软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(6):1819−1834 [doi: 10.13328/j.cnki.jos.005824]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向签到日志的用户行为模式交互探索*

李丛敏[1], 李 杰[1], 张 康[2], 陶文源[1]

[1](天津大学 智能与计算学部,天津 300354)

[2](The University of Texas at Dallas Computer Science Department, USA Texas 75080)

通讯作者: 李杰, E-mail: jie.li@tju.edu.cn

摘 要: 签到日志记录了用户对于某类设施的使用情况,从中发现用户行为模式,在很多领域如精确广告投放、犯罪团伙发现等方面都具有非常广泛的应用价值.但是,发现过程却较为困难,主要因为:(1) 日志数据体现为长时间序列且含有噪声,导致数据在高维空间分布较为稀疏,影响模式提取的准确性;(2) 行为模式往往与不同的时间尺度相关;(3) 多样的参数选择空间以及数据处理方式使得传统的机器学习方法很难获得可信且易于理解的行为分析结果.提出一种面向签到日志的用户行为模式交互探索的方法,该过程采用动态子空间策略,动态改变用于分析相似行为模式的时间片,从而减少人为设定参数对于分析结果的影响.方法集成了一个可视分析工具以支持该过程,利用该工具,分析人员可以实时了解方法每一步发现的模式,及时调整分析过程、直观理解和验证分析结论.包含了一个基于真实数据集的案例分析和一个来自不同领域专家的评审,其结果验证了方法的有效性.

关键词: 签到数据;群体行为模式;子空间探索;可视分析;交互探索

中图法分类号: TP181

中文引用格式: 李丛敏,李杰,张康,陶文源.面向签到日志的用户行为模式交互探索.软件学报,2019,30(6):1819−1834. http://www.jos.org.cn/1000-9825/5824.htm

英文引用格式: Li CM, Li J, Zhang K, Tao WY. Interactive exploration of behavior patterns from check-in logs. Ruan Jian Xue Bao/Journal of Software, 2019,30(6):1819−1834 (in Chinese). http://www.jos.org.cn/1000-9825/5824.htm

Interactive Exploration of Behavior Patterns from Check-in Logs

LI Cong-Min[1], LI Jie[1], ZHANG Kang[2], TAO Wen-Yuan[1]

[1](College of Intelligence and Computing, Tianjin University, Tianjin 300354, China)

[2](Computer Science Department, The University of Texas at Dallas, Texas 75080, USA)

Abstract: Check-in logs record how users access certain facilities. Discovering users' behavior patterns via logs has a wide range of applications, such as targeted advertising, criminal activity detection, etc. However, the discovery process is complex and challenging, due to the following reasons. (1) Log data is usually of long-term and contains noise, with sparse distribution of data in high-dimensional space. (2) Behavior patterns always relate to different time scales. (3) The variety of parameter selections and methods of data processing make traditional machine learning approaches difficult to obtain credible and understandable behavior analysis results. This study proposes an interactive approach to exploring behavior patterns from check-in logs. The process uses a dynamic subspace strategy which changes the time slices to analyze similar behavior patterns dynamically. The strategy reduces the effect of setting parameters artificially on the analytical results. The proposed approach integrates a visual analytical tool to support the process. Through visualization, analysts could understand the patterns found in each step-in real time, adjust the analysis process, comprehend and verify the results intuitively. The paper also presents a case study based on a real data set and a review of experts from different fields. The results confirm the effectiveness of the approach.

* 基金项目: 国家自然科学基金(61602340, 61572348); 国家重点研发计划(2018YFC0831700, 2018YFC0809800)

Foundation item: National Natural Science Foundation of China (61602340, 61572348); National Key Research and Development Program of China (2018YFC0831700, 2018YFC0809800)

收稿时间: 2018-05-26; 修改时间: 2018-09-17, 2018-12-19; 采用时间: 2019-01-17; jos 在线出版时间: 2019-03-27

CNKI 网络优先出版: 2019-03-27 16:54:32, http://kns.cnki.net/kcms/detail/11.2560.TP.20190327.1654.012.html

Key words: check-in data; group behavior pattern; subspace exploration; visual analytics; interactive exploration

在很多领域中,用户签到日志是一种常见的数据类型,这类数据直接记录了用户对于某种设施的使用情况,常见的使用场景包括宾馆入住记录、网吧上网登记和在线系统的登录日志等.从这类数据中挖掘出具有相似行为的用户群体并分析该群体的行为模式,在信息服务、在线搜索服务、医学诊断、网络安全、商业营销等方面具有非常重要的作用.

现有的方法常常基于统计对用户时序行为进行分组,然而由于日志数据、行为模式、统计方法等方面的限制,已有的方法往往很难获得准确且易于理解的结果,其挑战主要体现在:

- 签到日志的高维稀疏性:签到日志往往时间跨度较长,用户在不同时间点使用设施,形成了时间分布上的高维向量,使得行为特征在高维空间分布较为稀疏.因此,无论是传统的聚类等机器学习算法、社区发现算法,还是推荐系统算法往往都不能直接得到高质量的具有相似行为模式的群体;
- 行为模式具有多样性且与时间层次紧密相关:数据集中往往同时存在多个行为模式,且行为模式可能发生在任何特定的时间尺度上,例如白天、夜晚、假期、春季、下雨天等.多个行为模式交叉在一起,对其发现和理解带来了较大的挑战;
- 统计方法对参数和数据分布有要求:现有的分析方法往往需要事先设定某些先验参数[1],或者要求数据符合某些特定的分布.这些参数设定和前提假设往往需要复杂的数据验证,并且依赖分析人员对数据的理解和经验.这也加大了群体行为模式发现的难度.

越来越多的研究人员采用可视分析方法分析群体模式,然而这些工作更多地只是展示统计分析的结果,缺少相似行为模式发现的过程.与其不同,本文设计了动态探索群体模式的可视分析方法,主要贡献如下.

- 定义了一个动态迭代探索过程.该方法以一种"顺藤摸瓜"的迭代方式将用户逐步加入到群体中.本文引入了信息熵,动态地获得具有较好行为区分度的时间子区间,并探索在此区间内具有相似行为特征的群体;
- 开发了一个支持以上迭代方法的可视分析工具.通过该工具,使用者可以交互地控制分析过程,直观地理解和验证所获得的群体行为模式,并根据可视化反馈,实时主动地调整分析过程;
- 对群体在不同时间尺度上的统计和关联进行分析,并减少由于数据偶然性带来的噪声,帮助使用者对群体行为模式进行理解.通过迭代前后群体行为模式的对比,验证本文方法的有效性.

1 问题描述

1.1 数 据

签到日志包含大量人员在较长时间上的行为记录,其结构主要包括两方面信息,即设施使用时间和用户的基本信息.表 1 展示了某网吧 3 个用户的上网记录,其中,身份证号表示个人信息,上线和下线时间反映其在网吧上网的时间区间.大部分用户只是固定或不固定地、有限度地使用设施,因此行为记录在时间尺度上具有较为明显的稀疏性.用户签到日志数据,时间的跨度很广,并且绝大多数用户使用某设施的起止时间不同,造成用户使用设施时间没有对齐,描述用户行为的时间结构不统一,这给行为模式的探索造成了困难.

Table 1 User check-in logs in net bar

表 1 网吧用户签到日志

身份证号	姓名	上线时间	下线时间
3607261996****4715	A	2016-11-05 09:00:00	2016-11-05 11:10:35
3607261996****4715	A	2016-11-08 18:30:23	2016-11-08 23:20:22
3625021992****267X	B	2016-11-02 20:25:56	2016-11-03 00:01:02
3625021992****267X	B	2016-11-06 21:22:52	2016-11-07 01:09:42
6301021990****2511	C	2016-11-04 10:34:13	2016-11-04 12:24:28
6301021990****2511	C	2016-11-05 10:44:29	2016-11-05 14:12:41
…	…	…	…

1.2 任 务

本文为数据分析人员提供了可视分析工具,帮助分析和理解签到数据中存在的群体行为模式.如果某些用户经常同时使用设施,则可认为这些用户属于一个群体,并具有相似的行为模式.本文的主要任务是找到频繁在某些时间片上签到的用户群体.以网吧数据为例,有些用户经常在周末上网,有些则经常在晚上或凌晨上网.了解这些群体行为,有助于获取群体行为习惯,推断其身份,有针对性地开展行业应用.同一用户群体可能同时存在多种行为模式,这给模式的发现和理解造成了困难,因此,本文将这一探索过程分为 3 个不同层次的任务.

