虚拟学习社区中意见领袖识别模型研究

许 睿,李艳翠,訾乾龙,李宗儒,张平川
(河南科技学院 信息工程学院,河南 新乡 453003)

摘 要:虚拟学习社区是传统教育突破空间资源限制形成的便捷性学习环境,其中意见领袖是构成社区信息通路的重要角色,对其他用户有强大的影响力。为了准确识别社区中的意见领袖,构建出虚拟学习社区网络,分析各用户的中心性和社会网络角色特征,选取入度、出度、介数、特征向量中心性、用户活跃度、用户帖子转发量、用户帖子评论量等七个特征值作为筛选条件,结合基于 K-means 的用户聚类算法,提出基于 K-means 算法的意见领袖识别模型。最后,将该识别模型应用于某虚拟社区,根据各个聚类子类的特征向量,提取理论意义上的意见领袖集合。实验证明,获取意见领袖集合具有很高的准确性,识别出的意见领袖均处于中心者或桥梁位置,占据着社会网络的优势位置,在虚拟社区中承担着核心或中介等特殊作用。

关键词:意见领袖;识别模型;中心性;虚拟社区;K-means 算法
中图分类号:TP319 **文献标识码**:A **文章编号**:1673-629X(2020)05-0056-05
doi:10.3969/j.issn.1673-629X.2020.05.011

Research on Identifying Model of Opinion Leader in Virtual Learning Community

XU Rui,LI Yan-cui,ZI Qian-long,LI Zong-ru,ZHANG Ping-chuan
(School of Information Engineering,Henan Institute of Science and Technology,Xinxiang 453003,China)

Abstract:Virtual learning community is a convenient learning environment which breaks through the limitation of traditional educational space resources. Opinion leaders play an important role in the formation of community information channels and have a strong influence on other users. In order to accurately identify opinion leaders in the community,we construct a virtual learning community network,analyze the user-centered and social network role characteristics,and select in degree,out degree,betweenness centrality,eigenvector centrality,user activity,the amount of user posts forwarded,number of comments on user posts as screening conditions. Based on K-means user clustering algorithm,an opinion leader recognition model based on K-means algorithm is proposed. Finally,we use the model to process a virtual community,and extract the theoretical opinion leader set according to the feature vectors of each clustering subclass. Experiment shows that the collection of opinion leaders has high accuracy,and the identified opinion leaders are in the center or bridge position,occupying the dominant position of social network,and playing a special role of core or intermediary in the virtual community.

Key words:opinion leader;recognition model;centrality;virtual community;K-means algorithm

0 引 言

“意见领袖”这一概念最早出现于 1940 年,由拉扎斯菲尔德和贝雷尔森在著名的“伊里调查”中提出,他们发现观念常常是从大众观点流向意见领袖,然后通过意见领袖发声传递给不太活跃的人群,意见领袖在社会网络中是活跃分子,处于核心地位[1]。在网络教学平台的支撑和支持下,虚拟学习社区蓬勃发展,已经成为教育信息传播和共享的重要途径,为教师、学习者、管理者等提供了多种学习途径,这些用户可以自由、开放地发表观点、讨论、分享等,这一系列的互动方式最终形成了一个完整的社会网络[2]。剖析虚拟学习社区成员的日常行为可以发现,各类用户主要通过个人学习、分享学习、交流学习、指导学习等方式获取知识,其中意见领袖经常可以通过直接或者间接的社会关系,对其他用户的学习行为和学习习惯造成影响,促进在线知识的传播,提升传播的速度、广度以及范

收稿日期:2019-07-03 **修回日期**:2019-11-05 **网络出版时间**:2020-01-10
基金项目:国家自然科学基金(61502149);河南省高等教育教学改革研究与实践项目(2017SJGLX392);河南科技学院大学生创新创业训练项目(2018CX74)
作者简介:许 睿(1987-),男,讲师,研究方向为复杂网络、数据挖掘。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20200110.1123.044.html

围[3]。对于虚拟学习社区的研究,主要从两个方面进行分析。第一,以社区成员的行为特征为依据,通过聚类算法分析成员的活跃程度、响应值、浏览量等得到社区的意见领袖群[4];第二,从复杂网络的角度进行分析,衡量成员的中心性、特征属性等识别出意见领袖[5]。文中将结合以上两种思路,从社区拓扑结构和社会网络角色两方面提取特征,进行虚拟学习社区的意见领袖识别。

