基于 PageRank 算法的校园好友关系分析

王全民,赵亚康
(北京工业大学 计算机学院,北京 100124)

摘 要:随着校园信息化、数字化建设的普及,作为该建设进程重要组成部分的“校园一卡通”系统的建设也逐步深入,校园一卡通系统是校园信息化建设中信息采集的基础工程。利用校园一卡通系统采集的学生日常生活数据去分析学生的行为特点,挖掘出大数据中隐藏的价值,将在很大程度上帮助学校的日常管理并为学校领导提供相关决策依据。文中以学生校园生活中的一卡通消费数据为研究对象,首先应用高斯相似度函数构造相似度矩阵,系统地对学生间好友关系进行研究分析;然后应用 PageRank 算法找出疑似“孤立”学生或者是社交圈较小的学生群体,及早发现不善交友的学生群体,为学校的相关管理部门提供数据支撑,以便引导学生更健康的交友,构造和谐校园;最后通过实验进行了有效性验证。

关键词:一卡通;数据分析;高斯相似度;PageRank
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)01-0140-04
doi:10.3969/j.issn.1673-629X.2020.01.025

Campus Friend Relationship Analysis Based on Pagerank Algorithm

WANG Quan-min, ZHAO Ya-kang
(School of Computer, Beijing University of Technology, Beijing 100124, China)

Abstract: With the popularization of campus informatization and digitalization construction, the construction of “campus all-in-one card” system, which is an important part of the construction process, has been gradually deepened. The campus all-in-one card system is the basic project of information collection in the campus informatization construction. The daily life data collected by the campus one-card system are used to analyze the behavior characteristics of students and dig out the hidden value in big data, which will greatly help the daily management of the school and provide relevant decision-making basis for school leaders. The students' campus life of IC card consumption data is taken as the research object. First of all, we use the Gaussian similarity function to construct the similarity matrix, and carry on the research analysis to the friend relation between the students. Then we adopt PageRank algorithm to identify suspected “isolated” students or students with small social circles, timely detection of students who are not good at making friends, to provide data support for school related management department and guide students to healthier dating, constructing harmonious campus. Finally, the validity verification is carried out by experiment.

Key words: all-in-one card; data analysis; Gaussian similarity; PageRank

0 引 言

在社会信息化的大背景下,建设“智慧型”校园,不断推进以学校为主体的教育信息化进程,成为教育信息化的重要组成部分[1]。随着建设“智慧校园”理念的兴起,国内高校信息化建设开始推广,各高校普遍经历了校园网络建设阶段和以各类信息系统及相关资源建设为主的数字校园建设阶段。这意味着学生的在校生活“轨迹”可以很好地被记录下来,也为研究、分析并合理利用学校学生大数据提供了可能[2-3]。利用数字化校园产生的海量数据去分析学生的行为特点,挖掘出大数据中隐藏的价值,将在很大程度上有助于学校的日常管理并为学校领导提供决策依据。而校园“一卡通”是在大多数高校功能建设比较齐全的校园数字化系统,已经渗透到了学生生活中的方方面面,比如学生的餐饮、洗浴消费,图书借阅情况,门禁出入管理等,是学生校园生活轨迹的一个记录仪。

作为一个特殊的社会群体,当代大学生也面临许多问题,有对专业选择上的迷茫,有对新学习环境新学

收稿日期:2019-02-19 **修回日期**:2019-06-20 **网络出版时间**:2019-09-25
基金项目:国家自然科学基金(61272500)
作者简介:王全民(1963-),男,博士,副教授,硕导,CCF 高级会员(E200005398S),研究方向为网络与信息安全;赵亚康(1991-),女,硕士研究生,通讯作者,研究方向为网络与信息安全。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1520.010.html

习方式的不适,也存在人际交往能力上的问题以及对未来职业的选择问题等等。方方面面的压力很容易造成学生的心理问题。良好的交往行为,能建立良好的人际关系,提供良好的学习、生活氛围,通过好友间的交流及时疏通心理障碍,保证身心健康发展,促进专业知识的学习[4]。因此,挖掘学生校园中的生活数据,了解学生交友动态,引导和帮助学生积极健康学习、交友和生活对学生的成长意义重大[5]。同时找出疑似"孤立"学生,及时发现潜在的孤独症学生,并由辅导员及时的沟通和辅导,有利于构造和谐的校园生活。