- T_1:行为特征可视化.直观地可视化数据中个体和群体的设施使用行为.该任务是后续分析的基础,分析人员可以据此选择具有特定行为特征的用户,并交互探索与其具有相似行为的群体.所展示的行为特征应包括基本的行为时序特征,如周期性、趋势、高频使用阶段、行为的统计指标以及不同用户间的行为相似程度;
- T_2:用户群体发现.寻找频繁共同使用设施的群体.由于签到数据的稀疏性和用户行为的偶然性,该过程往往受到数据噪声的影响.方法应该提供必要的数据处理和过程控制,减少数据噪声带来的影响.探索过程应可视化并具有较好的交互能力,使分析人员实时全面地理解和调整分析过程.发现过程应该减少参数影响,参数能随中间结果的变化而变化,发现过程也应是参数不断优化的过程;
- T_3:群体行为模式理解.在发现共同行为模式的基础上,应进一步分析该模式在时间尺度上的分布特征.其目标是理解所发现模式的实际物理意义,辅助推断群体的行为习惯和可能的身份,并据此开展实际行业应用.方法应该能从不同的时间尺度(如周、天、小时等)对群体行为特征进行理解,能将发现的群体与初始数据进行对比,验证本文方法的有效性.

2 相关工作

签到日志在很多领域都具有非常重要的分析应用价值.有些研究通过对签到日志的分析,优化资源配置.Peng 等人[2]通过社交媒体签到日志,检查出租车高需求区域,改善出租车资源分配.Li 等人[3]通过行李托运日志,分析用户行李登机行为和行李需求特征,优化机场资源配置.有些研究通过对用户使用产品的行为和需求模式分析,改进产品设计.如 Leemans 等人[4]通过分析用户的软件事件日志得到在现实生活中用户操作软件系统的过程,从而发现软件存在的问题.Liu[5]和 Chen[6]等人通过分析社交媒体签到日志,为用户推荐其感兴趣的主题.一些研究通过对用户商店签到日志的分析,得到用户的消费模式,从而改善营销策略.如 Chen 等人[7]通过分析顾客使用商场 WiFi 的签到日志,分析时间对顾客选择商场偏好的影响,从而基于时间为顾客推荐商场.Doi 等人[8]通过商店签到日志的分析,得到消费者的偏好,改进营销方案.还有一些其他的研究在不同的领域中也具有重要的意义.例如,Yang 等人[9]通过分析游客使用社交媒体的签到日志,分析游客的旅游路线,帮助人们做出经济有效的旅行决策.Liu 等人[10]通过分析用户使用出租车的日志,找到放置广告牌的最佳位置.以上研究更偏重于对个人或整体签到日志的统计分析,很少有通过分析用户间相似度寻找分组行为模式的研究.

群体行为的发现往往根据个体之间的相似度,使用分组算法对数据分组.很多研究使用聚类的方法来寻找具有相似行为模式的分组.Frhan 等人[11]提出了模式聚类和关联聚类的方法来寻找用户行为相似的群体.Lei 等人[12]使用聚类方法寻找微博用户的行为模式.这些方法往往对数据分布有要求且较依赖参数.各类社区发现算法也是经常采用的方法.Bron 等人[13]用算法生成组,生成候选用户集,删除不符合派系定义的候选用户,算法的终止条件是生成了一个完全连通的图.Liu 等人[14]提出了一种基于网络连接强度的重叠社区发现算法,该算法从重要性最高的用户逐步扩展,直到满足终止条件.He 等人[15]使用 SimRank 相似性度量和 NMF 模型发现复杂网络中的社区.Zhou 等人[16]使用基于主题感知特性的隐式关系和基于互动行为的显示关系对动态社交用户网络模型进行扩展和完善,从而发现更为合理的社区.推荐系统是另一类典型的群体行为模式发现方法.Rohit 等人[17]使用基于潜在语义索引的推荐系统算法来寻找相似类型的博客.Maake 等人[18]利用选择性驱动的推荐系统算法为用户推荐需要的论文.Yi 等人[19]分别使用基于图形数据库和基于深度学习的方法为用户推荐同类型的电影.Hariadi[20]基于混合属性和个性的推荐系统算法为用户提供相关的书籍.这些分组算法往往使用用户间

的相似度分组,但相似度通常存在噪声和稀疏数据,且相似度的度量方法也会影响分组结果,因此这些算法的准确率不高.不仅如此,预设的参数也无法根据中间结果实时调整.

越来越多的研究采用可视分析探索用户行为模式.Liu 等人[10]通过热图表示用户在空间的行为模式.Saas 等人[21]将热图、树状图、折线图结合,分析游戏玩家的行为模式.Krueger 等人[22]使用围巾图和时空立方体图揭示访问者序列模式.Li 等人[23]使用柱状图、平行坐标图等视图发现犯罪数据的多个属性模式.Zhang 等人[24]将热图和饼图结合,展示在公共交通系统中用户的流动模式.Li[25]通过词云、时间流、地图等视图寻找文本时空模式.Zhao 等人[26]利用边缘重叠度概念,减少 MSV 的视觉混乱,同时保留网络通信的时变特征,分析动态网络的变化模式.Zhou 等人[27]基于地图发现移动学习者的行为模式.Chen[28]通过词云、平行坐标图来分析社交媒体中重大事件,将分析关联模式,将模式形成故事.Wei 等人[29]通过自组织映射将网络点击流投影到二维区域,研究用户浏览网页模式.Zhao 等人[30]通过多维可视评估,使用模糊聚类寻找群体行为模式.Li 等人[31,32]分别通过地图、散点图等多视图协同寻找共现模式和气象变化模式.这些研究更多是对分析结果的展示,用户无法直观了解探索过程.

综上所述,签到日志的研究偏重于统计分析,鲜有根据用户相似度寻找群体行为模式的研究.而关于分组算法的研究大多因数据的稀疏性,分组结果的准确率不高.同时,关于行为模式的可视化研究大多是对分析结果的展示,使用者无法了解探索分析过程.为了解决以上问题,本文使用动态子空间策略迭代探索具有相似行为模式的群体,并通过可视化工具使用户可以实时地控制探索过程,从而直观地理解和验证所获得的群体行为模式.

3 分析流程

根据数据特征和任务,本文设计了一个发现群体的迭代探索方法和一个支持迭代过程的可视分析工具.如图 1,本文输入签到数据,经过迭代和可视分析处理,输出找到的群体和群体行为模式.

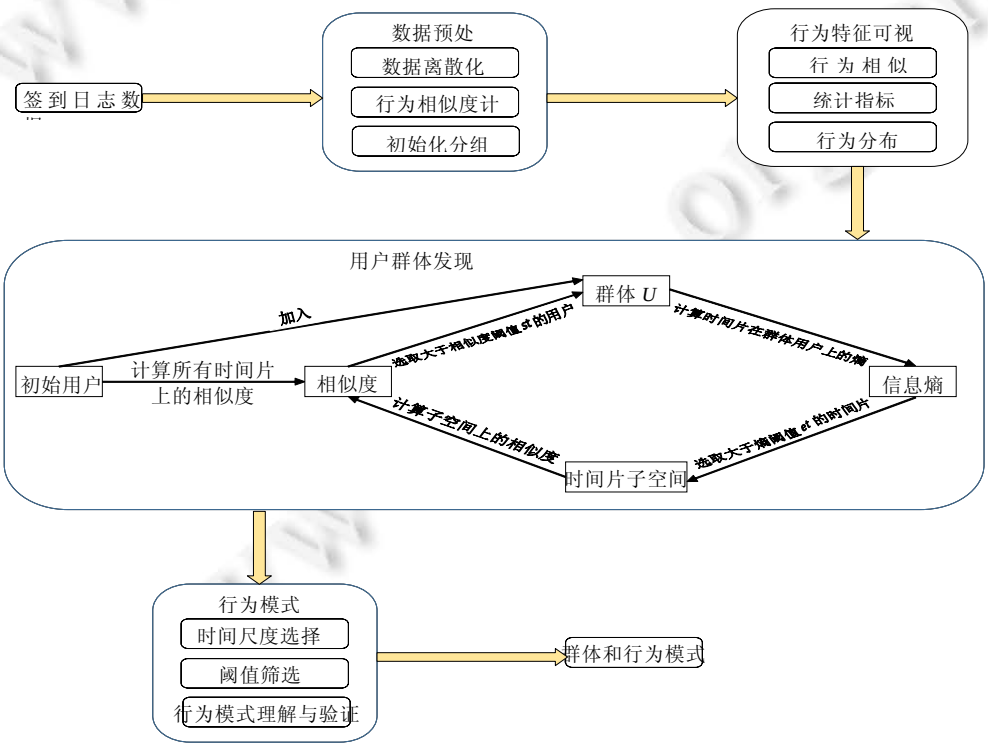


Fig.1 Analysis process

图 1 分析流程

分析流程分为如下 4 部分.

(1) 数据预处理

签到日志是用户使用设施的记录,不同用户使用设施的时间各不相同并且数据量很大,这给群体的寻找带来了困难.为了解决该困难,本文对数据进行预处理,将原始记录处理为时间对齐,结构统一地用于描述用户行为的特征向量,具体方法如下.