1　虚拟学习社区网络的拓扑结构

1.1　构建虚拟学习社区网络

在虚拟学习社区中,用户的社交行为主要有发帖、评论、转发、关注等,其中最直接的相互作用方式为"关注|被关注"[6]。依据用户间的关注关系构建用户关系网络 $G(V,E)$,其中 V 为网络中节点的集合,$v \in V$ 表示虚拟学习社区中某一个用户;E 为有向边集合,边$<a,b> \in E$ 表示该用户 a 对用户 b 存在关注行为,也可用"$a \rightarrow b$"表示由用户 a 指向被关注用户 b。虚拟学习社区用户关系网络是典型的有向网络。

1.2　用户关系网络的中心性分析

由于社会网络是复杂网络中一个重要分支,因此社会网络也具有复杂网络的两个重要特性,分别是小世界特性[7]和无尺度特性[8]。其中无尺度特性在社会网络中普遍适用,具有成长性和优先连接的特征,主要表现在社会网络中新增节点趋向于与高连接的节点相连接,高连接的节点数量少,但是和网络中大多数节点相联系,这些节点被称为节点中枢或集散节点[9]。在社会网络中,进行节点的中心性度量,有助于识别网络中的意见领袖,对于探索意见领袖的网络位置和分析信息流传播规律有重要的意义。常见的节点中心性度量方法包括:度中心性(degree centrality,DC)、介数中心性(betweenness centrality,BC)和特征向量中心性(eigenvector centrality,EC)。

(1)度中心性。

度中心性主要统计与该节点直接相邻的其他节点的数目[10]。

$$DC(i) = \sum_{j} C_{ij} \tag{1}$$

其中,C_{ij} 表示社会网络中节点 j 到节点 i 的权重值,可用转发量、回帖量、评论量等进行衡量。节点的度最容易获取,当某节点的度值越大,表明该节点参与的社交行为越多。在有向社会网络中,节点的度可分为入度和出度,分别反映该节点被关注、关注其他用户的数量。

(2)介数中心性。

衡量介数中心性的分式中,分子表示社会网络中通过某个节点 v 的最短路径的数量,分母表示网络中所有最短路径数[11]。

$$BC(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

用户节点的介数中心性反映了该用户在社会网络中是否处在枢纽位置,介数值越大,则表明有越多的最短路径通过该用户进行连接。介数值大的用户对于网络信息的传播有重要的意义,这类用户一旦被删除,将会导致网络直径增大,信息传播通路中断。

(3)特征向量中心性。

特征向量中心性是从社会网络整体的角度发掘具有全局性的核心用户节点[12]。

$$EC(u) = \alpha_{\max}(u) \tag{3}$$

有研究表明社会网络中核心节点的邻居节点比边界节点的重要性更高。特征向量中心性就是通过统计该用户的邻居的重要程度,反衬出该用户在社会网络中的中心程度。

文中选取入度、出度、介数、特征向量中心性作为评价用户节点中心性的四个特征参数,用于衡量虚拟学习社区用户的重要性程度。

2　基于 K-means 算法的意见领袖识别模型

2.1　意见领袖的社会网络角色分析

在社会网络中,众多用户之间存在着规则性的联系,在拓扑结构上也具有相似性,依据用户所处的社会网络位置,可以对用户的社会网络角色进行划分,主要包括:中心者、桥梁、边缘者[13]。

中心者处于社会网络的核心位置,拥有更多的社会资源,联系并影响着众多其他用户;中心者社会活跃度高,往往是社区舆论的发起者;中心者拥有较高的社会声望,有能力对网络信息的传播和网络环境的改变产生影响。

桥梁处于结构洞位置,用于连接社会网络中两个毫无关系的用户或者用户群,起到中间人的角色,控制信息的传递,容易识别并获取不同渠道或群体的信息资源,比其他位置上的用户更具有竞争力。这一类用户往往具有特殊的身份,如新闻评论员、微博大 V、论坛管理员、聊天群群主等,对于舆论传播和发展,有管理、识别、控制、引导等作用。

边缘者处于社会网络的边界位置,与其他用户较少联系。他们参与的网络活动主要包括浏览信息、回复评论、关注用户等,对于社会网络信息的识别、传播、控制的能力有限,社会影响力小。

综上所述,对于意见领袖的识别需要重点研究社会网络中的中心者、桥梁这类用户。

2.2 社区用户特征向量分析

依照虚拟学习社区用户关系网络的拓扑特性以及社会网络中意见领袖的角色特性,文中选取以下 7 个特征值作为筛选意见领袖的重要条件,并构建出用户特征向量 $\boldsymbol{U}_i = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ 。

