计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190649
57(11): 2442-2455, 2020

基于校园行为信息网络的生活习惯相似学生搜索

王新澳[1] 段 磊[1] 崔丁山[1] 卢 莉[1] 顿毅杰[2] 秦蕊琦[1]
[1](四川大学计算机学院 成都 610065)
[2](西北民族大学数学与计算机科学学院 兰州 730030)
(wangxinao@stu.scu.edu.cn)

Search of Students with Similar Lifestyle Based on Campus Behavior Information Network

Wang Xin'ao[1], Duan Lei[1], Cui Dingshan[1], Lu Li[1], Dun Yijie[2], and Qin Ruiqi[1]
[1](*School of Computer Science, Sichuan University, Chengdu* 610065)
[2](*School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou* 730030)

Abstract It is important to keep track of both the psychological and academic status of students in campus. Generally, student data covers a wide range of kinds such as students' interests, hobbies, and lifestyles, and these data can be collected via smart devices such as student e-cards by many campuses. With the rapid development of new generation of information technology, in recent years, researchers have explored novel ways to improve the quality of talent cultivation by utilizing the student data, such as applying big data analysis on the data to discover subtle but meaningful information as the guidance for better student management. Among such research, search of students with similar lifestyles can exert positive effect on the improvement of student management, as potential and insightful information can be found and may further provide some warnings for students at an early stage if anything unusual is found. Existing algorithms for searching students with similar lifestyle have two deficiencies. Firstly, they cannot explain the similarities between students because related semantic information is lost in the searching process. Secondly, they fail to integrate multiple data sources, while the student behavioral data is growing dynamically and only using one dataset may lead to biased results. To break these limitations, we first propose the concept of campus behavior information network to represent student behaviors in campus. Next, based on the constructed campus behavior information network, an algorithm named SCALE is proposed for similar campus lifestyle mining. SCALE calculates the student similarity by specific meta-paths with constraints. SCALE is strong and unique, not only in keeping the similarity semantics of the original data, but also in extensively integrating multiple data sources in a scalable way while retaining the original results of calculation. Due to the large scale of datasets, parallel strategy is further designed and

收稿日期:2019-09-02;**修回日期**:2020-01-17
基金项目:国家自然科学基金项目(61972268,61572332);国家重点研发计划项目(2018YFB0704301);四川省重点研发项目(2019YFG0491);四川省高等教育人才培养质量和教学改革项目(JG2018-92)
This work was supported by the National Natural Science Foundation of China (61972268, 61572332), the National Key Research and Development Program of China (2018YFB0704301), the Key Rsearch and Dvelopment Program of Sichuan Province (2019YFG0491), and Sichuan Higher Education Talent Training Quality and Teaching Reform Project (JG2018-92).
通信作者:段磊(leiduan@scu.edu.cn)

applied to SCALE for the sake of efficiency. Through extensive experiments on real campus behavior datasets, the effectiveness and execution efficiency of the SCALE are verified.

Key words campus behavior information network; heterogeneous information network; student behavior analysis; meta path; similar student search

摘 要 利用大数据分析、深度学习等新一代信息技术,通过掌握学生的兴趣、爱好、生活习惯等,提高人才培养质量已成为当前重要的科学研究问题.寻找具有相似生活习惯的学生对于心理状况及学业状况预警都有着积极的作用.已有的相似生活习惯学生搜索算法无法解释学生之间相似的原因,并且无法拓展性地融合更多数据源.为此提出了基于校园行为信息网络的生活习惯相似学生搜索算法 SCALE (similar campus lifestyle miner).SCALE 算法通过带约束的元路径计算相似度.SCALE 不仅能保留原始数据中的相似语义,同时可以在此基础上拓展性地融合更多数据源.进一步对算法各部分解耦,为 SCALE 算法设计了并行化策略以提高执行效率.通过真实校园环境数据集上的实验,验证了 SCALE 算法的有效性和执行效率.

关键词 校园行为信息网络;异构信息网络;学生行为分析;元路径;相似学生搜索

中图法分类号 TP311

随着 2018 年国家标准《智慧校园总体框架》发布,致力于构建校园工作、学习和生活一体化的智慧校园正在全国多个高校逐步成型,从课堂到生活的教育理念已经被广为接受.传统基于预制定教学计划的培养模式已不能满足当前创新性人才的个性化培养需求.以大数据分析、人工智能等信息技术为支撑的智慧教育模式已成为教育信息化的趋势[1],通过掌握学生的兴趣、爱好、生活习惯等,提高人才培养质量成为当前教育领域的重要研究问题.

生活习惯是学生心理状况、财务状况和兴趣爱好的综合体现,对学生的个人发展和学业表现有着重要的影响.分析学生的行为,掌握学生的生活习惯,对关爱学生心理健康、明晰学生财务状况、促进学生学业进步有非常重要的作用.例如:中国矿业大学根据学生校园生活状况,建立家庭经济困难学生数据库,提供精准资助依据①.西安电子科技大学利用大数据分析学生食堂用餐期间的消费记录,"隐性地"资助贫困学生②.

计算学生生活习惯的相似性,搜索相似的学生,可以支持包括下面 2 种场景的众多应用:

1) 场景 1.现有的大学寝室分配方法较单一,没有充分考虑学生的兴趣、性格、生活习惯等方面,容易造成矛盾.通过搜索生活习惯相似学生,调整寝室分配,对促进和谐校园、改善寝室氛围有着积极的作用.

2) 场景 2.学生进行社团选择、项目组队时信息来源较少.搜索与学生生活习惯一致的社员或队友,可以为学生的选择提供参考,同时有利于突破学生自身交际圈促成跨专业或跨学院的交流.

本文基于校园行为信息搜索具有相似生活习惯的学生.从技术上讲,使用校园行为数据分析学生生活习惯具有 2 方面挑战:

1) 学生在校行为数据是多源、异构且持续增长的,包含例如选课、成绩、消费、门禁等不同来源和不同结构,并会随时间逐渐增多数据.算法设计过程中需要充分考虑原始数据的这些特点.

2) 不同数据源之间的语义复杂,包括自习(图书馆门禁数据)、饮食(食堂消费数据)等.在计算相似性时需要保证语义清晰准确,即能够解释相似的原因.

目前教育数据挖掘领域绝大多数研究的关注点在于学生的学习过程和学习表现以及一些特殊任务,例如评估抑郁[2]、拖延症[3]、学业预警[4]或辅助奖助学金发放[5-6]等.文献[7]通过基于 LINE 的网络嵌入方法获得学生的低维向量表示,从而计算学生之间的相似性,但这种方法会损失原始数据中包含的语义信息,并且无法拓展性地融合更多的数据源.

使用异构信息网络可以很好地将学生和行为信息保存在一起.借鉴异构信息网络的思想和技术[8],我们构建校园行为信息网络(campus behavior infor-

① http://www.moe.gov.cn/jyb_xwfb/s6192/s133/s183/201612/t20161212_291588.html

② http://www.cpwnews.com/content-22-32315-1.html

mation network)来表达学生在校行为信息.并且在校园行为信息网络中,我们用具有明确语义信息的元路径度量学生之间的相似性,从而得到所有学生之间的相似关系.目前基于异构信息网络的相似性度量方法已较为成熟,但因为校园活动数据与常用于构建异构信息网络的数据不同,具有重复率高的特点(第 2 节做详细分析),目前的相似性度量方法并不完全适用于校园行为信息网络.

同时因为校园行为数据多源的特点,在单一数据源的行为信息网络中提取的相似语义信息往往是片面的.例如,仅使用图书馆的进出记录无法确定一个学生是否喜欢上自习,因为教学楼同样具有自习的功能.因此有必要集成多个网络中的相似信息来更全面地体现学生的在校行为.相应地,还需要设计将多个学生相似信息融合起来的方法,用于从整体上评判学生之间的相似性.

对此,本文提出 SCALE(similar campus lifestyle miner)算法用于解决在校园行为信息网络中搜索生活习惯相似学生的问题.主要工作有 4 个方面:

1) 单层学生相似子网络的构建.由单一数据源得到校园行为信息网络,提出一种带约束的元路径相似度计算方法,使用给定的元路径计算学生之间的相似度,构建单层学生相似子网络.