本文把每个用户的签到时间对应一个长度统一的离散化的签到时间片集合.首先,本文把签到时间划分成 m 个连续的时间片序列 $T=(t_1,t_2,t_3,...,t_m)$.为了便于计算,时间片采用固定长度,其时间跨度可以根据分析目标进行灵活设定,时间跨度越小,会得到越精确的时间片序列,但是时间片序列也会变长、更加稀疏,同时也增大计算复杂度.较长的跨度可能产生错误的行为记录,因此,使用者要根据数据特点灵活设定时间跨度.本文为每个用户生成一个签到时间片集合,用户 i 在时间片序列 T 上对应一个签到时间片集合 $c_i=(c_{i1},c_{i2},c_{i3},...,c_{im})$.如果用户 i 在时间片 t_j 内使用某设施,则向量对应位置的 c_{ij}=1;否则,c_{ij}=0.例如,本文将用户上网数据的时间跨度设为 30 分钟,因为根据统计大部分的有效数据,用户连续上网时间都超过了 30 分钟.如果用户 a 在 8:40~10:10 和 12:00~13:00 上网,那么生成的签到时间片集合如图 2 中的 c_a 所示.

2016-11-15

	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	
C_a	…	0	1	1	1	1	0	0	0	1	1	0	0	0	…
C_b	…	0	1	1	0	1	1	1	1	1	1	0	0	0	…

Fig.2 Set of check-in time slices for user a and user b

图 2 用户 a 和用户 b 的签到时间片集合

本文根据签到时间片集合计算两两用户之间的行为相似性,从而判断两个用户是否属于一个群体.行为相似性是后续迭代探索的计算依据.如果两个用户使用设施重合度较高,即签到时间片集合中“1”的重合度较高,则认为这两个用户具有很相似的行为.令 c_a 和 c_b 分别为用户 a 和 b 的签到时间片集合,a 和 b 之间的行为相似度定义见公式(1):

$$s_{ab}=\frac{|c_a\cap c_b|}{|c_a\cup c_b|} \tag{1}$$

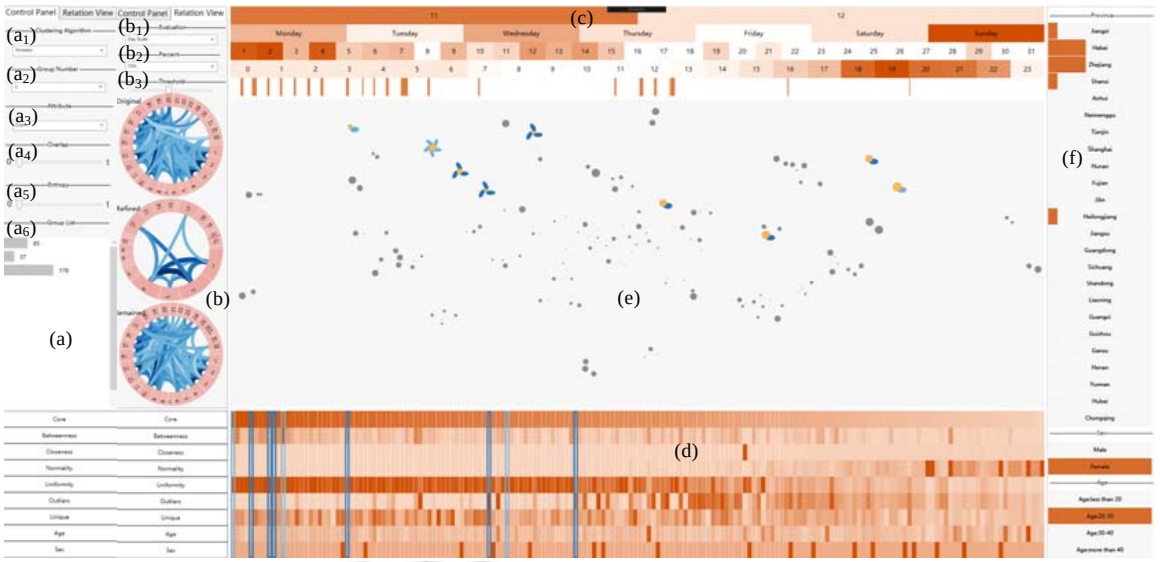
以图 2 为例,c_a 和 c_b 分别为 a 和 b 的签到时间片集合,则 $c_a\cap c_b$=5,$c_a\cup c_b$=9,s_{ab}=0.556.

为了提高后续的分析效率,在数据初始化时,可以依据用户在全部时间区间内的行为特征进行初始分组.初始分组可使用现有的聚类算法.聚类算法需设置较小的簇个数,以保证具有相似行为特征的用户不被分开,本文将初始化分组得到的组称为初始组,如图 3 中(a_6)有 3 个初始组.后续分析可以针对初始化得到的其中一个组开展.这一过程是可选的,当数据量不大或用户行为不存在明显差异无法得到清晰的簇时,可不进行初始化分组.

(2) 行为特征可视化

行为特征可视化的目的是直观地向使用者展示初始组的行为特征,为行为模式探索提供初始的依据.本文提供了多种可视化设计辅助使用者选择初始用户,可视化设计包含行为相似性、统计指标、行为分布这 3 部分.

首先,使用者通过用户行为相似性的可视化设计,即投影分布,观察用户间的相对关系,结合投影点的大小(点的大小映射某个统计属性)选择一个初始用户;第二,当使用者选择某个初始用户后,系统可以在底部统计属性视图中展示其多个量化指标,如 Betweenness,Closeness 等;第三,用户关系视图中展示初始组和初始用户的行为分布.使用者可根据行为特征动态地更换初始用户.在以上多种方式中,行为分布对于初始点选择非常重要.一个好的初始个体,应在时间尺度上具有较为集中的行为分布,通过观察行为特征视图可以了解其在不同时间尺度上的分布情况,有助于选出具有潜在行为模式的群体.依据这些行为特点,使用者可快速了解用户之间的相似程度,用户个体在初始组中的地位和使用设施的时间分布特征,初始组和初始用户在不同时间尺度的行为分布情况等信息.使用者将根据这些信息,在下一阶段选择合适的用户作为群体的初始用户.



(a) 控制面板 (b) 行为特征视图 (c) 子空间选择视图 (d) 统计属性视图 (e) 用户关系视图 (f) 组信息视图

Fig.3 Visual interface design

图 3 可视界面设计

(3) 用户群体发现

寻找具有相似行为模式的群体本质上是用户聚类的过程.由于时间片集合分布稀疏以及用户行为的偶然性,导致很多时间片对于群体的发现是没有作用的,因此,本文选择一种子空间探索的方法,挑选出时间片子集来取代整个时间片集合进行探索.子空间就是时间片子集,它相对于原数据来说,维度降低了很多,稀疏性也有了很大的改善.该方法解决了上文中提出的签到日志数据的高维稀疏性问题.在子空间中,群体使用设施的行为较为一致.不仅如此,本文设计了一个迭代探索过程,每一次迭代都会依据当前群体中用户行为数据的分布,动态改变用于探索的子空间.同时,迭代过程还把在子空间上与群体行为相似度较大的其他用户加入群体,从而保证新生成的群体使用设施的时间也能够集中在子空间上.每个时间片可看做一个离散随机变量,本文使用信息熵度量群体在不同时间片上使用设施的一致性,熵越大,表示群体在该时间片上的签到行为越一致,可以认为在该时间片上更有可能存在特定的签到行为模式.熵的计算如下:

$$e_{t_i} = -\sum_{j=1}^{n} P(u_j)\log_2 P(u_j) \tag{2}$$

其中,$P(u_j)$表示用户 u_j 在时间片 t_i 上使用某设施的概率,n 表示当前群体用户的个数. e_{t_i} 的值越大,表示群体在 t_i 时间片共同签到行为越一致.本方法会为熵设置阈值 et,只有熵大于 et 的时间片才会进入下一次迭代,以确保群体行为在时间片上具有较高的一致性.迭代过程见算法 1,迭代探索的具体流程如图 1 的“用户群体发现”所示.

算法 1. 迭代过程算法.

U 为初始组集合,U'为新生成的群体集合,a 为初始用户,T 为总的时间片集合,T'为根据熵选择的时间片子空间集合,

s_{ij} 为用户 i 和用户 j 之间的行为相似度,e_i 为在时间片 i 上的熵,st 为行为相似度的阈值,et 为熵的阈值.

Input: U, a, T, et, st;

Output: U'.