入度 v_1:表示其他用户对于该用户的关注数量,入度值的高低反映该用户与其他用户的交互情况,通常用来描述该用户间交互的主动性与积极性。

出度 v_2:表示该用户关注其他用户的数量,当一个用户关注的人越多,获取信息的能力就越强,对整个网络的信息传播有着重要意义。

介数 v_3:介数常用来衡量网络中某节点控制其他节点间交流的能力,通过比较其他节点的交流要通过该节点的次数来进行计算,可以有效地判别网络中节点的全局重要性。

特征向量中心性 v_4:特征向量中心性就是通过统计该用户的邻居用户的重要程度,反衬出该用户在社会网络中的中心程度。

用户活跃度 v_5:用户活跃度用来统计用户发帖的数量。用户活跃度越高,反映该用户的社会行为比较频繁,在对信息产生和传播的数量上有一定的比重,在筛选意见领袖时用户活跃度是一个关键的因素。

用户帖子转发量 v_6:用户帖子的转发量可以直观地反映信息的传播情况。转发量越大,说明该消息传播范围越广,发起者的意见可以直接或间接地影响读贴人对某一事件的看法,从而引发舆情的转变。

用户帖子评论量 v_7:评论量的高低体现出话题的热度。帖子评论量越高,说明该话题被其他用户关注的越多,帖子本身质量高并且更具有吸引力。

2.3 基于 K-means 的用户聚类算法

聚类是一种无监督的学习方式,包括划分法、层次法、模糊聚类法等,通过设定判断标准,把原数据集分割为多个单独存在的簇,同一簇中数据相似性大,不同簇中数据差异性大[14]。文中采用的 K-means 算法是最著名的聚类算法之一,具有高效、复杂度低的特点。算法描述如下:

输入:样本集 $D = \{x_1, x_2, \cdots, x_m\}$,聚类的簇值 k ,最大迭代次数 N

构建初始簇集 $C = \{C_1, C_2, \cdots, C_k\}$,随机选取 k 个样本点作为初始中心点 $Q = \{q_1, q_2, \cdots, q_k\}$;

repeat

计算样本集 D 到中心点集 Q 的距离,划分到各个簇中 $C_i(i \in [1,k])$;

重新计算各簇的中心点;

until 簇集 C' 无变化|迭代次数≥ N

输出:簇集 $C' = \{C_1, C_2, \cdots, C_k\}$

2.4 虚拟学习社区意见领袖识别

虚拟学习社区是教育信息化可持续发展的新途径。在社区中,意见领袖处于重要的社会网络位置,充当关键的社会网络角色,掌握了大量的强联结关系以及重要的弱联结关系,对于建立稳定的信息交流秩序和提高学习信息的传播有重要的作用[15]。

依据意见领袖的拓扑特性以及角色特性,文中设定意见领袖的筛选条件为:①簇成员数较小;②簇成员的特征向量 $\boldsymbol{U}_i = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ 的均值较大[16];③簇成员具有特殊的社会网络角色。

意见领袖识别过程如下:首先利用 K-means 算法获取聚类结果,簇集 $C' = \{C_1, C_2, \cdots, C_k\}$;然后选取成员数较小的簇作为备选意见领袖集合,再分别计算各备选集合中特征值 $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ 的均值,并降序排列;最后分析均值最大的簇中各成员的社会网络角色,最终得到意见领袖集合。具体过程如图 1 所示。

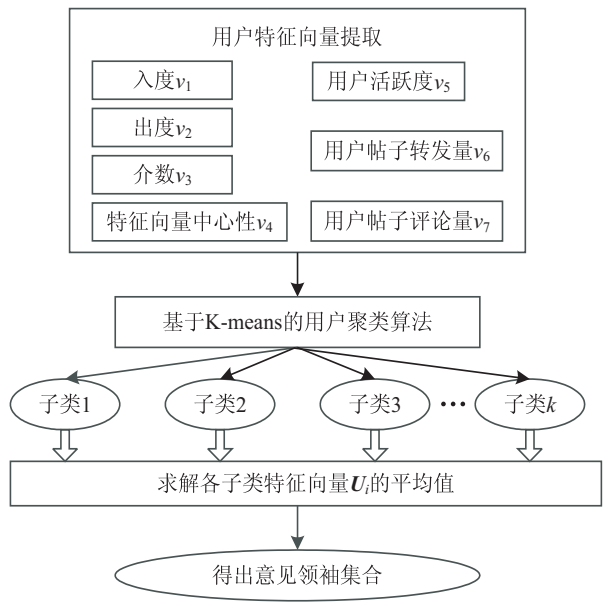

图 1 基于 K-means 算法的意见领袖识别模型

3 虚拟学习社区中意见领袖识别实证研究

3.1 实验数据

文中选取国内某虚拟社区的用户数据作为研究对象,该社区数据具有用户多、特征量多、关系复杂等特点。采用 Python 语言编写爬虫工具,利用 Scrapy 突破反爬虫机制,有效地提升爬取速度,对于该社区 2018 年 6 月 1 日至 2018 年 12 月 31 日的数据进行抓取,共爬取了 49 996 条关系信息,通过筛选得到 40 231 条用户信息。文中分别计算每个社区用户的特征值,构建用户的特征向量,并将这些用户的特征向量作为 K-means 算法的测试数据。典型的社区用户特征向量数据如表 1 所示。

