

基于时空关联的高校社会网络关系挖掘方法研究

李有增¹, 周 全¹, 蒋鸿玲²

(1 首都师范大学, 北京 100048; 2 中国航天科技集团有限公司 物联网技术应用研究院, 北京 100094)

摘 要: 以高校学生刷卡消费数据为分析对象, 定义了学生之间呈现出的时空关联关系, 设计了算法检测时空关联关系, 利用检测出的时空关联关系构建社会关系网络, 并对社会关系网络进行优化, 进而挖掘出学生之间的社会网络关系, 通过判断学生是否在同一个班级来验证挖掘出的社会网络关系, 实验表明本方法能够较为准确地挖掘出学生的社会网络关系。

关键词: 时空关联; 社会网络关系; 消费数据

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2018)12-0137-04

Research of Method on Mining Social Network Relationship in Colleges Based on Spatio-Temporal Correlation

LI You-zeng¹, ZHOU Quan¹, JIANG Hong-ling²

(1 Capital Normal University, Beijing 100048, China;

2 Internet of Things Institute, CASC, Beijing 100094, China)

Abstract: Based on the data of university students' consumption, spatio-temporal correlation between students is defined. We designs an algorithm to detect spatio-temporal association, and the social relations network is constructed by using the spatiotemporal relation detected, and the social network is optimized, and social network relationship between students is found out. The social network relationship is verified by judging whether the students are in the same class. Experiments show that this method can accurately extract the social network relationship of students.

Key words: spatio-temporal correlation; social network relationship; consumption data

1 引言

近几年来, 学术界掀起了对社会网络分析的研究热潮. 社会网络分析是对社会网络的关系结构及其属性进行分析的方法^[1], 主要关注参与者之间的关系等. 当前, 社会网络分析广泛应用于微博用户分析^[2,3]、研究者团体分析^[4,5]等方面. 对高校学生的社会网络研究较少. 而随着高校信息化建设进程的推进, 使得高校各类信息系统产生了大量的数据, 这些数据为挖掘学生社会网络关系提供了数据支撑. 通过对这些大数据的分析, 可以分析出学生行为的时间、空间关系, 进而挖掘出学生的友好关系及社会网络.

本文以高校学生刷卡消费数据为例, 研究了基于时空关联分析的社会网络关系挖掘.

2 消费大数据中的时空关联关系

本文分析了某高校的校园消费流水数据. 每条记录表示该学生账号的一次消费行为. 由于学生好友经常一起去食堂或者超市消费的概率较大, 因此可以根据学生是否经常同时一起消费, 来推断学生之间的友好关系, 进而挖掘出学生的社会网络关系.

选取学生消费数据学生账号(记为 s)、商户账号(记为 $chapmanID$)、POS 机编号(记为 $PosID$)、刷卡时间(记为 t).

从消费数据中, 本身并看不出学生的友好关系,

但通过时间和地点的记录,能够推断出学生刷卡消费的时空关联关系.

本文定义的时空关联关系,是指学生 U 和 V ,在较短的时间间隔内,在同一个地点刷卡,则称为学生 U 和 V 具有时空关联关系. 如果 2 个学生多次呈现时空关联关系,则这两个学生之间可能是好友,据此可以构建学生的社会网络关系.

3 时空关联关系检测

要挖掘学生的社会关系网络,首先要检测学生之间是否存在时空关联关系. 判断任意两个学生是否存在时空关联关系,依据他们的刷卡时间间隔 T 和刷卡地点 L . 如果两个学生 U 和 V 在同一个地点 L 刷卡,并且刷卡时间间隔 T 小于时间阈值 T_c ,则学生 U 和 V 可能有时空关联关系.

时空关联关系检测首先按照地点 L 将一卡通消费记录分为不同的时间序列,每个 L 对应一个刷卡记录时间序列. 对于每个地点 L_i ,对应的刷卡记录序列为:

$$\{ \langle s_{1i}, t_{1i} \rangle, \langle s_{2i}, t_{2i} \rangle, \dots, \langle s_{ni}, t_{ni} \rangle \}$$

(1)

式中, $\langle s_{1i}, t_{1i} \rangle$ 是 L_i 对应的时间序列中的第一个记录; s_{1i} 是学生账号; t_{1i} 是该记录的刷卡时间.