2) 学生相似网络的构建.增量式地将单层学生相似子网络构建为一个多层结构的学生相似网络,并通过带偏随机游走的方式生成每个学生的上下文语义.

3) 基于网络嵌入的相似学生搜索.使用 Skip-Gram 模型将所有学生的上下文语义嵌入到一个低维向量空间中,将每位同学的相似信息向量化.通过计算向量之间的相似度搜索相似学生.

4) 通过真实校园环境数据集上的实验,验证了 SCALE 算法的有效性和执行效率.

1 问题定义

我们首先引入一些用于表示学生行为的概念.

我们用一个三元组〈时间(t),地点(l),事件类型(c)〉表示事件实例,描述学生在时间 t、地点 l 参与了一个类型为 c 的事件.校园中常见的事件类型包括:上课、自习、就餐等.例如〈2019-08-19T11:05:00,一食堂,就餐〉表示了一名学生于 2019 年 8 月 19 日 11:05:00 在一食堂就餐的事件实例.

考虑到校园行为一般以教学周为周期迭代进行,我们用时间约束(τ)描述一对时间条件$\{W(t)=T_{\mathrm{dow}}, T(t)\in T_z\}$,其中 T_{dow} 表示 1 周中的某一天,T_z 表示 1 天中的某个时间区间.满足此约束的时间 t 记作 $t \sqsubset \tau$.例如$\{W(t)=\mathrm{Monday}, T(t)\in[11:00, 13:00)\}$为一个具体的时间约束.

满足同一个时间约束且在相同地点发生的同类型事件实例的集合体现了相似的行为,由一个行为实例表示,记作〈时间约束(τ),地点(l),事件类型(c)〉.对于$\forall t \sqsubset \tau, l=l, c=c$,都有$\langle t,l,c\rangle \in \langle \tau,l,c\rangle$.

例 1. 有属于学生 1 和学生 2 的 3 个事件实例.

1) 学生 1:〈2019-08-26T11:05:00(t_1),一食堂,就餐〉;

2) 学生 1:〈2019-08-26T12:45:00(t_2),一食堂,就餐〉;

3) 学生 2:〈2019-08-25T11:55:00(t_3),一食堂,就餐〉.

对于时间约束 τ:$\{W(t)=\mathrm{Monday}, T(t)\in[11:00,13:00)\}$,$t_1$,$t_2$ 满足时间约束 τ,而 t_3 不满足 τ.因此,学生 1 的 2 个事件实例均属于同一个行为实例〈$\{W(t)=\mathrm{Monday}, T(t)\in[11:00,13:00)\}$,一食堂,就餐〉.且学生 1 参与了此行为实例 2 次,学生 2 没有参与此行为实例.

校园行为信息网络是 1 个保存了学生行为信息的有向图 $G=(V,E,W)$,具有 1 个对象类型映射函数 $\phi: V\to\mathcal{A}$,1 个链接类型映射函数 $\psi: E\to\mathcal{R}$ 和 1 个属性类型映射函数 $\theta: W\to\mathcal{W}$,并且 $|\mathcal{A}|>1$ 或 $|\mathcal{R}|>1$,$|\mathcal{W}|>0$.其中,V 代表对象集合,$\mathcal{A}$ 代表对象类型,$E\subseteq V\times V$ 代表链接集合,$\mathcal{R}$ 代表链接类型,W 代表链接权重的集合,$\mathcal{W}$ 代表属性值的类型.

校园行为信息网络包含了 5 种典型的对象类型:学生(s)、时间约束(τ)、地点(l)、事件类型(c)、行为实例(b).时间约束、地点及事件类型为行为实例的属性.网络还包括 4 种类型的链接:学生与行为实例之间具有参与几次或者被参与几次的关系,行为实例和时间约束之间存在“发生”或者“发生在”的关系,行为实例和地点之间存在处于或发生的关系,行为实例与事件类型之间存在属于或包含的关系.容易看出,校园行为信息网络是一个带权重的异构信息网络[9],包含了 4 种权重类型.学生与行为实例之间链接的权重为学生参与此行为实例的次数,时间约束、地点和事件类型为行为实例的属性,它们与行为实例之间链接的权重均为 1,且任一行为实例必须与其对应的时间约束、处于的地点及属于的事件类型对象相连.图 1 为校园行为信息网络的一个

示例,时间约束、地点、事件类型与行为实例之间链接的权重被省略.

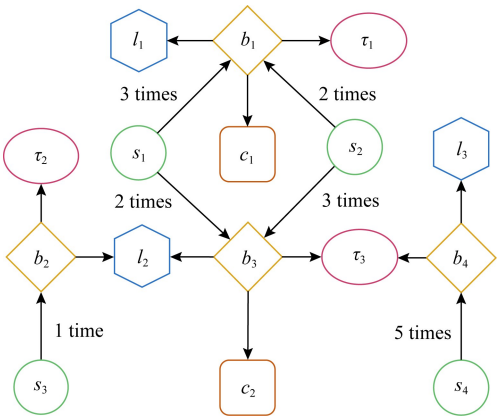

Fig.1 An example of campus behavior information network

图 1 校园行为信息网络示例

在校园行为信息网络中,2 个对象可以通过多条不同的路径相连,连接 2 个对象的某一条路径蕴含了这 2 个对象之间的某种语义关系,且不同路径表达着不同的语义关系,称这些路径为元路径,记作 P.若元路径 P 上的链接带有权重,则 P 为带权重元路径[9].

若校园信息网络中存在 1 条与元路径 P 的对象类型和链接类型全部对应的路径 p,则称 p 为元路径 P 的实例,记作 $p\in P$.

考虑元路径 P:"学生—行为实例—地点—行为实例—学生",在要求路径中对象不重复的情况下,图 1 中存在着 2 条元路径 P 的实例.p_1:"s_1—b_3—l_2—b_2—s_3";p_2:"s_2—b_3—l_2—b_2—s_3".

在校园行为信息网络中使用元路径查找相似语义时,存在不同类型行为的路径并不能表达相似,因此要求元路径中出现的行为实例为相同事件类型.具有较强相似语义信息的元路径有 3 条:

1)"学生—行为实例—时间约束—行为实例—学生".2 个学生在相同的时间约束下具有相同类型的行为,例如图 1 中包含的实例"s_1—b_3—τ_3—b_4—s_4",语义为 s_1 和 s_4 在相同的时间约束(τ_3)下有相同类型的行为(b_3,b_4 的事件类型同为 c_2).

2)"学生—行为实例—地点—行为实例—学生".2 个学生在相同的地点具有同样类型的行为,例如图 1 中包含的实例"s_1—b_3—l_2—b_2—s_3",语义为 s_1 和 s_3 在相同的地点(l_2)有同样类型的行为(b_2,b_3 的事件类型同为 c_2).

3)"学生—行为实例—学生".2 个学生在相同的时间约束下和相同的地点有相同的行为,例如图 1 中包含的实例"s_1—b_3—s_2",等价于同时存在前 2 种元路径的情况,即同时存在实例"s_1—b_3—τ_3—b_3—s_2"和"s_1—b_3—l_2—b_3—s_2",语义为 s_1 和 s_2 在相同的时间约束(τ_3)下和地点(l_2)中有相同的行为(b_3 的事件类型为 c_2).

可以发现,上面 3 种元路径与其反向的元路径是相同的,我们称这种元路径为对称元路径[8].对于一个给定的对称元路径 P,文献[8]给出了 2 个相同类型对象 o_s 和 o_t 之间基于实例数的元路径相似性度量方式 PathSim.

$$Sim(o_s,o_t,P)=\frac{2\times|\{p_{o_s\Rightarrow o_t}|p_{o_s\Rightarrow o_t}\in P\}|}{|\{p_{o_s\Rightarrow o_s}|p_{o_s\Rightarrow o_s}\in P\}|+|\{p_{o_t\Rightarrow o_t}|p_{o_t\Rightarrow o_t}\in P\}|},\quad(1)$$

其中,$p_{o_s\Rightarrow o_t}$ 表示 o_s 和 o_t 之间的路径实例,$p_{o_s\Rightarrow o_s}$ 表示 o_s 和 o_s 之间的路径实例,$p_{o_t\Rightarrow o_t}$ 表示 o_t 和 o_t 之间的路径实例.