$U \leftarrow U-\{a\}$

$U' \leftarrow U'+\{a\}$

$q \leftarrow 1$

$T'\leftarrow T$
while $q>0$ **do**
if *length*[U']>1
$T'\leftarrow null$
  for $k\leftarrow 0$ to *length*[T]
    计算群体 U'在 $T[k]$上的熵 $e_{T[k]}$,如公式(2)
    if $e_{T[k]}>et$ **then** $T'\leftarrow T'+T[k]$
  end
$q\leftarrow 0$
for $k\leftarrow 0$ to *length*[U']
  for $x\leftarrow 0$ to *length*[U]
    计算用户 $U'[k]$与用户 $U[x]$在 T'上的行为相似度 $s_{U'[k]U[x]}$,如公式(1)
    if $s_{U'[k]U[x]}>st$ **then**
      $U\leftarrow U-\{U[x]\}$
      $U'\leftarrow U'+\{U[x]\}$
      $q\leftarrow q+1$
    end
  end
end

在迭代开始之前,使用者选择阈值 *st* 和 *et*,此时初始状态仅一个用户,无群体模式,为了不失一般性,此时不计算熵,而是在全部时间片上寻找与其具有相似行为的用户加入 U',进行群体的初始化,在后续迭代过程中计算熵,并通过熵选择时间片.迭代开始后,本方法首先使用当前 U'中的用户计算所有时间片 T 的熵,选取大于 *et* 的时间片,得到子空间 $T'\subseteq T$,然后计算 U'和 U 中两两用户在 T'上的相似度,选择 U 中相似度大于 *st* 的用户加入到 U'中.如果没有新用户加入到 U',则停止迭代,得到群体 U';否则,按以上步骤执行下一次迭代.迭代过程也可由使用者控制结束.

(4) 行为模式理解

完成探索之后,本文帮助使用者理解群体的行为模式.群体行为模式中经常存在多个行为模式交叉的问题,为了解决该问题,本文将群体使用设施的时间按照不同的时间尺度进行划分,如“小时”、“周”、“日”等,之后,对不同的时间尺度采用同一个分析框架,分别对不同时间区间上的用户进行统计,并分析这些区间上用户的关联程度,帮助使用者分析群体在不同时间尺度上的行为模式.

在理解群体行为模式时,为了便于描述不同用户在不同时间区间上的签到分布,本文统计用户在不同时间区间上的签到比例(用户在某时间区间上使用设施的时长占该用户使用设施总时长的百分比),不同用户会有不同的行为偏好.为了描述所找到的群体的共同的行为偏好,本文使用弦图描述群体在哪些时间区间同时签到以及在这些区间上签到的相同的用户个数.例如,群体中只包含两个用户,假设用户在周一~周三的签到比例为50%,49%,1%,此时在弦图中,周一~周三这 3 个时间区间上,两两都有连线且连线的粗度和颜色都是一样的.明显地,该用户绝大部分时间在周一和周二上网,在周三上网具有很强偶然性,因此周三对理解行为模式的理解不仅没有帮助,反而会干扰使用者的理解.为了减少噪音和突出重要的组群内的模式,本文设置了“25%”,“50%”,“75%”这 3 个阈值,以选出群体内前 x%高的分布的时间区间进行绘图.本文使用两两时间区间上的相同用户个数来表示用户在时间区间上的关联程度,通过相同用户的绝对个数和相对个数来表示不同时间区间上用户的绝对和相对关系.绝对个数为两两时间区间上相同用户的个数,相对个数为相同用户的个数与两时间区间上用户并集元素个数的比值.最后,本文通过弦图将统计结果和关联程度可视化,如图 3(b)所示.

4 可视设计

为了让使用者实时全面地了解并灵活地控制行为模式探索过程,我们开发了一种可视分析工具.本文将从分析流程出发,分别介绍 6 个与探索流程相关的视图.

(1) 统计属性视图

统计属性视图用来帮助使用者了解初始组中的个体在统计属性上的特征,如图 3(d)所示.这些统计特征包括用户个体在初始组内的作用和地位以及使用某设施时间的分布特征.使用者可以通过该视图了解每个用户在统计属性上的特征,并将其作为选择迭代探索的初始用户的依据之一(T_1).

本文使用了 9 个统计属性来描述个体的特征,如图 3(d)所示.这些属性包括:

1) Core 指点度中心性(degree centrality),它描述了个体位于组中“核心”位置的程度;
2) Betweenness 是中介中心性(betweenness centrality),是指个体在组中起到的“桥梁”或“中介”作用的程度,描述了该个体与其他个体交往的能力;
3) Closeness 表示接近中心性(closeness centrality),反映了组中个体与其他个体之间的接近程度;
4) Normality 描述个体使用设施的时间符合正态分布的程度;
5) Uniformity 反映了个体使用设施的时间分布的稳定程度;
6) Outliers 用来衡量时间分布中离群值的个数;
7) Unique 是个体使用设施的次数在时间分布上唯一值的个数,表示数据的唯一性;
8) Age 为初始组用户年龄分布,共有“<20”,“20~30”,“30~40”,“>40”这 4 个年龄段,在图 3(d)中,表示这 4 个年龄段的颜色依次变深;
9) Sex 为初始组用户的性别分布,在图 3(d)中,表示 Male 的颜色比表示 Female 颜色浅.

本文通过一个热力图表格来表示各个用户属性值特征,表格的每一列代表一个用户,从上到下依次是各个属性的值对应的矩形,矩形的颜色越深,表示对应属性值越大.最左侧标有属性名的按钮控制用户的顺序,点击其中一个按钮,可视化工具会按照对应属性值的大小对用户排序.在迭代过程中,本文使用对应迭代次数颜色的矩形框来表示加入群体 P'的用户.

(2) 用户关系视图

用户关系视图是本文的主视图,用来帮助使用者了解初始组用户的行为相似性(T_1)以及迭代探索的步骤(T_2),如图 3(e)所示.在每次迭代中,群体的变化、某个用户是在第几次迭代被加入群体的以及在迭代过程中群体的某个用户与其他用户的关系等信息都可从该视图中得到.

本文根据预处理阶段得到的用户签到时间片集合对初始组数据降维,将结果投影到二维的用户关系视图中.降维算法[33]有很多种,比如线性方法 PCA,LDA、非线性方法 MDS,T-SNE 等.其中,T-SNE[34]又称为 *t* 分布随机领域嵌入算法,它是用于探索高维数据的非线性维数降低算法.它将多维数据映射到适合人类观察的两个或多个维度,主要是保证高维空间中相似的数据点在低维空间中的距离尽量较近.MDS[35]同样用于高维非线性降维,但它更适合用于没有特征矩阵只有相似矩阵的情况.由于签到时间片集合是特征矩阵同时又是高维数据,同时,本文希望降维之后在高维中相似的点在低维空间也能保持相对关系,综合以上考虑,本文选择 T-SNE 算法.用户关系视图中,每个点代表初始组中的一个用户,点之间的相对位置表示用户行为相似性.其中,碰撞算法[36]用来减少点的重叠.视图中点的大小由控制面板 Attribute 的值来确定,若复选框中值为 Core,那么用户的 Core 值越大,对应到视图中的点越大.

在迭代开始之前,使用者在控制面板视图 3(a)中选择相似度的阈值 *st* 和熵的阈值 *et*,然后根据用户行为相似性,图 3(b)中用户行为分布以及图 3(d)中统计属性上的特征,进行初始用户的选择.迭代过程中,如果某个用户已被加入到群体 U',那么该用户对应点的颜色变浅,该用户周围也会生出花瓣,如图 3(e)所示.图中花瓣个数表示该用户与群体中其他用户相似度大于阈值 *st* 的用户个数,花瓣的颜色用来表示迭代的次数,颜色越深,迭代次数越大.不同于其他分组算法,本文的方法将使用者考虑其中,通过交互控制迭代进度,进入下一次迭代或返回上一次迭代,或终止迭代.使用者还可以在群体中加入或剔除某个用户.

(3) 子空间选择视图

子空间选择视图是对子空间中时间片在不同时间尺度上的统计,用于了解迭代过程中子空间的变化(T_2),如图 3(c)所示.该视图的前 4 行是对子空间 T'在“月”、“周”、“日”、“小时”的统计,颜色深浅代表时间片的个数.该视图的最后一行是对时间片分布的展示,该行被分为 m 个小矩形 $R=(r_1,r_2,r_3,...,r_m)$,对应在数据预处理时 m 个连续的时间片 $T=(t_1,t_2,t_3,...,t_m)$,如果某个时间片的熵大于阈值,即 $e_i>et$,那么 r_i 被染上色;否则,r_i 为无色.

(4) 组信息视图

每一次迭代,群体中的用户个数都会发生改变,相应地,用户在基本属性上的分布也会发生变化.组信息视图就是用来帮助使用者了解在迭代过程中不断变化的群体在籍贯、性别和年龄段这 3 个基本属性上的分布(T_2),如图 3(f)所示.使用者可以通过该视图的设计找到群体在基本属性上分布的特征,从而更好地理解群体的行为模式(T_3).在迭代过程中,本方法会根据群体的变化调整柱状图的分布.柱状图中矩形的长度 $lr=\frac{count_a}{Count}\times l$,$count_a$ 是当前群体符合某个属性的用户个数,$Count$ 是当前群体用户的个数,l 是矩形的最大长度.