表 1　典型的社区用户特征向量

编号	特征向量值
1	{0,1,0,0,5,0,0}
2	{123,12,0.000 097,0.028 7,5,0,0}
3	{19,1,0.000 021,0.003 9,0,0,0}
4	{37,14,0.000 095,0.010 1,271,1,105}
5	{1,9,0,0,0,0,0}
6	{100,13,0.000 163,0.028 6,756,5,76}
7	{72,1,0.000 007,0.021 7,0,0,0}
8	{1,4,0.000 006,0.002,6,0,1}
9	{128,271,0.002 161,0.026 5,809,8 793,1 396}
10	{72,25,0.000 709,0.017 0,102,1,4}

3.2　实验结果验证及分析

文中采用 K-means 算法进行聚类分析,将用户特征向量 $\boldsymbol{U}_i$ 中的 7 个特征值作为原始输入数据。首先,选取 DBI 指数(Davies-Bouldin 指数)作为评价聚类优劣的指标,进行聚类中心数的选取[14]。当 DBI 指数越小,所得到的聚类结果越好。如图 2 所示,当 $k=6$ 时,DBI 值最小,选取 6 作为聚类中心数。

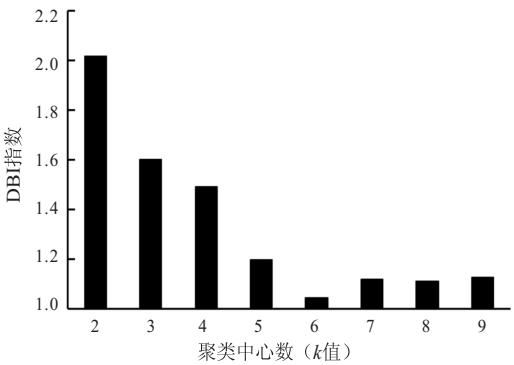

图 2　K-means 算法聚类中心数选取

运行 K-means 算法程序,对输入的用户特征向量反复训练和测试后,原始输入数据被聚类为 6 个不同的子类,如表 2 所示。

表 2　社区用户聚类情况

子类	用户数	百分比/%
C_0	28	0.07
C_1	2 957	7.35
C_2	36 484	90.69
C_3	551	1.37
C_4	7	0.02
C_5	204	0.51

分析社区用户聚类结果,可见子类 C_4、C_0 中成员的个数较少,占整个社区用户的比例分别为 0.02%、0.07%,符合筛选条件的第一条,这两个子类中必有一个子类能最大程度满足意见领袖的特征需求。为了进一步筛选出最适合的意见领袖集合,对于 6 个子类中的成员,分别统计出特征值{入度 v_1,出度 v_2,介数 v_3,特征向量中心性 v_4}的均值,如表 3 所示。

表 3　子类特征值{ v_1,v_2,v_3,v_4 }的均值

子类	入度 v_1	出度 v_2	介数 v_3	特征向量中心性 v_4
C_0	159.651 8	188.561 4	0.000 441	0.041 431
C_1	9.971 9	4.460 3	1.42E-08	0.000 067
C_2	0.317 4	0.958 3	1.24E-09	1.20E-07
C_3	78.991 7	21.204 3	0.000 005	0.009 779
C_4	315.058 7	274.792 5	0.001 296	0.194 850
C_5	72.838 6	117.164 5	0.000 037	0.023 051

表 3 中,子类 C_4 的特征值均值比其他子类更具有优势,其入度、出度的均值最高,说明 C_4 成员与其他用户间存在着大量的直接关注关系,其发表的言论会被其他用户直接接收,容易产生强大的影响力;介数均值远高于其他用户,说明 C_4 成员是信息交流过程中的重要通路,控制着信息的流动;特征向量中心性高,说明 C_4 成员在社会网络全局角度也具有很强的重要性。综上所述,子类 C_4 中的社区用户比其他用户更具有意见领袖的特征。