例 2. 对于图 1 中的校园行为信息网络 G 和元路径 P:"学生—行为实例—学生".学生 1(s_1)与学生 2(s_2)之间的 Pathsim 相似度计算如下:

1) 学生 1 与学生 2 之间元路径 P 的实例有 2 条,分别为"s_1—b_1—s_2"和"s_1—b_3—s_2",因此 $|\{p_{s_1\Rightarrow s_2}|p_{s_1\Rightarrow s_2}\in P\}|=2$.

2) 学生 1 与学生 1 之间元路径 P 的实例有 2 条,分别为"s_1—b_1—s_1"和"s_1—b_3—s_1",因此 $|\{p_{s_1\Rightarrow s_1}|p_{s_1\Rightarrow s_1}\in P\}|=2$.

3) 学生 2 与学生 2 之间元路径 P 的实例有 2 条,分别为"s_2—b_1—s_2"和"s_2—b_3—s_2",因此 $|\{p_{s_2\Rightarrow s_2}|p_{s_2\Rightarrow s_2}\in P\}|=2$.

4) 因此,$Sim(s_1,s_2,P)=\frac{2\times2}{2+2}=1$.

通过基于元路径的相似度计算方式,我们可以基于给定元路径从校园行为信息网络中计算得到所有学生之间的相似度.以学生作为节点、相似度作为权重,构建单层学生相似子网络.单层学生相似子网络是一个无向带权重图 $B=(S,\Theta)$,其中每个节点 $s\in S$ 代表 1 个学生,每条边 $e\in\Theta$连接 2 个相似的学生,e 上带有的属性 w 代表 2 个学生的相似度.

但是获得多个子网络之后,单层学生相似子网络的权重并不能表达学生之间的相似度.因此为了度量学生在多个子网络中表现出的相似性,我们构建多层结构的学生相似网络并使用网络嵌入的方法得到学生的向量表示,从而通过计算向量之间的距离得到学生之间的相似性.

本文的研究问题(Top-k 生活习惯相似学生搜索):给定查询学生 s,基于校园信息网络集合 $\mathcal{G}$ 和元路径集合 $\mathcal{P}$,找到与 s 相似度最大的 k 个学生.

2 相关工作

本文基于异构信息网络,以信息网络的形式重构校园行为数据,构建了校园行为信息网络,使用结合元路径方法的网络嵌入方法来研究校园行为信息网络中的相似搜索.因此,本节将从基于异构信息网络的相似性度量和教育数据挖掘 2 个方面介绍本文的相关工作.

2.1 基于异构信息网络的相似性度量

异构信息网络被定义为由多种类型的实体和关系构成的网络.区别于传统的网络,异构信息网络包含了不同的类别信息,它们能用来表达路径中丰富的语义信息.因此在大部分现实场景下,异构信息网络更适合用于对现实世界进行抽象表示.近些年,为了研究复杂网络中节点之间丰富的联系,基于异构信息网络的数据挖掘任务成为了研究热点,其中包括聚类[10]、分类[11]、链接预测[12]和相似搜索[13]等.比如,Sun 等人[10]将元路径与融入了用户偏好的聚类相结合,从而对网络中对象聚类;Ji 等人[11]基于在 1 个类中排位更高对象应该有更重要作用的思想,提出了基于排序的分类方法 RankClass;Kuo 等人[12]通过综合的统计方法,将异构信息网络中不同类别的信息建模到一个多层的图中,并推理出隐藏的链接.侯泳旭等人[13]构建了包含疾病、基因和病症节点的疾病信息网络,并设计了基于元路径的相似基因搜索算法 gSim_Miner.在这些任务中,异构信息网络的相似性度量是一个基本并且重要的功能.在下文中,我们将总结异构信息网络的相似性度量的相关工作.

不少研究者已经意识到基于异构信息网络的相似性度量的重要性.Ni 等人[14]在利用科学文献中丰富的元数据构建有向图的基础上,设计了一个有路径约束的随机游走算法(path-constrained random walks, PCRW)来测量任意类型节点对之间的相似性.Sun 等人[8]考虑到不同类型对象组成的元路径能表达语义,提出了 PathSim 算法,该算法通过对称的元路径计算 2 个相同类型对象之间的相似性.Shi 等人[15]结合 PCRW 和 PathSim 算法,设计了 HeteSim 算法,通过用户给定的任意的元路径计算相同或不同类型的对象相关性.注意:校园行为信息网络与其他常见的异构信息网络存在不同,学生常在几个固定的场所活动,很少前往没有去过的地点,且对于熟悉的地点,学生通常会频繁前往,即重复率高,所以在校园信息网络中需要以边上权重的方式存储学生与某地之间产生联系的频度,且一般情况下权重会比较高.若使用以上的方法计算元路径相似度,边上的权重信息就会被丢失,例如偶尔去 1 次图书馆和经常出入图书馆会被相似度评价方法视作相同的行为.因此以上方法不适用于本问题.近年来,Shi 等人[9]介绍了 SemRec 算法,并提出用带有权重的元路径来精细地描述路径语义,在计算实例数时要求对称的 2 个关系所具有的权重相等,从而保证被计算的实例能够表达 2 个对象之间相似的语义.但是 SemRec 适用于评分的场景,对于重复率高的校园数据来说,只计算权重相等的实例太过严格,会丢失过多的语义.

网络嵌入是将对象嵌入到低维稠密的向量空间中的技术,能有效捕捉对象的重要信息.因此,许多研究工作将基于元路径的方法融入网络嵌入来得到节点唯一的向量表达.Metapath2vec[16]和 HIN2Vec[17]通过元路径的随机游走得到节点的序列,并结合 Skip-gram 模型从而得到网络节点的嵌入.HEBE[18]提出了异构信息网络中事件的概念,它将参与同一个事件的对象看为 1 个整体,即 1 个事件,并用超边表示对象之间的多种关系,从而得到对象的近似.TransPath[19]借用了知识图谱中的平移机制的思想,将元路径当作源结点到目标节点的平移操作,用于得到元路径和节点的嵌入.但是此类方法的拓展性普遍较差,在融合更多数据源的数据时已有的计算结果将被全部重新计算.

2.2 教育数据挖掘

近年来,由于学生相关数据越来越多,教育数据挖掘(educational data mining, EDM)已成为一个新兴的跨学科研究领域.EDM 指在教育环境中利用数据挖掘的技术解决实际的教育教学问题,从而改善和提高学生学习质量,完善学习过程与教育管理[20].

在教育数据挖掘中,大部分研究关注于学生的学习过程[21-26]和学习表现[27-33].这些方法通过分析线下或线上的学习活动所产生的数据来进行建模,从而研究和预测学生的学习行为和学习成绩.除了学生的学习过程和学习表现,校园生活等也引起了研究者的注意.Resnik 等人[2]分析对大学生的问卷调查,使用文本分析主题建模以预测学生中的抑郁者.

Zhu 等人[3]提出了一个从行为画像到预测抑郁的无监督学习模型(动态 RP),该模型通过分析大学生在图书馆的借阅记录来评估学生的拖延症.Sattar 等人[4]介绍了一个框架,该框架利用了多组不同类型的变量,包括了家庭背景、中学信息、注册登记和学分,以预测学生退学的概率.Ye 等人[5]给出了多模型多标签的方法,来辅助大学提供学生奖学金和补助金的分配.Guan 等人[6]设计了 Dis-HARD 框架,用于预测学生应给的补助等级.Hang 等人[7]将学生的 Check-In 数据(WIFI 访问日志)整合到二部图,并编码学生、兴趣点(point of interest, POI)和活动之间的相关性,用以预测 POI、查询相似学生.

据我们所知,在教育环境下的研究工作只有文献[7]针对有着相似生活行为学生的搜索,与本文最为相似.但文献[7]提出的算法基于 LINE 进行向量嵌入,计算时会丢失语义信息,并且无法拓展性地融合更多数据源.本文将在实验部分与文献[7]提出的算法进行对比.

3 SCALE—生活习惯相似学生搜索

SCALE 是基于校园行为信息网络的生活习惯相似学生搜索算法.学生的校园行为是多种多样的,因此描述学生在校行为的数据也是多源的,对于单个数据源可以构建出一个校园行为信息网络,通过给定的元路径能得到单层学生相似子网络.显然,单层学生相似子网络所包含的信息是片面的,无法从整体上对学生之间的相似性进行表达.因此需要构建多层结构的学生相似网络,并使用网络嵌入的方法将所有学生映射到低维的向量空间中,从而使相似学生搜索问题得到简化.