(5) 行为特征视图

行为特征视图用于对初始组和个体行为分布的描述、群体行为模式的理解和探索结果的验证(T_3),如图 3(b),该视图对不同时间尺度上行为分布和关联进行统计分析.在迭代探索开始之前,本文需选择初始用户,该视图对初始组和初始用户在不同时间尺度上的行为分布进行统计分析,结合图 3(d)和图 3(e)中对统计属性和行为相似性的可视化,帮助使用者选择初始用户.在迭代过程中,该视图会随着群体的变化而变化.使用者可结合用户关系视图,调整迭代过程中群体中的用户.使用者分析群体行为的分布和关联,得到群体的行为模式.使用者在该视图中对初始组、群体、剩下组(初始组用户减去群体用户得到的组)的行为模式进行对比,从而验证本文方法的正确性.同时,使用者还可以通过该视图和子空间选择视图中时间的对应关系,验证动态子空间策略的正确性.

在行为特征视图中,使用者在图 3(b_1)Evaluation 复选框中的“日”、“周”、“小时”这 3 个时间尺度上选择以后,视图会展示对应时间尺度的关系图.图 3(b_2)的第 2 个复选框 Percent 是对重要用户的百分比进行筛选,视图对筛选结果进行统计.图 3(b_3)中,Threshold 用来控制连线的多少,弦图中的连线会随着滑动条值的增大去掉颜色比较浅的线(也就是相对用户个数比较少的连线).行为特征视图主要由弦图构成,图 3(b)共有 3 个弦图,从上到下依次初始组关系图、群体关系图、剩下组关系图.群体关系图和剩下组关系图都会随着迭代过程不断变化.弦图的弧长代表在对应时间上用户的个数.连接弧的弦具有颜色和粗细两个特征,它们分别代表两个弧中相同用户个数的相对值和绝对值.颜色越深,表示两个弧相同用户的相对值越大;线越粗,表示绝对值越大.

(6) 控制面板

控制面板视图包含使用者可控的所有变量,用于变量选取.使用者对该视图的操作贯穿了本文的大部分工作,包含分组算法选取、初始组集合表示、初始组选取、用户关系视图中点大小的表示、阈值选取和新群体的表示.使用者可在图 3(a_1)中选择分组算法(kmeans,spectral clustering,decision tree 等)生成初始分组,并在 Group Number 中选择生成初始组的个数.若数据量较少,初始数据也可不进行分组.图 3(a_3)中,Attribute 复选框包含“Core”“Betweenness”“Closeness”“Normality”“Uniformity”“Outliers”“Unique”这 7 个统计属性,使用者可按需选择一个属性,用户关系视图中点的大小将映射该属性值的大小.图 3(a_4)和 3(a_5)中的“Similarity”和“Entropy”两个滑动条控制迭代过程中的两个阈值,分别为时间行为相似度的阈值 st 和熵的阈值 et.只有大于 st 的用户和大于 et 的时间片才会进入下一次迭代.使用者若希望得到关系紧密的群体,可把阈值调大;反之,可调小.在该视图最下方的柱状图是组的列表,每个小矩形代表一个组,矩形的长代表组中用户的数量.图 3(a_6)记录了全体用户的整体分组情况,其状态会随着探索结束后产生的新群体发生变化.图 3(a_6)展示了初始状态(全体个体被分为 3 个组),迭代结束之后,产生新的状态(包含 178 个用户的组分为两个分别包含 168 个用户和 10 个用户的新组).

5 案例分析

本节利用真实的网吧上网数据,分别从群体发现和群体行为理解两个方法验证方法的有效性.

5.1 群体发现

本文首先对上网数据进行初始化.在数据离散化时,本案例将时间跨度设为 30 分钟,因为根据统计,大部分用户的连续上网时间都超过了 30 分钟.由于上网记录的数据量较大,本案例根据数据量将数据分成 3 个组,生成初始组的用户个数分别为 85,37,178.本案例通过 3 个组中用户个数的比较,得出用户个数为 178 的组数据量最大,分组结果可能最为粗糙,因此,本案例选择初始用户个数为 178 的初始组进行迭代探索.

首先,本文根据用户的行为相似性、统计指标以及行为分布为初始组选择初始个体(T_1).因为在初始组中重要的用户是该组的核心,与很多用户都有关联,同时,迭代方法是通过用户之间的相似度将用户加入到群体的,所以本案例使用 Core 值来映射用户关系视图中点的大小.如图 4(b)所示,被圆形框标记出来的点较大,表示该点对应的用户在初始组中比较重要,并且该点位于用户关系视图的中心,周围环绕着很多的用户,表示与其行为相似的用户有很多.如图 5 所示,被矩形框出的用户对应图 4 中被圆形框标记的点,该点在初始组中“Core”“Betweenness”“Closeness”的值较大,表示在初始组中的“重要性”“桥梁”作用、与其他点的接近程度方面的值较大.并且该点使用设施在时间上的分布较为集中,离群值较少.因此,本案例选择该点作为初始个体.

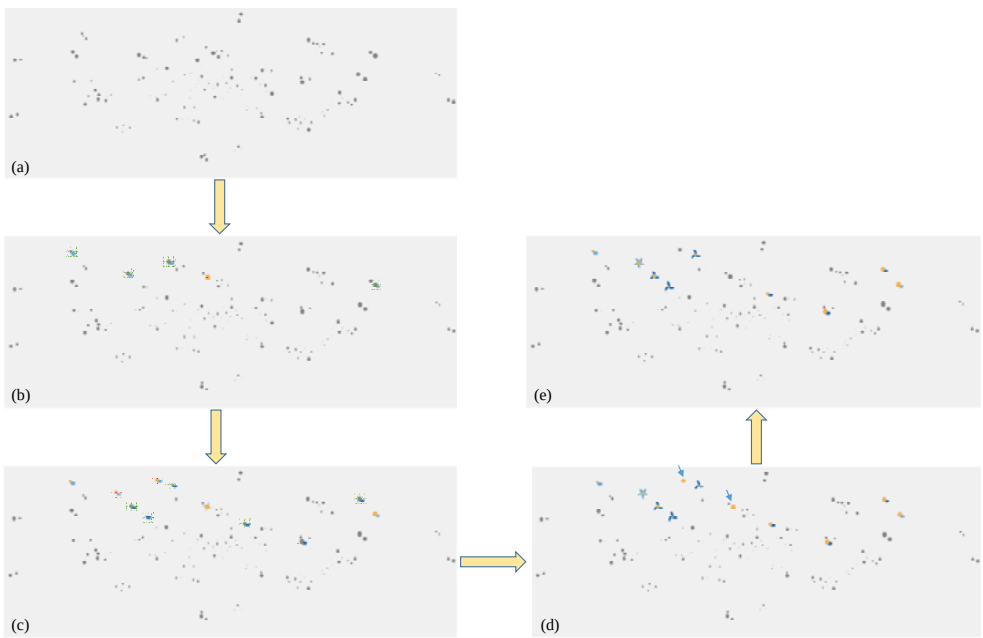


Fig.4 Iteration process

图 4 迭代过程

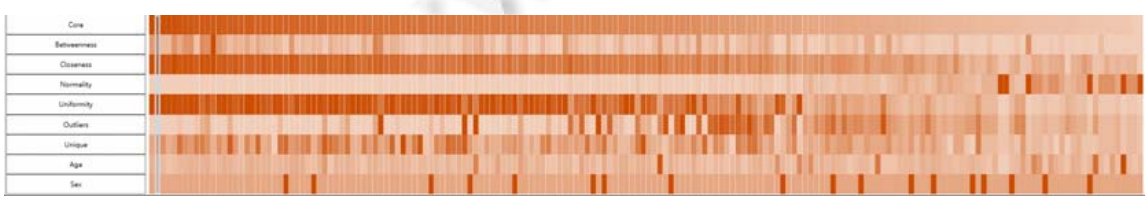


Fig.5 Statistical attributes of initial user

图 5 初始用户的统计属性

然后,本文进行迭代探索(T_2).如图 4 所示,已加入群体的点用圆形框标记,群体中点相似度大于阈值的点用

方形框标记,它们会在下一次迭代中加入群体.如图 4(b)所示,与初始点相似度大于阈值的点有 4 个,在图 4(c)中,这 4 个点被加入群体,此时,初始点对应的点周围有 4 个花瓣,表示该点与刚进入群体的点相似度均较大;刚进入群体的 4 个点都只有一个花瓣,表示这些点只与初始点相似度较大,4 个点之间相似度不大.通过图 4(b)~图 4(d)这 3 次迭代后,已无点被加入群体.由于子空间被不断改变,群体中的点可能在新的子空间中,与其他点的相似度均小于阈值,即无花瓣的颜色较浅的点,如图 4(d)中被箭头标记的点,本文将这些点从群体中去除.如图 4(d)所示,最终本文得到了一个用户数为 10 的群体.

