网络用户节点 V_2 的信息文本特征向量. 通过计算两个向量的余弦相似度来代表用户节点间的相似度 ss .

$$\cos\theta = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

$$SS(v_1, v_2) = \cos(tfidf_{v_1}, tfidf_{v_2}) \tag{5}$$

3.2 建立基于内容相似度的网络加权图

首先我们把网络用户映射到图模型中,其具体定义如下:

定义 1 设 $V = \{V_1, V_2, \dots, V_i, \dots, V_n\}$ 为网络用户集合,无序偶对 (V_i, V_j) 是用户 V_i 与用户 V_j 之间的边,表示用户 V_i 与用户 V_j 之间的联系,如何两个网络用户之间有回帖行为,即代表存在这种联系. $G(C, E)$ 是以 V 为用户集合,以 $E \in \{(V_i, V_j) \mid V_i, V_j \in V\}$ 为联系集合的无向图.

定义 2 用户 V_i 的度 D_i 是指与其他用户相关联的边数, $D_i = |\{(V_i, V_j) \mid (V_i, V_j) \in E, V_i, V_j \in V\}|$.

用户的聚集系数是指与该用户相连的近邻用户之间互联的比例,如公式(6)所示:

$$C_i = 2K_i / (D_i(D_i - 1)) \tag{6}$$

式中, $K_i = |\{(V_j, V_k) \mid (V_i, V_j) \in E, (V_i, V_k) \in E, V_j, V_k \in V\}|$.

定义 3 设 $V = \{V_1, V_2, \dots, V_i, \dots, V_n\}$ 为网络用户集合,无序偶对 (V_i, V_j) 是用户 V_i 与用户 V_j 之间的边, W_{ij} 为用户 V_i 与用户 V_j 之间的信息文本相似度数值,作为边的权值,则 $WG(V, E, W)$ 是以 V 为用户集合,以 $E \in \{(V_i, V_j) \mid V_i, V_j \in V\}$ 为边集合, $W = \{W_{ij} : (V_i, V_j) \in E\}$ 为权值集合的无向加权图. 用户的加权度反映了该用户与其他用户的连接强度,如公式(7):

$$WD_i = \sum_{(v_i, v_j) \in E} W_{ij} \tag{7}$$

用户的加权聚集系数 WC_i 如公式(8):

$$WC_i = 2WK_i / (C_i(C_i - 1)) \tag{8}$$

式中, $WK_i = \sum_{(v_j, v_k) \in E} W_{jk} \tag{9}$

$$R = \{(V_j, V_k) \mid (V_i, V_j) \in E, (V_i, V_k) \in E, V_i, V_j, V_k \in V\}.$$

最后根据用户的加权聚集度和加权聚集系数得到加权综合特征值公式(10).

$$WCF_i = \alpha WC_i + (1 - \alpha)WD_i / N \tag{10}$$

式中, α 为可调的参数, $0 < \alpha < 1$, N 为图中用户节点个数. 在此基础上使用谱聚类方法对该图进行划分,得到的网络用户社区的划分结果.

4 实验结果分析

4.1 实验过程

本文以某网络论坛作为数据采集源,爬取论坛

讨论帖子链接,如果一个用户回复了一个主题帖,则该回复用户和发帖用户之间建立一条边. 在实际实验中,最终获取帖子数据 1 026 条,实际用户数量为 75. 按照前面所述的方法建立无向图,节点数 75 个,边数为 276,如图 1 所示.



图 1 网络拓扑结构

将用户的帖子文本按 3.1 和 3.2 节所述方法计算,得出相似度矩阵如表 1 所示.

表 1 内容相似度矩阵

节点	1	2	3	...	75
1	0	0.21	0.6	...	0.05
2	0.21	0	0.59	...	0.02
3	0.6	0.59	0	...	0.12
...	...	...	...	...	...
75	0.05	0.02	0.12	...	0

按照定义 3 生成加权图,如图 2 所示,图中边的粗细代表权值的大小.