图 2 展示了 SCALE 算法的主要流程.

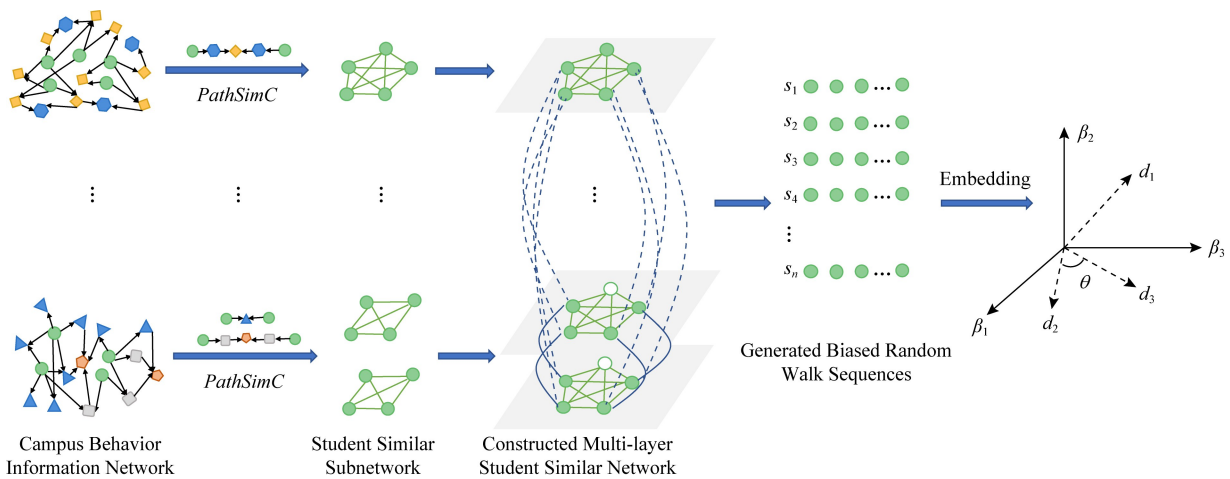

Fig.2 Algorithm flow introduction of SCALE

图 2 SCALE 算法流程介绍

3.1 单层学生相似子网络的构建

根据不同的数据源,构建校园行为信息网络的方式有很多种.对于行为信息来说,我们首先可以把学生的所有事件划分为多个独立的行为实例,用行为实例作为事件实例的载体保存在网络中.同时,为保证能够在网络的元路径中提取到明确的语义,我们按如下方式构建校园行为信息网络:

1) 根据校园生活存在的周期性和具体情况设置时间约束.不失一般性,我们采用与文献[7]相同的方式将所有的时间划分到以 1 周 7 天为周期,每天 4 个时间段(从 0 点开始,每 6 h 为 1 个时间段)所组成的 28 个时间约束中.

2) 将同一个时间约束下,同一个地点发生的相同类型的事件实例保存在同一个行为实例对象中存入校园行为信息网络.并与对应的时间约束、地点和事件类型对象相连,链接的权重为 1.

3) 将每个学生作为 1 个对象存入网络,并与参与的行为实例对象相连,链接的权重为参与的次数.

自然地,所有的行为实例都具有时间约束、地点及事件类型属性.因此上述的校园行为信息网络构建方式对于所有的校园行为都适用.但校园行为信息网络的表达能力是可拓展的.针对一些具有特殊属性的行为实例,也可以将这些属性作为节点加入到校园行为信息网络中,使网络包含更丰富的语义.例如,对于学生的消费行为,可以将“消费金额范围”作为行为实例的属性存储在校园行为信息网络中,

从而使元路径“学生—行为实例—消费金额范围—行为实例—学生”表达2个学生消费金额相近的语义.

根据上述的方式在单数据源下构建校园行为信息网络后,我们可以通过基于元路径的相似性度量方式计算学生之间在此网络中的相似度.本文提出一种基于权重相似度的方式对元路径的实例数进行计算.

给定1个对称带权重元路径 P,2个相同类型对象 o_s 和 o_t 及阈值 $\alpha \geqslant 1$,对于 o_s 和 o_t 之间的路径 p_k 中任意2个对称的路径 $o_i \xrightarrow{\delta_i(R_i)} o_{i+1}, o_{j-1} \xrightarrow{\delta_j(R_i^{-1})} o_j$,若

$$\frac{1}{\alpha} \leqslant WeightSim(o_i, o_j, R_i) \leqslant \alpha, \tag{2}$$

$$WeightSim(o_i, o_j, R_i) = \frac{|\delta_i(R_i)|}{outdeg(o_i, \delta_i(R_i))} \Big/ \frac{|\delta_j(R_i^{-1})|}{indeg(o_j, \delta_j(R_i^{-1}))}, \tag{3}$$

则 p_k 为对象 o_s 和 o_t 之间带权重元路径 P 满足约束 C 的1个实例,记作 $p_k \in P|C$.$|\delta_i(R_i)|$ 代表此路径上权重的值,$outdeg(o_i, \delta_i(R_i))$ 代表 o_i 在关系 $\delta_i(R_i)$ 上出度的和,$indeg(o_j, \delta_j(R_i^{-1}))$ 代表 o_j 在关系 $\delta_j(R_i^{-1})$ 上入度的和.则带约束 C 的元路径相似度的计算公式为

$$PathSimC(o_s, o_t, P) = \frac{2 \times |\{p_{o_s \Rightarrow o_t} | p_{o_s \Rightarrow o_t} \in P|C\}|}{|\{p_{o_s \Rightarrow o_s} | p_{o_s \Rightarrow o_s} \in P|C\}| + |\{p_{o_t \Rightarrow o_t} | p_{o_t \Rightarrow o_t} \in P|C\}|}. \tag{4}$$

使用带约束的元路径相似度计算公式可以得到所有学生相互之间的相似度值,从而构建学生相似子网络.

例3. 对于图1中展示的校园行为信息网络 G,给定元路径 P:“学生—行为实例—地点—行为实例—学生”及权重相似度阈值 $\alpha = 2$.构建基于(G, P)的单层学生相似子网络的步骤为:

1) s_1 与 s_3 之间有1条路径实例 p_1.“s_1—b_3—l_2—b_2—s_3”,因行为实例对象只会和1个地点对象相连且链接的权重均为1,即 $WeightSim(b_2, b_3, \text{locate at})$ 恒为1,一定符合权重约束条件.由此可见行为实例对象和时间约束、地点、事件类型对象之间的链接不需要考虑权重相似度的约束.且 $WeightSim(s_1, s_3, \text{participate}) = \frac{2}{5} \Big/ 1 = \frac{2}{5} \notin \left[\frac{1}{2}, 2\right]$, $p_1 \notin P$.s_1 与 s_3 之间无元路径实例.

2) s_2 与 s_3 之间有1条路径实例 p_2.“s_2—b_3—l_2—b_2—s_3”,$WeightSim(s_2, s_3, \text{participate}) = \frac{3}{5} \Big/ 1 = \frac{3}{5} \in \left[\frac{1}{2}, 2\right]$, $p_2 \in P$.s_2 与 s_3 之间有1条元路径实例.

3) 同理,s_1 与 s_2 之间有2条元路径实例,s_4 与其他学生对象之间无元路径实例.s_1, s_2, s_3 与自身之间分别有2条、2条、1条元路径实例.

4) 使用 w_{ij} 代表 s_i 与 s_j 的相似度,有

$$w_{12} = PathSimC(s_1, s_2, P) = \frac{2 \times 2}{4 + 3} = \frac{4}{7},$$

$$w_{13} = PathSimC(s_1, s_3, P) = \frac{0}{4 + 1} = 0,$$

$$w_{23} = PathSimC(s_2, s_3, P) = \frac{2 \times 1}{3 + 1} = \frac{1}{2}.$$

5) s_4 与其他学生对象之间的相似度均为0.

以每一个学生作为对象,学生之间相似度作为链接的权重,构建基于(G, P)的单层学生相似子网络.