得到具有相似行为模式的群体之后,本案例通过行为特征视图对群体行为模式进行理解(T_3).用户上网的偶然性给群体行为模式的理解带来了困难,因此,本案例对在不同时间尺度上的用户进行筛选,通过比较群体在“25%”“50%”和“75%”这 3 个阈值上弦图的效果之后,我们发现阈值为“25%”时,弦图中弧的分布更为清晰集中,并且细小的连线和颜色较浅连线也减少了很多.这表示在阈值为“25%”时,用户上网模式更为明显,并且不同时间上的关联也较为紧密,因此,本案例使用阈值为“25%”时的行为特征视图对群体行为模式进行理解.

图 6 是对初始组、群体和剩下组在“小时”“天”“周”上行为分布的展示.图 6(a)中,在“小时”上,群体中大部分用户在 16~21 时上网,且连线呈完全图,因此群体明显集中在 16~21 时上网.如图 6(b)所示,在“日”上,与其他两组相比,群体在时间分布的比重上有了很大变化,大部分用户分布在 1~4 日、12 日、17 日,且群体在这些时间的比重明显高于其他时间.初始组和剩下组的时间分布差别不大,时间之间的关联比较混乱,没有明显的规律.在“周”上,群体在周三、周四、周末上网的比重较大.周末有很多用户一起上网,周三、周四也有较多用户一起上网.群体中,上网的人的籍贯主要分布在河北、浙江,性别均为男性,并且年龄全部在 20 岁~30 岁之间,如图 3(f).综上所述:群体成员主要在月初(1~4 日)和月中(12 日、17 日),周三、周四、周末,16~21 时上网.

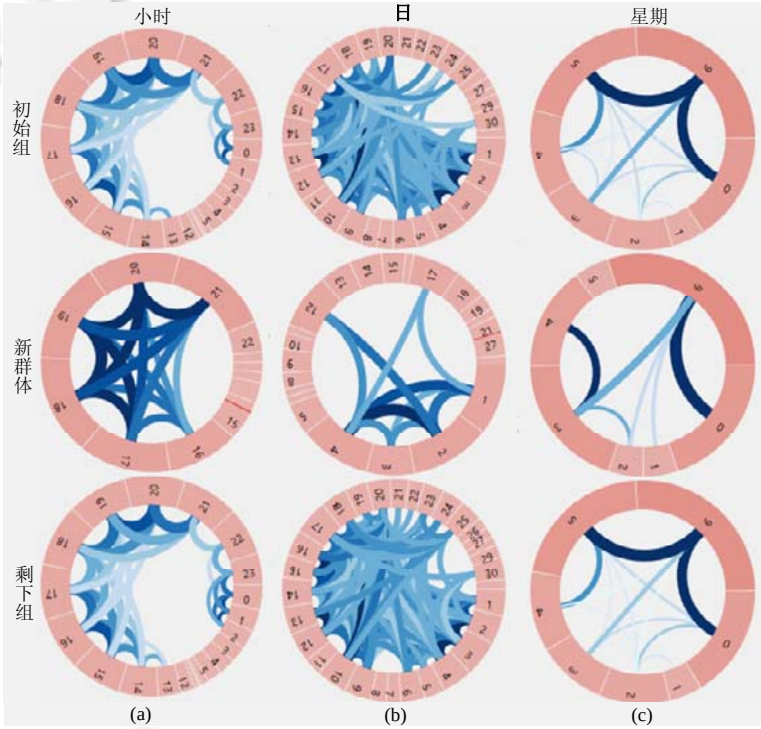


Fig.6 Behavior distribution which threshold is 25%

图 6 阈值为 25%的行为分布

群体的行为模式可得出如下结论:用户主要在周三、周四、周末和傍晚、晚上上网,并且没有熬夜.同时,我们根据基本属性分布可知,群体均为男性且年龄在 20 岁~30 岁之间,因此该群体可能为课余时间较多大学生或

上班时间较为松散的上班族.对比初始组、群体和剩下组中用户在时间上的分布,我们可以明显看到群体中的用户上网时间更集中,并且关联也更紧密清晰,这也验证了本文方法的正确性.

5.2 行为特征理解

本案例通过行为特征视图中不同时间尺度上人数的统计和关联,以及子空间选择视图中时间片在不同时间尺度的分布,对群体行为特征进行进一步的理解(T_3).本案例对网吧初始分组的另一个组进行迭代探索,该组有 85 个用户,探索得到的群体中有 12 个用户.该组群体发现流程与第 5.1 节大致相同,因此本案例不再详细描述.图 7(a)是群体分别在时间尺度为“月”“小时”“日”“周”上的行为特征图.图 7(b)是群体的子空间在“月”“周”“日”“小时”上的统计分布.

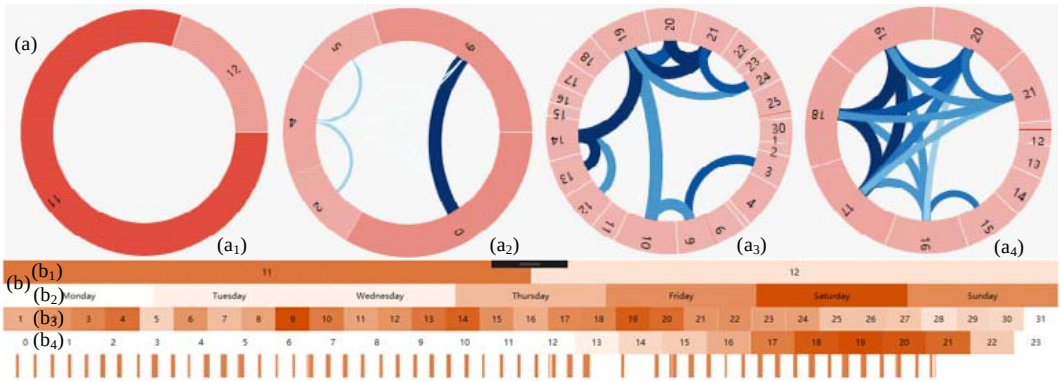


Fig.7 Understanding the features of group behavior

图 7 群体行为特征的理解

图 7(a_1)为群体在“月”上的行为特征视图.在图 7(a_1)中,11 月上网的用户明显多于 12 月.对应图 7(b_1)中,时间片的分布也是如此,两图时间分布相对应.图 7(a_2)中,群体中大部分用户在周末上网,且上网人数相差不大,说明群体中大部分用户在周末一起上网.对应图 7(b_2)中时间片的分布,即周末分布较多,群体在周末一起上网的概率较大.图 7(a_3)为群体在“日”上的行为特征视图,群体中大部分用户 3 日、4 日、9 日、10 日、14 日、18 日~20 日上网,在 4 日、10 日、19 日分布更多,连线更粗且构成完全图,说明群体中有更多用户在 4 日、10 日、19 日一起上网,与图 7(b_3)中的时间分布大致相对应.但时间片在 9 日分布最多,而图 7(a_3)中,9 日的用户分布却不是最多的.如图 7(a_4),在“时”上,群体上网的人数大多分布在 17 时~21 时,并且两两之间都有连线,构成一个完全图,表明群体该时间上网的用户有一部分是相同的,即群体中有一部分用户经常在 17 时~21 时一起上网.18~20 时之间的连线明显比其他连线粗,表明在 18 时~20 时,群体中有更多用户一起上网.在图 7(b_4)中,17 时~21 时的时间片分布较多,说明群体在该时间段一起上网的概率较大,与图 7(a_4)中部分用户一起上网的时间相对应.18 时~20 时的时间片分布更多,也与图 7(a_4)对应.

根据上述分析,群体在不同时间尺度上网规律如下:在“小时”上,群体经常在 17 时~21 时上网,并且上网时间更集中在 18 时~20 时;在“日”上,群体在月初(3 日、4 日、9 日)和月中(14 日、18 日~20 日)上网,且上网时间更加集中在 4 日、10 日、19 日;在“周”上,群体集中周末上网.根据群体上网模式,我们发现:群体通常在周末和晚上上网,并且可能在 17 日~21 时连续上网,说明群体可能在工作日有工作要做,因此我们推测该群体的身份为上班族.

图 7(a)和图 7(b)统计中,两图在时间上基本能够相互对应;同时,由于图 7(b)只是对子空间的统计,而图 7(a)是对全部时间片集合的统计,并且群体只是在子空间的时间片上一起上网的概率较大,一起上网不是必然事件,因此,两视图不能完全对应.总体来说,两图时间上基本相互对应,间接验证了本文动态子空间策略的正确性.

6 专家意见

为了对本文方法的可用性进行评估,我们进行了一个实验.我们邀请了 15 位参与者(5 位女性、10 位男性,年龄在 24 岁~49 岁),为了避免模糊指代,本文根据研究领域对参与者编号.参与者包含 2 位来自数据可视化领域的教授(编号 V_1,V_2),5 位来自数据可视化方向的研究生(编号 V_3~V_7),1 位来自人工智能领域的副教授(编号 A_1),3 位来自人机交互领域的专家(编号 H_1~H_3),3 位来自大数据领域的研究员(编号 D_1~D_3),1 位来自虚拟现实研究领域的副教授(编号 R_1).他们之前均未使用过本文方法.我们首先向参与者介绍本文提出的问题和解决方法,然后参与者使用可视化工具寻找上网数据的群体行为模式.最后,我们对参与者进行访谈.