针对每个时间序列,如果两个学生的刷卡记录的时间间隔小于时间阈值 T_c ,则视为这两个学生具有时空关联关系. 图 1 是时空关联关系检测示意图,横坐标表示刷卡时间,纵坐标表示刷卡地点. 对于每个地点 L 的刷卡记录序列,包括学生账号 s 、刷卡时间 t ,检测时空关联关系算法的步骤如下:

- (1) 将时间序列的第一条记录 $\langle s_1, t_1 \rangle$ 作为当前记录 $\langle s_p, t_p \rangle$,第二条记录 $\langle s_2, t_2 \rangle$ 作为待检测记录 $\langle s_q, t_q \rangle$,如果 $t_2 - t_1 \geq T_c$,则跳过该条记录,从第 2 条记录作为当前记录 $\langle s_p, t_p \rangle$,下一条记录作为待检测记录 $\langle s_q, t_q \rangle$,重新判断;如果 $t_q - t_p < T_c$,则执行(2);
- (2) 如果 $s_p \neq s_q$,则将 $\langle s_p, s_q \rangle$ 加入到时空关联关系队列中;否则,则执行(3);
- (3) 当前记录 $\langle s_p, t_p \rangle$ 不变,将下一个记录作为待检测记录 $\langle s_q, t_q \rangle$,如果 $t_q - t_p < T_c$,则执行(2),否则将当前记录的下一条记录作为最新的当前记录,继续执行(2);
- (4) 直到所有记录都检测完毕为止.

在上述算法中,对于时间,比较容易判断时间 T 是否小于阈值 T_c ,但对于空间,判断两个刷卡记录

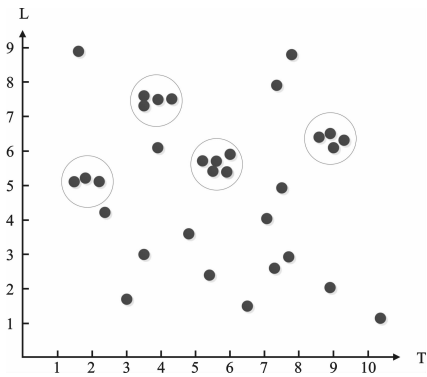


图 1 时空关联关系检测示意图

是否在同一个地点较为困难. 由于学校有多个校区,在消费数据中并没有体现校区,因此需要从数据中挖掘出校区信息. 不同校区的学生之间呈现频繁时空关联关系的概率较低,同一个校区的学生之间呈现频繁时空关联关系的概率较高,因此,要对消费空间进行聚类分析,并选取其中一个校区的学生刷卡消费数据作为分析对象.

选取了其中一个校区后,需要对校区内的刷卡地点进行区分. 比如在同一个食堂同一个楼层,不同的窗口消费,虽然在不同的 POS 机上刷卡,但实际上是同一个地点. 因此,在对地点 L 进行聚类时,也考虑了时间的因素,如果一个学生一个食堂刷卡,前后两次刷卡记录的时间间隔不会太长. 通过对 L 聚类,进而判断哪些 L 同一个地点,从而确定是否在同一个地点消费.

4 社会关系网络挖掘

检测出时空关联关系后,通过构建社会关系网络来挖掘学生间的社会网络关系. 本文方法中分为两步,第一步构建初始社会关系网络,第二步对社会关系网络进行优化.

(1) 构建初始社会关系网络

首先,构建一个初始的学生社会网络. 顶点表示学生,边表示学生之间的时空关联关系,边的权值表示两个学生呈现时空关联关系的次数. 构建步骤如下:

第一步,初始状态下,该网络为空,即没有节点和边. 当检测出学生 U 和 V 呈现时空关联关系时,则向网络中添加节点 U 和 V ,并将 U 和 V 之间的边 $E(U, V)$ 添加至网络中,同时将该边的权值 $w_{u,v}$ 置为 1.

第二步,当检测出新的时空关联关系对 $\langle U, V \rangle$ 时,如果当前网络中没有节点 U 或 V ,或者没有

这两点之间的边,则添加 U 或 V 到网络中,并添加这两点之间的边 $E(U, V)$;如果当前网络中存在节点 U 和 V ,并且这两点之间存在边 $E(U, V)$,则将该边的权值 $w_{u,v}$ 加 1.

第三步,扫描该时间段内的所有刷卡记录,将具有时空关联关系的学生找出来,并添加到学生社会网络中,初始社会网络即构建完成,可用矩阵 W 表示学生的初始社会网络,如公式(2)所示:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \quad (2)$$

式中, $w_{u,v}$ 为边 $E(U, V)$ 的权值,在初始社会网络关系中即为学生 U 和 V 呈现出时空关联关系的次数.

(2) 社会关系网络优化

在构建出的初始学生社会关系网络中,由于将所有具有时空关联关系的学生都加入到网络中,因此会存在大量偶然因素出现时空关联关系的学生,他们可能并不是好友关系,因此要将这类边过滤掉.