3.2 学生相似网络的构建

单层学生相似子网络只反映了从1个数据源中通过1条元路径语义表达的学生相似性,将得到的多个单层学生相似子网络整合起来,形成1个多层结构的学生相似网络.因为每个学生一定是和自身完全相似的,所以通过权重为1的边将多层网络中相同的学生对象连接起来.从而获得1个多层的网络结构表达学生之间的相似关系.

SCALE在学生相似网络中采取带偏的随机游走算法生成每个学生的上下文语义.因为网络是多层的,因此随机游走的过程中会出现2种情况:1)算法根据随机生成的概率选择留在本层,以更大概率游走到和当前节点更相似的节点,即与当前节点由更大权重的边相连的节点;2)算法选择游走到网络中的其他层,则此步不再做其他操作.通过上述的随机游走算法,可以为每一个学生生成1个由相似节点组成的序列,表达其他节点与当前节点之间的相似关系.

3.3 基于网络嵌入的相似学生搜索

通过带偏的随机游走算法在学生相似网络中获得每个学生与其他学生的相似关系之后,SCALE采用Skip-Gram模型对所有的随机游走序列进行嵌入学习.从而将所有学生映射到1个低维的向量空间中,使得每个学生嵌入的向量保留了学生相似网络中体现的相似性.

得到所有学生的向量表示之后,对于每一个查询学生,利用余弦相似度计算此学生向量与其他所有向量之间的距离,得到距离最小的 k 个向量,所对应的 k 个学生即为SCALE的搜索结果.

需要注意的是,SCALE 在单层学生相似子网络构建时采用基于带约束的元路径相似度计算方法度量节点间相似性,在学生相似网络生成上下文语义和网络嵌入时使用带偏随机游走和 Skip-Gram 模型将学生映射到低维向量.在其他的应用中,可以根据使用场景,更换上述度量方式或表达学习方法.

SCALE 算法的整体流程如算法 1 所示.

算法 1.SCALE 算法.

输入:校园行为信息网络的集合 $\mathcal{G}$、给定的元路径集合 $\mathcal{P}$、查询学生 s、参数 k;

输出:学生 s 的 k 个最相似学生集合 $\mathcal{S}$.

① $\mathcal{N}\leftarrow\varnothing$; /* 初始化学生相似网络集合 */

② FOR $G\in\mathcal{G}$,$P\in\mathcal{P}$ DO

③ 计算 G 中每一对学生的 $PathSimC(s_i,s_j,P)$;

④ $N\leftarrow$由 G 构建的单层学生相似子网络;

⑤ $\mathcal{N}\leftarrow\mathcal{N}\cup\{N\}$;

⑥ END FOR

⑦ $\mathcal{M}\leftarrow$连接 $\mathcal{N}$ 中不同层的相同节点; /* 得到多层结构的学生相似网络 */

⑧ $\mathcal{R}\leftarrow$在 $\mathcal{M}$ 上进行带偏随机游走; /* 得到随机游走序列 */

⑨ $\mathcal{V}\leftarrow$使用 Skip-Gram 模型对 $\mathcal{R}$ 中每一个节点进行向量嵌入; /* 得到所有学生向量表示集合 */

⑩ $\mathcal{S}\leftarrow$计算向量间余弦相似度得到 s 的 Top-k 个相似学生;

⑪ RETURN $\mathcal{S}$.

3.4 并行化

SCALE 算法有 3 处是可解耦的,因此可以针对本算法设计并行化处理方法,从而提高算法效率.

1) 学生相似度计算.在构建学生相似网络的过程中,需要对任意 2 个学生之间计算相似度.而不同对学生之间计算相似度的过程是互不影响的,因此在学生相似度计算时,即单层学生相似子网络的构建过程中可以使用多进程(线程)提升程序运行效率.

2) 学生相似网络构建.构建不同的学生相似子网络的过程是相互独立的,在相似网络构建的过程中网络之间不会互相影响,因此可以使用多进程(线程)完成学生相似网络的构建过程.

3) 构建学生相似网络之后,需要针对每个学生使用带偏随机游走算法生成大量的随机游走序列,此处可用 2 个思路实现并行化:①每个进程(线程)都对所有的学生生成部分随机游走序列,全部运行完成后将结果进行拼接得到 1 个学生所有的随机游走序列.②每个进程(线程)只对部分学生生成所有的随机游走序列,全部运行完成后得到所有学生的随机游走序列.

同时,因为 SCALE 算法构建学生相似网络的过程是解耦的,因此 SCALE 算法是一个可拓展的方法.当添加新的数据源或元路径时,只需将新获得的单层学生相似子网络加入到之前已经构建好的学生相似网络中即可进行后续计算.之前计算得到的学生相似网络无需重新进行计算,由此节约了运算资源.

4 实 验

本文利用真实的数据集验证校园行为信息网络的适用性和相似学生搜索算法 SCALE 的有效性以及执行效率.实验源码存放于 https://github.com/hdwxa/SCALE.git.

4.1 数据集介绍及实验设置

本文使用 2018 年 3 月 1 日—11 月 30 日期间,四川大学 3 个校区内采集到的 6 个不同学院共 2 449 名学生在校行为数据进行本次实验.该数据包含 2 个数据源:1)后勤集团数据(source1).学生在校园内食堂、便利店及澡堂等地点的消费记录,共包含 1 276 806 个事件实例.2)保卫处数据(source2).学生进出教学楼、球场、寝室楼的门禁记录,共包含 752 361 个事件实例.表 1 分别展示了相关的事件实例数为 Top-5 的地点和事件类型,及它们对应的事件实例数.

Table 1 Top-5 Locations and Event Types with Highest Number of Event Instances

表 1 发生事件实例数 Top-5 的地点和事件类型

Top-5 Location	Count	Top-5 Event Type	Count
1F,Huaxi Eastern Canteen	193 368	Repast	949 596
Eastern 4 Dorm	118 750	Dorm Entrance	752 018
1F,Jiang'an Western Second Canteen	102 616	Water	269 334
Jiang'an Boiler Room	101 834	Recharge	51 780
Wangjiang Student Centre	101 780	Traffic	39 507

表 2 列出了通过每个数据源构建的校园信息网络的具体规模.

Table 2 Size of Campus Behavior Information Networks

表 2 校园行为信息网络规模

Item	Number		
	Source1	Source2	Total
Nodes in Network	3 324	3 417	6 741
Edges in Network	191 128	63 861	254 989

为验证 SCALE 算法的有效性和执行效率,本文在真实数据集上运行 SCALE 算法,挖掘 Top-k 生活习惯相似学生.从有效性测试、模型简化测试以及应用实例 3 方面说明 SCALE 算法的有效性.并验证 SCALE 算法采取的并行化策略对执行效率的提升效果.

4.2 SCALE 有效性测试

与本文工作相似的最新工作是由文献[7]提出的 EDHG 算法,对于给定的查询学生 s、向量嵌入维度 d 和负采样个数 m,EDHG 可以找到 Top-k 个相似学生,但无法提供对结果相似的语义解释.

同时,本文还将校园行为信息网络转化为矩阵的形式记录学生在 2 个数据源中参与某个事件类型、时间约束和地点的行为实例的次数,针对每位学生构建 9×28×101 的 3 维张量.其中后勤集团数据包含 6 种事件类型及 44 个地点,保卫处数据包含 3 种事件类型及 57 个地点,时间约束个数均为 28.通过主成分分析得到每位学生在事件类型、时间约束和地点维度上的第 1 主成分作为每位学生的向量表示,以此搜索 Top-k 的相似学生,与 SCALE 算法进行效果对比,从而说明 SCALE 算法获取校园行为信息网络中信息的准确性.3 种算法分别记为 PCA-c,PCA-τ,PCA-l.

文献[7]提出使用共现行为,即 2 位学生在很短的时间内(本次实验设置为 2 min)同时出现在同一个地点,作为学生之间是否在行为上相似的一种评判方式.2 位学生之间共现行为越多,则这 2 位学生生活习惯就更为相似.本文采取与文献[7]相同的方式作为评估模型效果的指标.以共现行为最高的 k 个学生为标准,对 SCALE 算法找到的 Top-k 个相似学生使用平均相关排名(mean reciprocal rank, *MRR*)进行评估.平均相关排名的计算方式为

$$\frac{1}{|U|}\sum_{i=1}^{|U|}\sum_{j\in F_i}\frac{1}{Rank(j)}, \tag{5}$$

其中,U 为全部查询学生的集合,F_i 为使用共现行为找出学生 i 的 $|F_i|=k$ 个相似生活习惯的学生,$Rank(j)$为学生 j 由 SCALE 算法计算出的排名.*MRR* 得分越高,说明 SCALE 算法的效果越好.