大多数参与者认为本文可视界面美观,操作流程简单流畅,视图含义易于理解,有较强的可用性.他们指出:多视图协同展示迭代过程,可帮助他们多方位实时了解数据信息.9 个参与者(V_1,V_2,V_4,V_6,H_2,H_3,D_1,D_3,R_1)指出:用户关系视图可帮助他们利用位置判断用户行为相似性,并在本文方法的理解上起到了关键作用.7 个专家(V_1,V_3,V_7,H_1,D_2,D_3,R_1)认为:本文的行为特征视图,简单易懂,不仅可帮助他们了解在迭代过程中群体模式的变化,而且 3 个弦图的对比,可明显地看出群体与其他两组的区别,从而验证本文方法的正确性.同时,他们还指出:行为特征视图使用弦图,直观展示了不同时间上的分布和关联,能容易地找到具体细致的行为模式.V_1 认为:用户关系视图中花瓣的设计新颖美观,点会随着迭代过程改变颜色,添加花瓣易引起注意,使复杂的迭代过程变得易于理解.V_2 指出:若数据量很大,聚类算法分组后每组用户数仍很多,由于可视界面可容纳的用户数有限,会出现点重叠等问题.经测试,本方法可容纳数千用户,满足大部分应用场景的需要.如果数据集包含了更多的样本,可通过提升初始聚类的个数,以减少单个初始簇中用户个数.

大多数参与者认为:信息熵用来度量活动的稳定性,在很多领域有应用,如检测网络异常、图像处理等,本文将熵用于检测群体在某时间上使用设施的一致性是可取的.他们还指出:动态子空间策略相当于在中间过程中改变参数,是对分组算法的创新.D_1 认为:动态子空间策略虽新颖,可以改变过程中的参数,但对于该策略的验证不够直接,应设计进一步的验证.D_2 指出:本文所提方法需要构建初始聚类,以缩小探索空间和提高后期迭代分析的效率,并提供了多个候选聚类算法,但不同的聚类算法可能产生不同的聚类结果.本文提供了多个候选聚类算法,并采用欧式距离作为用户相似度指标.虽然不同的聚类算法和距离指标可能产生不同的结果,但由于初始聚类只是对用户进行粗略的分组,且聚类算法设置的簇个数较小,具有相似行为特征的用户被分到不同簇的可能性较小,因此,使用不同的聚类算法对后续具有相同行为模式群体的探索影响不大.此外,这一过程是可选的,当数据量不大或用户行为不存在明显的差异无法得到清晰的簇时,可不进行初始化分组.

H_1 认为:本文交互操作方便有效,他们可灵活探索群体,可依需选择阈值,从而控制群体用户的个数和相似程度,通过交互控制迭代探索过程;同时,可根据自己的判断和需要从群体中增删用户.这些交互设计新颖特别,将人的智慧融入其中.4 个参与者(V_3,V_4,H_1,H_2)认为:他们虽可通过鼠标交互控制迭代过程,但鼠标点击敏感,一次无意识的点击就会改变迭代进程,如果本文使用其他的交互方式可能会更好.因为本文的迭代方法是可逆的,因此该问题可通过另一交互操作返回上一迭代进程来解决.6 个参与者(V_5~V_7,H_2,H_3,R_1)认为:他们虽可通过交互控制探索进程,但交互操作太多,不易记忆,且未在探索过程中用到全部交互操作.H_3 认为:本文案例中,时间跨度是两个月,但使用者可能只对某时间段比较感兴趣,因此,若本文可动态选择时间段,这将会有更好的体验.

综上所述,大多数参与者对本文方法表示了欣赏,一些参与者对本文方法提出了中肯的建议.我们会根据这些建议,在未来的工作中找到合理的方案来调整本文的设计.

7 讨 论

本节对方法中潜在的问题进行分析,并提供可能的解决方法.

- 数据噪声.若某用户长期占用设施,则其日志于行为模式发现是无用的,迭代探索时,很多用户会因该用户加入群体,使其他用户与该用户关联很强,其他用户之间的关联很弱.但本文会对用户间关系可视化,如图 3(e),若某个点有很多花瓣,而群体内其他点仅一个花瓣,表明其他用户只与该用户有关,可通过交

互去掉该用户;

- 可视重叠.本文通过降维,将数据映射到用户关系视图中,但映射会造成一些相似点的重叠.为了减少重叠,本文使用碰撞算法调整点的相对位置,但位置变化会对用户间关系的判断造成一定的影响,且用户数越多影响越大.本文考虑用气泡代替某些区域,在需要时再将该区域放大,当区域变大时,用户间的重叠就会相对减少;
- 可视化空间有限.由于可视化空间有限,可视化工具不能无限制地容纳数据,数据量越大,视图中点重叠问题越严重,算法调整后,点的位置变化越大.本文可增加聚类算法设置的簇个数,从而减少初始组的数据量;
- 阈值选择的主观性.迭代开始之前,使用者要选择熵和相似度的阈值,由使用者主观决定,因此有两个极限情况:当阈值都选择为 0 时,初始组的用户都会进入群体,造成迭代探索失效;当阈值都为 1 时,群体中只包含最初选择的一个用户.由于迭代方法是可逆的,因此在遇到这两种情况时,使用者可交互地回到最初状态,调整阈值;
- 可扩展性.本文方法仅根据数据的时间属性探索行为模式,并未结合空间等其他信息.若方法结合其他信息,可能会得到更加准确的群体;同时,也使行为模式更易于理解.因为本文数据均来自于一个网吧,因此本文仅使用了时间属性.作者将来会分析签到日志的时空模式,将时间先后顺序和空间拓扑关系纳入分析范畴.

8 总结与展望

本文设计了一个行为模式探索流程和一个可视分析工具,该流程使用动态迭代方法逐步将用户加入群体,同时引入熵的概念,挑选时间子空间,逐步提升迭代效果.可视分析工具将迭代过程可视化,帮助使用者实时了解数据的变化.使用者根据这些变化将自身的判断融合进迭代过程,通过交互对迭代过程进行调整,并对探索结果进行理解和验证.最后,本文通过两个案例分析以及专家意见验证本文方法的可用性.在未来的工作中,我们将尝试将不同的行为记录结合在一起,通过不同方面的属性对用户进行分析,得到更为准确的群体.对于阈值的选取,我们将为使用者推荐更为合理的阈值作为参考.现在的工具通过弦图向使用者描述用户在时间上的分布,未来我们也将改进可视化工具,使行为模式更加易于理解.我们还将对动态子空间策略进行进一步的验证.

References:

[1] Sun JG, Liu J, Zhao LY. Clustering algorithms research. Ruan Jian Xue Bao/Journal of Software, 2008,19(1):48−61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

[2] Peng XF, Pan YM, Luo JB. Predicting high taxi demand regions using social media check-ins. In: Proc. of the IEEE Int'l Conf. on Big Data. Boston: IEEE, 2017. 2066−2075. [doi: 10.1109/BigData.2017.8258153]

[3] Li ZY, Bi J, Zhang J, Li QW. Analysis of airport departure baggage check-in process based on passenger behavior. In: Proc. of the Int'l Symp. on Computational Intelligence and Design. Hangzhou: IEEE, 2017. 204−207. [doi: 10.1109/ISCID.2017.149]

[4] Leemans M, Aalst WMPVD, Brand MGJVD. Recursion aware modeling and discovery for hierarchical software event log analysis. In: Proc. of the 25th Int'l Conf. on Software Analysis, Evolution and Reengineering. Campobasso: IEEE, 2018. 185−196. [doi: 10.1109/SANER.2018.8330208]

[5] Liu Y, Ester M, Qian YQ, Hu B, Cheung DW. Microscopic and macroscopic spatio-temporal topic models for check-in data. IEEE Trans. on Knowledge and Data Engineering, 2017,29(9):1957−1970. [doi: 10.1109/TKDE.2017.2703825]

[6] Chen S, Li J, Andrienko G, Andrienko N. Visual exploration of spatial and temporal variations of tweet topic popularity. In: Proc. of the EuroVis Workshop on Visual Analytics. Bron: The Eurographics Association, 2018. 7−11. [doi: 10.2312/eurova.20181105]

[7] Chen YY, Zhang JY, Guo MY, Cao JN. Understanding customer behaviour in urban shopping mall from WiFi logs. In: Proc. of the IEEE Int'l Conf. on Pervasive Computing and Communications Workshops. Kona: IEEE, 2017. 50−53. [doi: 10.1109/PERCOMW.2017.7917519]

[8] Doi C, Katagiri M, Ishii A, Konishi T, Araki T, Ohta K, Ikeda D, Inamura H, Shigeno H. Estimating customer preference through store check-in histories and its use in visitor promotion. In: Proc. of the 10th Int'l Conf. on Mobile Computing and Ubiquitous Network. Toyama: IEEE, 2017. 1–6. [doi: 10.23919/ICMU.2017.8330107]