如果某条边的权值较小时,说明该边对应的两个学生呈现时空关联关系的次数很少,很可能是偶然因素出现的,而并不是真正的好友关系.因此将边的权值小于 k 的边过滤,然后再进行后续分析处理.

两个学生 U 和 V 之间的好友关系,不仅仅要考虑 U 和 V 之间呈现的时空关联关系次数,还要考虑与其他人的关联关系.如果学生 U 与较多其他人都呈现出时空关联关系,那么 U 和 V 之间的关系就不是那么密切,相反,如果 U 与其他人呈现的时空关联关系较少,而与 V 呈现的时空关联关系较多,那么 U 和 V 关系更密切,互为好友的概率更大.因此,本文在关系网络优化时,定义 U 和 V 之间的交际系数 $r_{u,v}$,如公式所示:

$$r_{u,v} = \frac{w_{u,v}}{(\sum_i^N w_{u,i} + \sum_j^N w_{v,j}) * (S_u + S_v)}$$
$$w_{u,i} \geq k, w_{v,j} \geq k \quad (3)$$

式中, $w_{u,v}$ 是边 $E(U, V)$ 的权值; N 是网络中节点个数; S_u 和 S_v 分别表示和节点 U 和 V 有时空关联关系的节点个数.当 U 和 V 都只有对方一个好友时,通过公式计算 $r_{u,v}$ 的值为 0.25,是最大值.

交际系数 $r_{u,v}$ 反应了学生 U 和 V 之间关系的紧密程度.这种度量方式不仅考虑了 U 和 V 之间的紧

密程度,还考虑了 U 和 V 与网络中其它节点的紧密程度,因此不会因为与多个节点有时空关联关系的“活跃”节点,而造成假性好友关系的出现.如果 $r_{u,v}$ 越大,则说明学生 U 和 V 的好友关系越紧密.

5 实验

本文选取了某高校一段时间的消费流水数据进行分析,其中学生账号数量为 2169,消费流水数据量为 109629 条.

从消费数据中提取出两个字段 $\langle \text{chapmanID}, \text{PosID} \rangle$ (记为地点 L) 来确定刷卡消费的空间.首先计算每个学生在所有 L 上是否有刷卡记录,如果两个消费地点 L_i 和 L_j 上刷卡的学生重复性较高,则说明 L_i 和 L_j 在同一个校区.然后根据是否有相同学生在 L 上刷卡来计算所有 L 的相似性,形成所有 L 的相似性矩阵.由于原始数据存在一些噪声,本文对原始数据进行过滤,重新计算过滤后的 L 之间的相似度,以及对 L 进行聚类后的结果如图 2 所示.

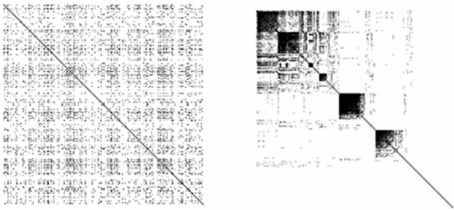


图 2 过滤后的 L 的相似性矩阵及 L 的聚类结果

从图 2 右图可以看出, L 可以被划分为 6 个区域,图中前 4 个区域能够形成一个较大的区域,因为这 4 个校区距离比较近,学生可能去相近的校区刷卡消费,该结果符合本文所分析的高校的校区个数.

为了检测学生之间的社会网络关系,本文选取其中一个校区作为分析对象.从图 2 右图可见,前 4 个校区距离较近,区分度不大,第 5 个校区与其他校区相聚较远,比较独立,因此本文后续分析以第 5 个校区为分析对象进行分析.

对上述第 5 个校区的 L 进行分析,同一个食堂同一楼层可以认为是在同一个地点消费,但同一个食堂同一个楼层刷卡时,其刷卡记录中 L 可能是不同的,因此要对 L 进行聚类,找出楼层相同的 L .

图 3 是对 L 的相似性矩阵及其聚类的结果.从图中可以看出三个明显的簇,实际中该校区只有一个食堂,食堂有 3 个楼层,与实际情况相符.这三个簇的刷卡记录占了 85%,因此本文的实验针对这三个簇的 L 对进行分析.