文中通过对校园一卡通消费数据的研究分析,发现学生交友情况并找出疑似孤独症学生,为学校相关部门的管理提供决策依据。

1　一卡通消费数据分析模型

目前,大部分高校的"一卡通"功能建设趋于成熟,可以较为详细地记录到每个在校学生的生活轨迹信息。文中所用的数据为某高校脱敏后的一卡通消费信息,包括食堂、浴室、校园超市等消费地点的详细信息和消费时间以及金额。经过数据预处理从中删除跟研究无关的用户群及特征值,之后对数据进行格式转换,利用高斯相似度分析对学生的消费数据进行处理生成学生相互间的关系矩阵,可以通过这个关系矩阵找到某学生关联值较高的同学,推测其为该生的好友或者有相同作息习惯的同学,并为学生提供好友推荐;基于生成的关系矩阵,利用 PageRank 算法对学生进一步进行关联度分析,进而推测疑似孤独症学生。分析框架如图 1 所示。

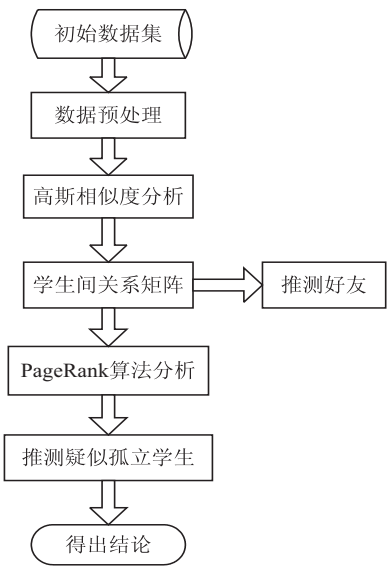

图 1　分析模型框架

1.1　数据介绍和预处理

文中需要分析的是学生的一卡通消费数据,会预先删除教职工的消费信息。其次将消费地点进行聚类后按顺序编号,将学生 ID 聚类后按顺序编号。并将原始数据里的消费时间转换为数字编码。处理后的数据格式如下:

第一列为学生分配的新 ID 编号,第二列为转化格式便于计算的消费时间,第三列为消费金额,第四列为相应的消费地点。处理后的数据集格式如图 2 所示。

1	2	3	4
2	4.3030e+04	1	7
2	4.3030e+04	2.5000	7
3	4.3027e+04	16	7
3	4.3026e+04	0	6
3	4.3027e+04	50	6
3	4.3031e+04	1.9000	7
3	4.3031e+04	5.5000	7
4	4.3022e+04	7	3

图 2　数据清洗结果

1.2　高斯相似度分析

高斯分布(Gaussian distribution)又称正态分布(normal distribution),也称"常态分布",而自然界中实例以及实例属性的分布很多情况下都是正态分布。显然文中所研究的学生每一次的消费时间概率也满足正态分布。那么采用高斯相似度函数对学生之间进行相似度计算,也将得到比较有效的实验结果。

高斯相似度(Gaussian similarity)表示的是两个零均值高斯分布间的相似程度,基于这种度量的算法称为高斯相似度分析(Gaussian similarity analysis, GSA)[6-7]。将每个学生看成一个点,将所有的点都相互连接起来,就构成一个校园学生关系图,同时所有的边的权重设置为相似度。即将任意两个样本点之间的高斯相似度形成权值矩阵。这种高斯相似度函数能够较好地反映实际中的相邻关系。

在所研究的数据集中,决定学生关联性的除了同一地点消费,还需要考虑消费时间,若是经常同时就餐,那么两位学生之间的消费时间差就会很小。学生 i 、j 的某次消费时间分别为 v_i 、v_j ,两人之间的消费时间差可表示为 $\| v_i - v_j \|$,则基于时间差的相似度函数为:

$$f(x) = e^{-\frac{(v_i - v_j)^2}{2\sigma^2}} \tag{1}$$

两位学生间的消费时间间隔越长, $\| v_i - v_j \|$ 越大,$f(x)$ 值越小,表示两人之间的关联度就越小。将每一次消费相似度进行累加,得到两个同学间的关联值。

1.3　PageRank 算法

通过高斯相似度分析,可以得到一个全校由学生间相似度作为权值的关系矩阵。可以通过 ID 信息检索得到所属学生"好友"推测。但仍需要从这个庞大的校园学生关系网中筛选出疑似孤立的学生,以为相关管理部分提供有用信息。

PageRank 算法是在 1998 年 4 月举行的第七届国际万维网大会上由 Sergey Brin 和 Larry Page 提出的[8]。随着对这种算法的深入学习,研究者渐渐地从对网页数据的分析扩展到了对人群数据的研究。Riquelme F 等[9]基于用户间的交互关系,通过改进 PageRank 算法建立影响力的评估方法;孙红等[10-11]根据微博用户之间的交流关系利用 PageRank 模型计算出微博用户的影响力;周飞等[12]用 PageRank 算法计算网络社区“知乎”用户的影响值;张欣等[13]结合专利的被引用次数和年限对原始的 PageRank 算法进行改进来识别核心专利。这些研究都是通过研究对象之间所存在的关联关系来分析个体在整个研究群体中的价值、影响力或者活跃值。最终找到活跃值最高的、影响力最大的。那么疑似“孤立”学生的选取也可理解为在校园关系网中活跃值最低的那位同学。就可采用该模型去挖掘疑似“孤立”学生。