[9] Yang K, Wan WG, Xia TY, He X. Urban tourism research based on the social media check-in data. In: Proc. of the IEEE Int'l Conf. on Smart and Sustainable City. Shanghai: IEEE, 2017. 1–3. [doi: 10.1049/cp.2017.0124]

[10] Liu DY, Weng D, Li YH, Bao J, Zheng Y, Qu HM. SmartAdP: Visual analytics of large-scale taxi trajectories for selecting billboard locations. IEEE Trans. on Visualization & Computer Graphics, 2016,23(1):1–10. [doi: 10.1109/TVCG.2016.2598432]

[11] Frhan AJ. Visualization and analysis of user behavior patterns for multimedia content view in social networks. In: Proc. of the Int'l Symp. on Electrical and Electronics Engineering. Galati: IEEE, 2017. 1–7. [doi: 10.1109/ISEEE.2017.8170685]

[12] Lei K, Zhang K, Xu K. Understanding Sina Weibo online social network: A community approach. In: Proc. of the IEEE Global Communications Conf. Atlanta: IEEE, 2013. 3114–3119. [doi: 10.1109/GLOCOM.2013.6831550]

[13] Bron C, Kerbosch J. Algorithm 457: Finding all cliques of an undirected graph. Communications of the ACM, 1973,16(9):575–577. [doi: 10.1145/362342.362367]

[14] Liu HT, Zhao CY, Jian J, Chen L, Zhang DS. The overlapping community discovery algorithm base on link strength. In: Proc. of the IEEE Int'l Conf. on Semantics, Knowledge and Grids. Beijing: IEEE, 2017. 1–7. [doi: 10.1109/SKG.2017.00009]

[15] He CB, Fei X, Li HC, Liu H, Tang Y, Chen QM. Community discovery in large-scale complex networks using distributed SimRank nonnegative matrix factorization. In: Proc. of the IEEE Int'l Conf. on Advanced Cloud and Big Data. Shanghai: IEEE, 2017. 226–231. [doi: 10.1109/CBD.2017.46]

[16] Zhou XK, Wu B, Jin Q. Analysis of user network and correlation for community discovery based on topic-aware similarity and behavioral influence. IEEE Trans. on Human-Machine Systems, 2017,48(6):559–571. [doi: 10.1109/THMS.2017.2725341]

[17] Rohit, Anil KS. Accuracy enhancement of collaborative filtering recommender system for blogs using latent semantic indexing. In: Proc. of the IEEE Information and Communication Technology. Gwalior: IEEE, 2017. 1–4. [doi: 10.1109/INFOCOMTECH.2017.8340646]

[18] Maake BM, Sunday O, Tranos Z. Toward altmetric-driven research-paper recommender system framework. In: Proc. of the IEEE Int'l Conf. on Signal-Image Technology & Internet-Based Systems. Jaipur: IEEE, 2017. 63–68. [doi: 10.1109/SITIS.2017.21]

[19] Yi NN, Li CF, Feng X, Shi MY. Design and implementation of movie recommender system based on graph database. In: Proc. of the Web Information Systems and Applications. Liuzhou: IEEE, 2017. 132–135. [doi: 10.1109/WISA.2017.34]

[20] Hariadi AI, Nurjanah D. Hybrid attribute and personality based recommender system for book recommendation. In: Proc. of the IEEE Int'l Conf. on Data and Software Engineering. Palembang: IEEE, 2017. 1–5. [doi: 10.1109/ICODSE.2017.8285874]

[21] Saas A, Guitart A, África P. Discovering playing patterns: Time series clustering of free-to-play game data. In: Proc. of the IEEE Int'l Conf. on Computational Intelligence and Games (CIG). Santorini: IEEE, 2016. 1–8. [doi: 10.1109/CIG.2016.7860442]

[22] Krueger R, Heimerl F, Han Q, Kurzhals K, Koch S. Visual analysis of visitor behavior for indoor event management. In: Proc. of the Hawaii Int'l Conf. on System Sciences. Kauai: IEEE, 2015. 1148–1157. [doi: 10.1109/HICSS.2015.139]

[23] Li DC, Wang YJ, Wu S, Qi JH, Wang TT. A visual analysis approach to explore criminal patterns based on multidimensional data. In: Proc. of the IEEE Int'l Conf. on Geoscience and Remote Sensing Symposium. Fort Worth: IEEE, 2017. 5563–5566. [doi: 10.1109/IGARSS.2017.8128264]

[24] Zhang X, Wang QY. PeopleVis: A visual analysis system for mining travel behavior. In: Proc. of the IEEE Int'l Conf. on Computer Supported Cooperative Work in Design. Wellington: IEEE, 2017. 463–468. [doi: 10.1109/CSCWD.2017.8066738]

[25] Li J, Chen S, Chen W, Andrienko G, Andrienko N. Semantics-Space-Time Cube. A conceptual framework for systematic analysis of texts in space and time. Journal of Latex Class Files, 2018,14(8). [doi: 10.1109/TVCG.2018.2882449]

[26] Zhao Y, She YM, Chen WJ, Xia JZ, Chen W, Liu JR, Zhou FF. EOD edge sampling for visualizing dynamic network via massive sequence view. IEEE Access, 2018,6(1):53006–53018. [doi: 10.1109/ACCESS.2018.2870684]

[27] Zhou DB, Li H, Liu S, Song B, Hu XH. A map-based visual analysis method for patterns discovery of mobile learning in education with big data. In: Proc. of the IEEE Int'l Conf. on Big Data. Boston: IEEE, 2018. 3482–3491. [doi: 10.1109/BigData.2017.8258337]

[28] Chen S, Li J, Andrienko G, Andrienko N, Wang Y, Nguyen, Phong H, Cagatay T. Supporting story synthesis: Bridging the gap between visual analytics and storytelling. IEEE Trans. on Visualization and Computer Graphics (Early Access), 2018. [doi: 10.1109/TVCG.2018.2889054]

[29] Wei J, Shen Z, Sundaresan N, Ma KL. Visual cluster exploration of web clickstream data. In: Proc. of the IEEE Conf. on Visual Analytics Science and Technology. Seattle: IEEE, 2012. 3–12. [doi: 10.1109/VAST.2012.6400494]

[30] Zhao Y, Luo F, Chen M, Wang Y, Xia J, Zhou F, Wang Y, Chen Y, Chen W. Evaluating multi-dimensional visualizations for understanding fuzzy clusters. IEEE Trans. on Visualization & Computer Graphics, 2019,25(1):1–10. [doi: 10.1109/TVCG.2018.2865020]

[31] Li J, Chen S, Zhang K, Andrienko G, Andrienko N. COPE: Interactive exploration of co-occurrence patterns in spatial time series. IEEE Trans. on Visualization and Computer Graphics (Early Access), 2018. [doi: 10.1109/TVCG.2018.2851227]

[32] Li J, Zhang K, Meng ZP. Vismate: Interactive visual analysis of station-based observation data on climate changes. In: Proc. of the Visual Analytics Science & Technology. Paris: IEEE, 2014. 133–142. [doi: 10.1109/VAST.2014.7042489]

[33] Camastra F. Data dimensionality estimation methods: A survey. Pattern Recognition, 2003,36(12):2945–2954. [doi: 10.1016/S0031-3203(03)00176-6]

[34] Gisbrecht A, Mokbel B, Hammer B. Linear basis-function *t*-SNE for fast nonlinear dimensionality reduction. In: Proc. of the Int'l Joint Conf. on Neural Networks. Brisbane: IEEE, 2012. 1–8. [doi: 10.1109/IJCNN.2012.6252809]

[35] Shang Y, Ruml W. Improved MDS-based localization. In: Proc. of the Joint Conf. of the IEEE Computer & Communications Societies. Hong Kong: IEEE, 2004. [doi: 10.1109/INFCOM.2004.1354683]

[36] Fan ZW, Wan HG, Gao SM. A fast collision detection algorithm in image space. Journal of Computer-Aided Design & Computer Graphics, 2002,14(9):805–809 (in Chinese with English abstract). [doi: 10.3321/j.issn:1003-9775.2002.09.001]

附中文参考文献:

[1] 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48−61. http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

[36] 范昭炜,万华根,高曙明.基于图像的快速碰撞检测算法.计算机辅助设计与图形学学报,2002,14(9):805−809. [doi: 10.3321/j.issn:1003-9775.2002.09.001]

李丛敏(1991－),女,河北邯郸人,硕士,主要研究领域为数据可视化,可视分析.

张康(1959－),男,博士,教授,博士生导师,主要研究领域为计算美学,图文法,可视化语言,生成艺术.

李杰(1984－),男,博士,讲师,CCF 专业会员,主要研究领域为时空可视分析,大数据探索工具,社交媒体,公共安全,环境科学.

陶文源(1971－),男,博士,教授,博士生导师,主要研究领域为物联网理论与技术,数字媒体技术.