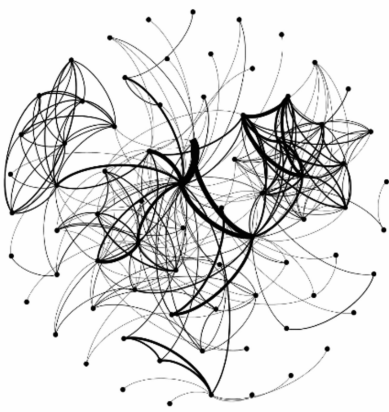


图 2 加权网络

将初始社团数分别设置为 2, 5, 8 和 11, 对加权图进行谱聚类. 图 3 为社团初始值为 5 时的社区划分结果.

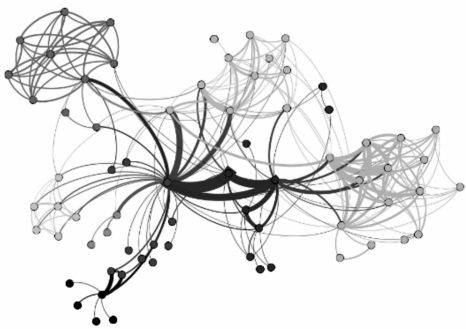


图 3 社区划分

4.2 结果分析

我们用模块度和平均相似度两个指标来评价社区划分的结果. 模块度的值在[0,1] 之间, 模块度越大说明社区划分的内部连通性越好, 这一指标可以评价社区划分结构上的好坏, 具体计算如公式(11),

$$Q = \sum_{s=1}^m \left[\frac{L_s}{L} - \left(\frac{D_s}{2L} \right)^2 \right]$$
 (11)

式中, L_s 代表社区内部的边; L 代表用户网络中所有边的数量; D_s 代表社区内部所有节点度的和.

平均相似度是已经划分好的网络内部节点相似性的判断指标, 用来衡量社区中节点的聚类程度. 其计算如公式(12).

$$Sim = \frac{\sum S_i}{N}$$
 (12)

式中, S_i 为节点两两相似度的均值; N 为划分好的社区数目.

对实验中的 4 次聚类分别计算它们的模块度和平均相似度, 计算结果如表 2 所示.

表 2 模块度和平均相似度比较

初始社区数	模块度	平均相似度
2	0.212	0.346
5	0.329	0.382
8	0.298	0.362
11	0.281	0.379

由表 2 可知, 初始社区数目为 5 时, 社区划分效果最好, 模块度和节点平均相似度最高, 从全部实验来看, 平均相似度指标比较稳定, 与之相比, 模块度指标则有一定波动性.

将本文算法与 fastunfolding 算法作比较, 发现本文算法在模块度上与 fastfolding 算法不相上下, 但在节点平均相似度指标上相比 fastfolding 算法具有较大优势.

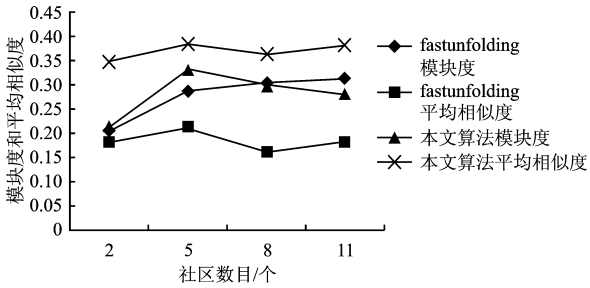


图 4 算法对比

5 结束语

本文提出了一种基于内容相似度的谱聚类算法, 以网络用户发布的文本信息的相似度来度量用户的相似性, 以信息丰富的文本内容为基础进行群体划分, 显然具有更高的簇内相似度. 通过实验证明, 即使在较纯粹的拓扑结构中, 也很难保证簇内部的节点有很高的相似性. 基于内容相似度的群体划分方法则不一样, 文本的大信息量和丰富维度, 使得簇内部的节点相似度更高且较稳定, 所能刻画的相似用户也更有解释力和说服力, 社区发现的质量会更好.