经过调查分析,子类 C_4 中 7 位用户的社会网络角色均为中心者或者桥梁(如表 4),其中 1、2、3、4、7 号用户是该社区的资深会员,5、6 号用户担任过版主,符合筛选条件的第三条。经过以上分析,说明子类 C_4 是该社区的意见领袖集合,其中 7 位成员均为意见领袖。

表 4　子类 C_4 中用户的社会网络角色

序号	ID	社会网络角色
1	1736360395	中心者
2	2159919174	中心者
3	2567257287	中心者
4	2803712700	中心者
5	3399780702	桥梁
6	5057069172	桥梁
7	5349058224	中心者

3.3　虚拟学习社区中意见领袖的作用

通过验证分析,发现社会网络中意见领袖多处于中心者或桥梁位置,占据着社会网络的优势位置。意见领袖通常具备丰富的学识、广阔的视野、先进的教育理念和专业的分析能力等,掌握更多的教育资本,可以将拥有的教学信息分享给其他用户,提供更多的课程资料,扩充社区的学习资源。通过意见领袖专业的信息加工和解释,可以有效地降低用户的理解难度,开拓视野,促进用户的专业发展。意见领袖并不是以独立个体的形式存在,他们是具有相似的社会网络拓扑特征和角色特征的用户集合。在虚拟学习社区的可持续

发展中,应该充分利用意见领袖集合对整个社区的影响力,协调社区中的各种资源,发挥其舆论引导能力,有效地控制社区中信息流动,营造良好的学习氛围,促进社区的自我组织和良性成长。

4 结束语

从社会网络拓扑结构角度,提取入度、出度、介数、特征向量中心性等特征参数,结合社会网络角色特性,得到社区用户特征向量,作为识别意见领袖的重要标准之一,通过 K-means 聚类算法并结合意见领袖的筛选条件,提出基于 K-means 算法的意见领袖识别模型。通过实例验证,获取的意见领袖集合具有很高的准确性,集合中各成员的社会网络位置和角色均符合意见领袖的特点。意见领袖群体注重各个领域间的融合,促使学习资源由单一的实用性转向多元化发展;引导用户利用闲暇时间学习,满足用户的日常学习需求;打破师生间时空分离的学与教,注重用户学习间的协作与共享。意见领袖以特殊的社会网络角色,促进大数据、人工智能等新技术与在线学习有效结合,推动虚拟学习社区的可持续发展。

参考文献:

[1] LAZARSFELD P F, BERELSON B, GAUDET H. The people's choice[M]. New York: Columbia University Press, 1948:434-445.

[2] 戴心来,刘聪聪. 基于结构洞理论的虚拟学习社区信息交互中介性研究[J]. 现代远距离教育,2018(3):21-28.

[3] 陈 远,刘欣宇. 基于社会网络分析的意见领袖识别研究[J]. 情报科学,2015,33(4):13-19.

[4] 刘 敏,胡凡刚,李兴保. 教师虚拟社区意见领袖的社会网络位置及角色分析[J]. 中国电化教育,2014(2):46-53.

[5] 张佳乐,张秀芳,张桂玲. 基于模糊综合策略的用户行为评估方法[J]. 计算机技术与发展,2017,27(5):138-143.

[6] 万新贵,李玲娟. 基于结构与属性的社区划分方法[J]. 计算机技术与发展,2017,27(8):97-101.

[7] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature,1998,393(6684): 440-442.

[8] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science,1999,286(5439):509-512.

[9] 吴 果,房礼国,李 中. 基于多指标综合的复杂网络节点重要性评估[J]. 计算机工程与设计,2016,37(12):3146-3150.

[10] JEONG H, MASON S P, BARABASI A L, et al. Lethality and centrality in protein networks[J]. Nature, 2001, 411(6833):41-42.

[11] JOY M P, BROCK A, INGBER D E, et al. High-betweenness proteins in the yeast protein interaction network[J]. Journal of Biomedicine & Biotechnology, 2005(2): 96-103.

[12] BONACICH P. Power and centrality: a family of measures[J]. The American Journal of Sociology, 1987, 92(5):1170-1182.

[13] 刘嘉琪,齐佳音,陈曼仪. 基于社会网络分析的意见领袖与在线群体影响力关系研究[J]. 情报科学,2018,36(11):138-145.

[14] 童莉莉,李荣禄,闫 强. 在线知识社群中的意见领袖识别模型研究[J]. 中国电化教育,2019(3):97-103.

[15] 梁云真. 在线交互网络中个体重要性评估及对学习成效的影响[J]. 中国电化教育,2018(11):94-102.

[16] 陈晓威,史昱天. 社会网络中关键节点的识别——基于符号网络的 PageRank 算法改进[J]. 数据分析与知识发现,2017,1(8):68-75.