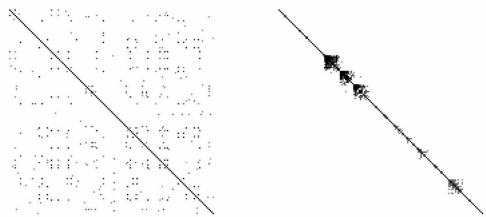


图3 对第5个校区L的相似性矩阵及其聚类结果

为了更清楚看出 $w_{u,v}$ 的变化趋势,图4采用了
对数形式, $w_{u,v}$ 的分布随着 $w_{u,v}$ 的增大逐渐降低.

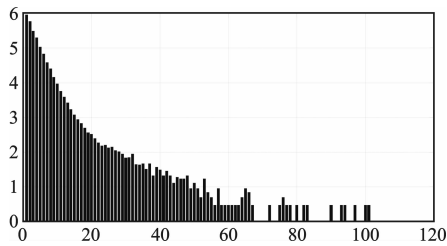


图4 $w_{u,v}$ 值的分布情况(纵坐标取对数)

社会关系网络构建完成后,要对网络进行优化,
首先过滤掉 $w_{u,v}$ 小于 k 值的边.不过滤时,网络的平均
度分布为1 068.89.在正常的校园环境中,平均一个
人拥有的好友个数不可能这么多,因此要对网络
进行优化,通过 k 过滤掉一些噪音边.实验中, k 值
分别取11—20之间的值时,学生社会关系网络的平
均度分布(AvgDegree)情况如表1所示.

表1 k 取不同值时的平均度分布

k	11	12	13	14	15	16	17	18	19	20
AvgDegree	10.8	8	6.1	4.8	3.9	3.3	2.9	2.6	2.3	2.1

从表1可以看出,随着 k 值的增大,平均度分布
不断减少.当 k 取[11,20]时,AvgDegree在2到11
之间,符合一个正常校园环境中,学生平均拥有的好
友个数.

过滤掉噪音边后,要计算学生U和V的交际系
数 $r_{u,v}$. $r_{u,v}$ 取不同值,好友关系级别不同.从公式(3)
可知, $r_{u,v}$ 越大,好友关系越密切.实验中,分别计算
了 $r_{u,v}$ 大于0、0.1和0.2时,检测出的好友关系对中
同班级的比例(sameClassRate).图5为 k 取[11,20]
时, $r_{u,v}$ 分别大于0、0.1和0.2时的同班级比例.

从图5中可以看出,当 $r_{u,v}$ 大于0时,随着 k 的
增大,同班级比例呈增大趋势.当 $r_{u,v}$ 大于0.1和0.
2时,同班级比例变化不是很明显.实验中以 $k=15$

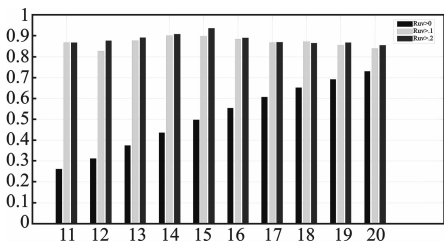


图5 同班级比例

为例,检测出的好友关系对有6 066对.从中随机选
取了8对好友关系作为样例,根据学生账号s,找出
所对应的学生的基本信息.

6 结束语

本文以高校学生消费流水数据为例,提出了一
种基于时空关联的社会网络关系挖掘方法.通过分
析消费数据中的时间和空间等来检测学生之间的时
空关联关系,通过时空关联关系进一步挖掘社会网
络关系.本文通过是否为同一个班级的学生来验证
好友关系.实验结果表明,本文方法能有效挖掘出学
生的社会网络关系.未来的研究中,将会把相关理论
方法应用在更广泛的数据中,在更高的维度挖掘出
社会网络关系.

参考文献:

[1] Furht B. Handbook of Social Network Technologies and Applications[M]. Springer US, 2010.
[2] 徐志明,李栋,刘挺,等. 微博用户的相似性度量及其应用[J]. 计算机学报, 2014, 37(1):207-218.
[3] 毛佳昕,刘奕群,张敏,等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014(4):791-800.
[4] 谢晓芹,韩帅,陈敏,等. 基于社会网络的团队生成方法研究[J]. 计算机学报, 2017, 40(3):712-728.
[5] 陈绍宇,宋佳兴,刘卫东,等. 关系网格:一种基于小世界模型的社会关系网络[J]. 计算机应用研究, 2006, 23(5):194-197.

作者简介:

李有增 男,(1962-),硕士,副研究员.研究方向为智慧校园、智慧教育、人才培养.
周全 男,(1982-),硕士,助理研究员.研究方向为智慧校园、智慧教育、人才培养.
蒋鸿玲(通讯作者) 女,(1986-),博士,高级工程师.研究方向为智慧校园、大数据. E-mail: hangtian_jhl@163.com.