PageRank 算法的最初目的是通过计算页面链接的数量和质量来确定网站重要性的粗略估计,创立之初是应用在 Google 的搜索引擎中,用来标识网页重要性的一种方法,即网页排名。它是根据网页之间的相互链接结构实现的。简而言之,如果一个网页可被其他很多个网页链接到,就说明这个网页比较重要,排名就靠前。算法的表达式为:

$$\mathrm{PageRank}_{(P_i)}=\frac{1-q}{N}+q\sum_{M_{(p_i)}}\frac{\mathrm{PageRank}_{(p_i)}}{|L_{(p_j)}|} \qquad (2)$$

其中,$p_1, p_2, \cdots, p_n$ 表示网页;$M_{(p_i)}$ 表示待研究页面 p_i 的页面链入数;$L_{(p_j)}$ 表示页面 p_j 的页面链出数;N 表示网络中的所有页面数量;$\mathrm{PageRank}_{(P_i)}$ 表示页面 p_i 的 PageRank 值,所有页面的 PageRank 值构成网络的 PageRank 向量;q 表示用户继续浏览该页面之后的页面的概率,通常概率值取 0.85。

文中将 PageRank 算法的思想用于校园学生疑似孤立的预测。同学间生成的具有相互关联关系的矩阵就类似于网络有向图,然后每个学生就如同要研究的页面,通过 PageRank 算法得到每个学生的 PageRank 值。与网页排序同理,如果一个学生可以被很多学生联系到,证明该生比较活跃,反则,该生社交范围极小,甚至有可能存在长期孤独的状况。

2 实验结果分析

实验分析了该高校一学期的学生一卡通消费数据。首先对每天刷卡时段进行聚类分析,根据刷卡时间分布将一天的消费时间分成早、中、晚三个时间段。图 3 为时间聚类结果,由此分布情况将消费事件大致划分为早餐 6-10 点;中餐 10-14 点;晚餐 14 点后,根据聚类可以将一天分为 3 个时间片。

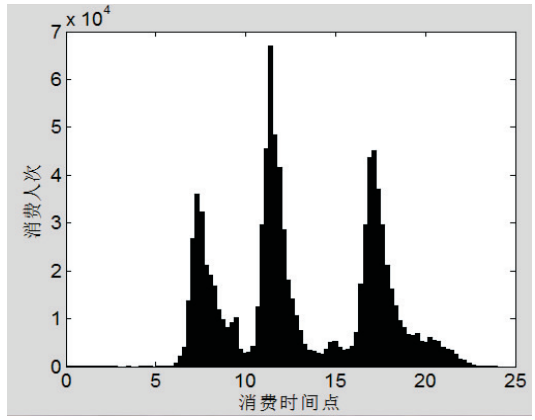

图 3 学生消费时间分布

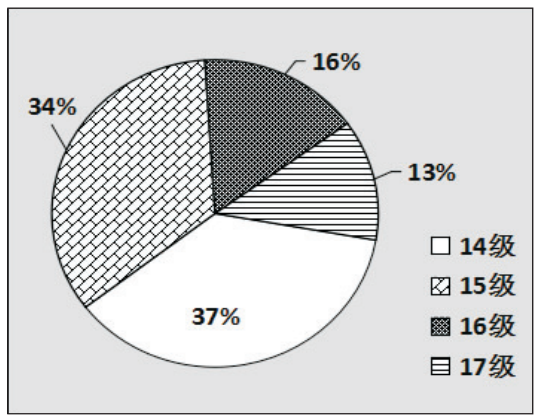

图 4 疑似孤立群年级分配比例

在活跃度排序结果中取活跃度最低的 300 名学生进行研究,除去 13 级延迟毕业的学生,活跃度较低群体的年级分布如图 4 所示。初步分析,高年级,特别是即将毕业的学生,由于找工作和实习不在校外的情况会较多,在所分析出来疑似孤立或存在交友问题的学生群里占得比例较大,低年级大部分时间都在校内上课,出现长时间单独出现的概率就会较少,在疑似孤立的学生群体中所占比例就会较小。实验结果与预期的理论分析结果完全吻合。