实验过程中,SCALE 算法需要设置的参数有:每次查询搜索的相似生活习惯学生个数 k、计算学生相似度时的权重相似度阈值 α、多层学生相似网络中对每个节点产生随机游走序列的个数 n,以及使用 Skip-Gram 模型进行向量嵌入的维度 d.为保证提取的相似语义充分且不重复,实验在元路径"学生—行为实例—学生"上计算相似度.表 3 记录了将四川大学学生在校行为数据分别应用于 PCA-c,PCA-τ,PCA-l,EDHG 算法和 SCALE 算法得到的结果.

Table 3 *MRR* Scores

表 3 *MRR* 评分

k	PCA-c	PCA-τ	PCA-l	EDHG	SCALE
2	0.003	0.015	0.016	**0.017**	**0.017**
4	0.009	0.036	0.040	0.044	**0.056**
6	0.018	0.051	0.062	0.066	**0.095**
8	0.026	0.066	0.082	0.085	**0.136**
10	0.035	0.080	0.101	0.104	**0.172**

Note: The best values are in bold.

在表 3 中可以看出,在 $k=2$ 时,SCALE 算法和 EDHG 算法的效果相近,且都比 PCA-c,PCA-τ,PCA-l 效果好.随着 k 的增大,5 种算法的 *MRR* 得分都呈现增长趋势,并且 SCALE 算法的得分始终高于其他 4 种算法,说明本文提出的 SCALE 算法在寻找相似生活习惯学生的任务上比其他 4 种算法效果更好.在 $k=10$ 时,SCALE 算法相对于 PCA-c,PCA-τ,PCA-l,EDHG 算法的效果提升分别达到了 391%,115%,70.3%,65.4%.同时可以发现,在 k 增大时,SCALE 算法相对于其他 4 种算法效果提升得更为明显,说明 SCALE 算法的效果在 k 取较大的值时更有优势.

图 3(a)~(c)分别展示了在完整数据集下参数 α,n,d 对于 SCALE 算法效果的影响.图 3(a)中可以看出,随着权重相似度阈 α 变大,算法的效果呈现先升后降的趋势,在 $\alpha=1.4$ 时,SCALE 算法取得最好的效果,因此默认情况下设置 $\alpha=1.4$.由图 3(b)可以看出随着每个节点产生随机游走序列个数 n 的增大,SCALE 的效果也逐渐变好,但当 n 由 128 增大至 256 时,模型效果的提升很微弱,因此本次实验默认将 n 设置为 128.由图 3(c)观察可知,当 $d=32$ 时 SCALE 效果最好,因此默认设置 $d=32$.

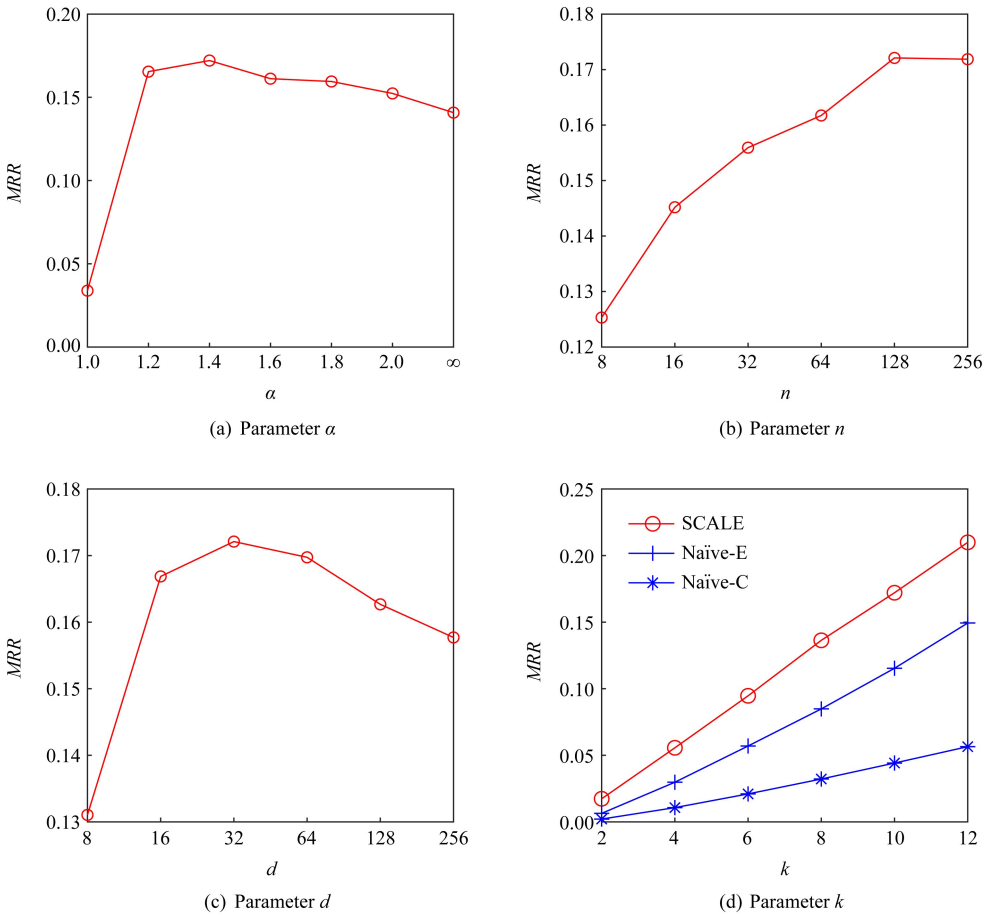

Fig. 3 Influence on SCALE with respect to parameters

图 3 参数对 SCALE 有效性的影响

4.3 模型简化测试

在图 3(a)中,当权重相似度阈值 $\alpha=1$ 时,*PathSimC* 等价于文献[9]提出的算法,当 α 为正无穷时,*PathSimC* 等价于 PathSim 算法,SCALE 算法的效果在 $\alpha=1.4$ 时获得最好效果,说明 *PathSimC* 相对于之前的方法可以更好地保留学生之间相似生活习惯的信息.

我们还将没有构建多层学生相似网络的单数据源 Naïve 算法与 SCALE 算法进行对比,说明在多数据源情况下使用 SCALE 算法的有效性.在本实验中,使用消费数据和门禁数据的 Naïve 算法分别记为 Naïve-C 和 Naïve-E,对比结果记录在图 3(d)中.可以看出,SCALE 算法的效果始终好于 2 种 Naïve 算法,说明使用多层结构的学生相似网络可以更好地保留多数据源中的学生生活习惯信息.

4.4 应用实例

SCALE 算法使用的相似度计算方法是基于元路径的,因此 SCALE 算法相对于 EDHG 算法的另一个优点就是还保留了原始数据中的语义信息.本实验展示 2 种应用场景下 SCALE 算法的 Top-k 搜索结果.

1) 在消费和门禁 2 个数据源中都使用元路径"学生—行为实例—学生"计算相似度.相似度高的学生说明他们更倾向于在同一时间、同一地点产生相同的行为.

2) 仅使用消费数据源,将"消费金额范围"作为行为实例的属性存储在校园行为信息网络中,使用元路径"学生—行为实例—消费金额范围—行为实例—学生"和元路径"学生—行为实例—地点—行为实例—学生"计算相似度,相似度高的学生说明他们消费金额相近且喜欢去相同的地方消费,即消费能力相似.