参考文献:

[1] 中国互联网络信息中心. 第 43 次中国互联网络发展状况统计报告[R]. 北京: 中国互联网络信息中心, 2019.
CNNIC. The 43rd statistical report on the development of internet in China[R]. Beijing: CNNIC, 2019.
[2] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2010, 486(3-5): 75-174. DOI: 10.1016/j.physrep.2009.11.002.
[3] FIEDLER M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(98): 298-305. DOI: 10.21136/CMJ.1973.101168.
[4] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826. DOI: 10.1073/pnas.122653799.
[5] SHANG R H, BAI J, JIAO L C, et al. Community-detection based on modularity and an improved genetic algorithm[J]. Physica A: Statistical Mechanics and Its Applications, 2013, 392(5): 1215-1231. DOI: 10.1016/j.physa.2012.11.003.
[6] LANCICHINETTI A, FORTUNATO S, KERTLESZ J. Detecting the overlapping and hierarchical community

- structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 033015. DOI: 10.1088/1367-2630/11/3/033015.
- [7] 解泊,汪小帆. 复杂网络的一种快速局部社团划分算法[J]. 计算机仿真, 2007, 24(11): 82-85. DOI: 10.3969/j.issn.1006-9348.2007.11.026.
- XIE Z, WANG X F. A fast algorithm for detecting local community structure in complex networks [J]. Computer Simulation, 2007, 24(11): 82-85. DOI: 10.3969/j.issn.1006-9348.2007.11.026.
- [8] 张思源,覃森,张智丰. 基于邻居节点相异性的社团发现新算法基于邻居节点相异性的社区发现新算法[J]. 杭州电子科技大学学报(自然科学版), 2018, 38(05): 98-102. DOI: 10.13954/j.cnki.hdu.2018.05.019.
- ZHANG S Y, QIN S, ZHANG Z F. A new algorithm for community detecting based on neighbor node dissimilarity[J]. Journal of Hangzhou Dianzi University (Natural Science), 2018, 38(5): 98-102. DOI: 10.13954/j.cnki.hdu.2018.05.019.
- [9] 姜雅文,贾彩燕,于剑. 基于节点相似度的网络社团检测算法研究[J]. 计算机科学, 2011, 38(7): 185-189. DOI: 10.3969/j.issn.1002-137X.2011.07.041.
- JIANG Y W, JIA C Y, YU J. Community detection in complex networks based on vertex similarities [J]. Computer Science, 2011, 38(7): 185-189. DOI: 10.3969/j.issn.1002-137X.2011.07.041.
- [10] 陈东明,王云开,黄新宇,等. 基于社团密合度的复杂网络社团发现算法[J]. 东北大学学报(自然科学版), 2019, 40(2): 186-191. DOI: 10.12068/j.issn.1005-3026.2019.02.007.
- CHEN D M, WANG Y K, HUANG X Y, et al. Community detection algorithm for complex networks based on group density[J]. Journal of Northeastern University (Natural Science), 2019, 40(2): 186-191. DOI: 10.12068/j.issn.1005-3026.2019.02.007.
- [11] 孟彩霞,李楠楠,张琰. 基于复杂网络的社区发现算法研究[J]. 计算机技术与发展, 2020, 30(1): 82-86. DOI: 10.3969/j.issn.1673-629X.2020.01.015.
- MENG C X, LI N N, ZHANG Y. Research on community detection algorithm based on complex network [J]. Computer Technology and Development, 2020, 30(1): 82-86. DOI: 10.3969/j.issn.1673-629X.2020.01.015.
- [12] HONG L J, DAVISON BD. Empirical study of topic modeling in Twitter[C]//Proceedings of the 1st Workshop on Social Media Analytics. Washington, DC, USA: ACM, 2010: 80-88. DOI: 10.1145/1964858.1964870.

作者简介:

钟磊 男(1980-),博士研究生,讲师.研究方向为信息管理和数据挖掘. E-mail:873770090@qq.com