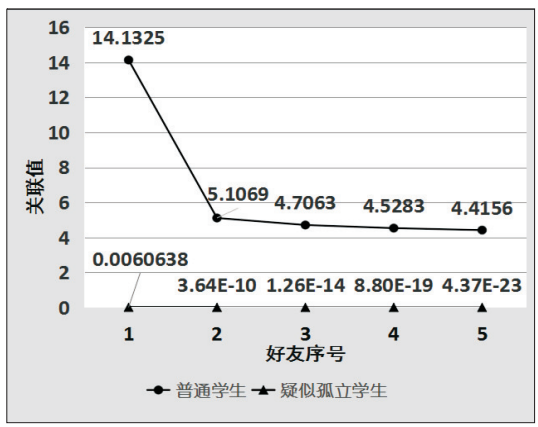

图 5 疑似孤立和正常学生好友情况分析对比

图 5 为一个普通学生与一个疑似孤立学生与“好友”之间关联值的比较。实验思路为分别取与这两位

学生关联值最大的五位同学,用 X 坐标表示“好友”,Y 坐标表示该学生与好友之间的关联度值,那么每个点由所研究学生与其好友的相似度值大小生成。可以明显看出,普通学生与其实验计算出的前五名“好友”之间的关联值较大,而“疑似孤立”学生与其对应的前五名“好友”的关联值极其小。这说明该学生与好友间共同消费行为较少,并且经常独自一人错峰消费。针对出现这种情况的学生,学校相关部门就需要深入了解一下这类学生的生活学习状态,针对有交友障碍的学生提供相应的疏导。

3　结束语

高斯相似度模型被广泛应用于基于相似度矩阵的聚类算法中。比如谱聚类算法,其主要思想就是对样本数据集生成的关联矩阵进行聚类,比起传统的 K-means 算法对数据分布的适应性更强[14-15]。文中的高斯相似度函数,将消费时间、地点看成空间上的一个点,每两个学生间的消费时间间隔较短的边的权值较高,这样可以有效地得出学生之间的关联程度。而在重要度、影响力排序中有颇多应用的 PageRank 算法可以综合计算出学生在校“影响力”,有助于分析出活跃度极低,疑似交友困难甚至“孤独症”的学生。

校园中的学生群体之间会形成一个巨大的关系图,为验证实验结果,在数据源所涉高校相关部门的帮助下对所分析的结果进行实体验证,随机选取该校学生进行访问。据调研学生验证,与其关联值较大的同学大多是其好友或是室友关系,但也存在生活习惯相近导致的“熟悉的陌生人”情况的发生,所以对数据一卡通消费数据的分析可以对有需求的同学提供一个好友推荐的平台,通过向其推荐与其生活习惯相近的同学来拓展该生的交际圈。

一卡通系统会源源不断地产生新的、大量的甚至是孤立无序的数据,需进一步采用科学合理的算法构建数学模型,找到数据之间的关联,为学校相关管理和学生的健康成长提供更多科学决策依据,这些还有待深入研究。

参考文献:

[1] 黄荣怀,张进宝,胡永斌,等. 智慧校园:数字校园发展的必然趋势[J]. 开放教育研究,2012,18(4):12-17.

[2] 黄　刚,刘　蓉,刘合富,等. 基于校园一卡通数据的人群画像分析[J]. 计算机与数字工程,2018,46(9):1881-1886.

[3] 甘　伟. 基于一卡通数据的高校学生行为分析与排名预测[J]. 现代计算机,2018(9):80-83.

[4] 马亚平. 当代大学生人际交往的问题归因及教育引导[J]. 教育与职业,2014(20):188-189.

[5] WAN L C. A research on the correlation of loneliness and social anxiety in college students[J]. Advances in Psychology,2016,6(4):391-397.

[6] WU Ji, WANG Zouying. A decision tree-structured algorithm of speaker adaptation based on Gaussian similarity analysis[J]. Chinese Journal of Electronics,2001,10(2):166-169.

[7] 吕　萍,王作英,陆　大. 基于高斯相似度分析的插值自适应算法[J]. 电子学报,2001,29(12A):1759-1761.

[8] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web[R]. Stanford: Stanford Digital Libraries,1998.

[9] RIQUELME F, GONZÁLEZ-CANTERGIANIP. Measuring user influence on Twitter: a survey[J]. Information Processing & Management,2016,52(5):949-975.

[10] 孙　红,左　腾. 基于 PageRank 的微博用户影响力算法研究[J]. 计算机应用研究,2018,35(4):1028-1032.

[11] 谢橙瞬,周莲英. 基于 PageRank 的微博用户影响力评估模型研究[J]. 信息技术,2018(5):75-78.

[12] 周　飞,高茂庭. 基于 PageRank 的网络社区意见领袖发现算法[J]. 计算机工程,2018,44(2):203-209.

[13] 张　欣,马瑞敏. 基于改进 PageRank 算法的核心专利发现研究[J]. 图书情报工作,2018,62(10):106-115.

[14] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[C]//Proceedings of the 14th international conference on neural information processing systems: natural and synthetic. Vancouver, British Columbia, Canada: MIT Press,2001:849-856.

[15] 李玲俐. 谱聚类算法及其应用综述[J]. 软件导刊,2016,15(7):54-56.