本文随机抽取了 3 位学生,并展示针对他们搜索得到的 Top-10 相似的学生来说明结果的合理性.为方便对比,展示时使用"专业—班号—学号后 2 位"代替学号.由表 4 的结果可以看出,在第 1 种应用场景下,寻找到的相似学生绝大多数都是相同专业甚至是相同班级的学生,这是因为相同专业和班级

Table 4 Top-10 Similar Students Found by SCALE

表 4 Top-10 相似学生展示

Student	Top-k Similar Lifestyle Students (k=10)	
	Meta-Path SBS	Meta-Path SBMBS & SBLBS
CS-201-10	(CS-201-53,0.990),(CS-201-21,0.990),(CS-201-54,0.988),(CS-201-72,0.988),(CS-201-30,0.988),(CS-201-39,0.986),(CS-201-47,0.986),(CS-201-28,0.986),(CS-201-32,0.986),(CS-201-79,0.986)	(MDAM-104-38, 0.766), (Nur-701-29, 0.745), (CS-101-78, 0.738),(SE-103-58,0.713),(MDAM-110-70,0.711),(Dra-203-66,0.706),(Ani-701-08,0.697),(MLS-404-06,0.694),(VCD-402-39,0.693),(RTP-601-19,0.693)
Acc-102-29	(Acc-103-73,0.994),(Acc-101-68,0.993),(Acc-201-49,0.993),(Acc-103-62,0.993),(Acc-104-34,0.992),(Acc-101-12,0.992),(Acc-103-36,0.991),(Acc-102-22,0.990),(Eng-0101-04,0.989),(Acc-1004-15,0.988)	(Acc-103-08,0.887),(OM-101-20,0.865),(ID-201-08,0.864),(CS-106-05,0.844),(OM-104-96,0.843),(ID-201-24,0.843),(IE-801-38,0.841),(Mus-801-11,0.840),(CS-111-07,0.840),(CS-107-36,0.839)
CM-304-14	(CM-304-21,0.968),(CM-304-04,0.963),(CM-304-80,0.920),(CM-506-32,0.916),(CM-304-22,0.853),(Nur-702-21,0.812),(CM-302-68,0.808),(CM-505-35,0.806),(CM-501-65,0.805),(CM-502-46,0.797)	(SE-110-31, 0.898), (HRM-601-77, 0.868), (MDAM-109-45, 0.868),(CF-601-09,0.862),(CS-105-66,0.850),(MDAM-103-59,0.864),(CS-102-82,0.843),(WRHE-105-01,0.843),(HRM-601-97,0.838),(MDAM-103-35,0.835)

Note: CS(computer science), MDAM(mechanical design manufacture and automation major), Nur (nursing), SE(software engineering), Dra(drawing), Ani (animation), MLS(medical laboratory science), VCD(visual communication design), RTP(radio and television programming), Acc(accounting), Eng(English), OM(oral medicine), ID(industrial design), IE(industrial engineering), Mus(music), CM (clinical medicine), HRM (human resource management), CF (computational finance), WRHE (water resources and hydropower engineering)

学生的上课时间安排及主要活动区域是一致的,因此他们更倾向于在相同时间前往相同的教学楼、食堂、宿舍等区域,说明 SCALE 算法在计算相似性时成功捕获了此类信息.同时我们可以发现一些有趣的现象:第 2 位和第 3 位查询学生在其搜索到的相似学生中都各自出现了 1 个非本专业的学生.我们通过查看以上学生的基本信息,发现第 2 位同学与其相似的非本专业相似学生性别都为女性,我们推测她们可能是好友.第 3 位同学与其相似的非本专业同学为不同性别(与其他相似同学均为同性别),推测他们可能是情侣.

而在第 2 种应用场景下,不再出现大多数相似学生专业、班级甚至性别属性相同的情况.这和常识相符,因为第 2 种场景下元路径所表达的语义为消费能力相似,与专业、班级或性别属性的相关性较小.

可见 SCALE 算法具有很好的灵活性,根据语义设置不同的元路径可以获取学生之间不同的相似性.

4.5 SCALE 执行效率

为了验证 SCALE 算法并行化策略对效率的提升效果,本文使用不采取并行化策略的 SCALE-Ser 算法和使用了并行化策略的 SCALE 算法在不同数据规模下对比执行时间.同时验证 SCALE 算法在数据规模上的拓展性,本实验在合成数据集上完成.

若无特殊说明,实验过程中参数设置与有效性实验中保持一致.并行化使用最大进程数为 10 的进程池实现.在图 4(a)中可以看出,SCALE 算法相对于 SCALE-Ser 算法有显著的效率提升.但只降低到了原时间规模的 40%左右,并没有在最大进程数为 10 的情况下将效率提升到预期的 10 倍.这是因为并行化方法只对 SCALE 算法的学生相似网络构建和随机游走部分进行了并行化,并没有对网络嵌入和 Top-k 搜索步骤采取并行测量,因此并行化并不能完全达到预期的效果.

同时我们可以发现,随着数据集规模的增大,SCALE 算法的耗时呈非线性关系增大趋势,这是因为在构建学生相似网络部分需要计算任意 2 个学生之间的相似度,通过 Skip-Gram 模型进行向量嵌入时也需要与其他所有学生作对比,因此当数据规模增大时需要进行的计算次数以平方规模增长,因此时间的增加呈现非线性趋势.

图 4(a)中还可以看出,SCALE 算法具有较好的拓展性,在学生规模达到 20 000 时仍然可以支持相似学生的搜索.真实环境下,在上万人中搜索相似学生已经可以满足绝大多数需求,因此本算法是具有现实意义的.

图 4 的(b)~(d)分别展示了参数 α, n, d 对 SCALE 算法效率的影响.图 4(b)中可以看出 α 对于 SCALE 算法效率的影响不大,只有在 α 较小时耗

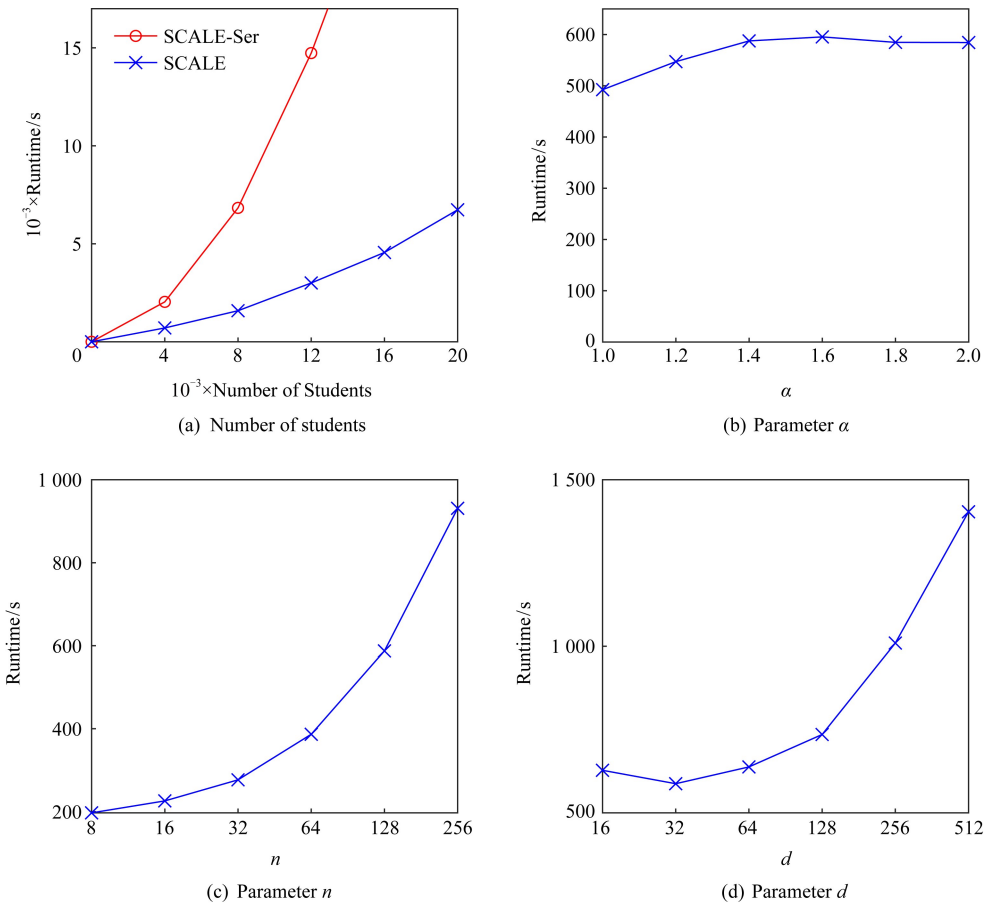

Fig.4 Scalability test and runtime with respect to parameters

图 4 拓展性测试及参数对 SCALE 效率的影响

时略低,这是因为在 α 很接近 1 时,构建学生相似网络过程中只有很少的学生之间有边连接,因此导致耗时较短.在 α 增长到 1.4 后 SCALE 算法的效率保持稳定.参数 n 对 SCALE 算法效率的影响在随机游走和网络嵌入 2 部分,图 4(c)中可以看出,参数 n 以乘方规模增大时,SCALE 算法耗时也呈非线性增长,但是增长速度没有达到乘方规模.图 4(d)中展示了 SCALE 算法随参数 d 的变化,整体上呈现非线性增长的趋势,但是在 d 由 16 增长至 32 时,耗时反而下降了,这可能是因为在 $d=16$ 时,Skip-Gram 无法快速收敛,因而导致效率降低.

5 结 论

搜索相似生活习惯的学生在教育数据挖掘领域是一个值得被关注的问题,但目前已有的研究存在着语义缺失或不适用于校园场景数据等问题,因此本文提出 SCALE 算法用于搜索校园场景下生活习惯相似的学生,在保留学生间相似语义的情况下设计带约束的元路径相似度计算方法解决校园场景数据中存在的密集性高的问题,最终得到所有学生的低维向量表示,从而搜索 Top-k 的相似生活习惯学生.同时,我们将 SCALE 算法的各部分解耦,通过并行化的方法提升效率.最后,我们在校园环境采集到的真实数据集中验证了 SCALE 算法的有效性和执行效率.

因为 SCALE 算法的设计是模块化、易拓展的,因此下一步可以考虑将更多的数据源纳入 SCALE,同时可以尝试在网络嵌入部分使用更为前沿的方法以提升模型的效果.在目前 SCALE 的算法流程中,并未考虑噪声对搜索结果的影响,如何在搜索过程中降低噪声的影响从而获得更准确的结果是未来需要进一步研究的工作.

参 考 文 献

[1] Zheng Qinghua, Dong Bo, Qian Buyue, et al. The state of the art and future tendency of smart education [J]. Journal of Computer Research and Development, 2019, 56(1): 209-224 (in Chinese)

(郑庆华, 董博, 钱步月, 等. 智慧教育研究现状与发展趋势[J]. 计算机研究与发展, 2019, 56(1): 209-224)

[2] Resnik P, Garron A, Resnik R. Using topic modeling to improve prediction of neuroticism and depression [C] //Proc of the 2013 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1348-1353

[3] Zhu Yan, Zhu Hengshu, Liu Qi, et al. Exploring the procrastination of college students: A data-driven behavioral perspective [G] //LNCS 9642: Proc of Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2016: 258-273

[4] Sattar A, Mahtab J F, Ratna B C, et al. Survival analysis based framework for early prediction of student dropouts [C] //Proc of the 25th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2016: 903-912

[5] Ye Hanjia, Zhan Dechuan, Li Xiaolin, et al. College student scholarships and subsidies granting: A multi-modal multi-label approach [C] //Proc of the 16th IEEE Int Symp on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2016: 559-568

[6] Guan Chu, Lu Xinjiang, Li Xiaolin, et al. Discovery of college students in financial hardship [C] //Proc of the 15th IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2015: 141-150

[7] Hang Mengyue, Pytlarz I, Neville J. Exploring student check-in behavior for improved point-of-interest prediction [C] //Proc of the 24th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 321-330

[8] Sun Yizhou, Han Jiawei, Yan Xifeng, et al. PathSim: Meta path-based top-k similarity search in heterogeneous information networks [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003

[9] Shi Chuan, Zhang Zhiqiang, Ji Yugang, et al. SemRec: A personalized semantic recommendation method based on weighted heterogeneous information networks [J]. World Wide Web, 2019, 22(1): 153-184

[10] Sun Yizhou, Norick B, Han Jiawei, et al. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks [C] //Proc of the 18th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1348-1356

[11] Ji Ming, Han Jiawei, Danilevsky M. Ranking-based classification of heterogeneous information networks [C] //Proc of the 17th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1298-1306

[12] Kuo T, Yan Rui, Huang Yuyang, et al. Unsupervised link prediction using aggregative statistics on heterogeneous social networks [C] //Proc of the 19th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 775-783

[13] Hou Yongxu, Duan Lei, Li Ling, et al. Search of genes with similar phenotype based on disease information [J]. Journal of Software, 2018, 29(3): 721-733 (in Chinese)

(侯泳旭, 段磊, 李岭, 等. 基于疾病信息网络的表型相似基因搜索[J]. 软件学报, 2018, 29(3): 721-733)

[14] Ni Lao, William W C. Fast query execution for retrieval models based on path-constrained random walks [C] //Proc of the 16th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 881-888

[15] Shi Chuan, Kong Xiangnan, Yu P S, et al. Relevance search in heterogeneous networks [C] //Proc of the 15th Int Conf on Extending Database Technology. New York: ACM, 2012: 180-191

[16] Dong Yuxiao, Chawla N V, Swami A. Metapath2vec: Scalable representation learning for heterogeneous networks [C] //Proc of the 23rd ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 135-144

[17] Fu Taoyang, Lee W, Lei Zhen. HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning [C] //Proc of the 2017 ACM Conf on Information and Knowledge Management. New York: ACM, 2017: 1797-1806

[18] Gui Huan, Liu Jialu, Tao Fangbo, et al. Embedding learning with events in heterogeneous information networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(11): 2428-2441

[19] Fang Yang, Zhao Xiang, Tan Zhen, et al. TransPath: Representation learning for heterogeneous information networks via translation mechanism [J]. IEEE Access, 2018, 6: 20712-20721

[20] Zhou Qing, Mou Chao, Yang Dan. Research progress on educational data mining: A survey [J]. Journal of Software, 2015, 26(11): 3026-3042 (in Chinese)

(周庆, 牟超, 杨丹. 教育数据挖掘研究进展综述[J]. 软件学报, 2015, 26(11): 3026-3042)

[21] Sun Xia, Zheng Qinghua. Educational resources search based on metadata expanded semantically [J]. Journal of Computer Research and Development, 2004, 45(12): 2170-2174 (in Chinese)

(孙霞, 郑庆华. 教育资源元数据语义扩展查找方法的研究[J]. 计算机研究与发展, 2004, 45(12): 2170-2174)

[22] Jiang Zhuoxuan, Zhang Yan, Li Xiaoming. Learning behavior analysis and prediction based on MOOC data [J]. Journal of Computer Research and Development, 2015, 52(3): 614-628 (in Chinese)

(蒋卓轩, 张岩, 李晓明. 基于MOOC数据的学习行为分析与预测[J]. 计算机研究与发展, 2015, 52(3): 614-628)

[23] Zhu Tianyu, Huang Zhenya, Chen Enhong, et al. Cognitive diagnosis based personalized question recommendation [J]. Chinese Journal of Computers, 2017, 40 (1): 176-191 (in Chinese)

(朱天宇, 黄振亚, 陈恩红, 等. 基于认知诊断的个性化试题推荐方法[J]. 计算机学报, 2017, 40(1): 176-191)

[24] Liu Qi, Huang Zai, Huang Zhenya, et al. Finding similar exercises in online education systems [C] //Proc of the 24th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1821-1830

[25] Shi Yuling, Peng Zhiyong, Wang Hongning. Modeling student learning styles in MOOCs [C] //Proc of the 2017 ACM Int Conf on Information and Knowledge Management. New York: ACM, 2017: 979-988

[26] He Juan. Application research on personalized recommendation of books based on user portrait and group portrait [J]. Information Studies: Theory & Application, 2019, 42(1): 129-133, 160 (in Chinese)
(何娟. 基于用户个人及群体画像相结合的图书个性化推荐应用研究[J]. 情报理论与实践, 2019, 42(1): 129-133, 160)

[27] Minn S, Yu Yi, Desmarais M C, et al. Deep knowledge tracing and dynamic student classification for knowledge tracing [C] //Proc of the 18th IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2018: 1182-1187

[28] Yao Huaxiu, Nie Min, Su Han, et al. Exercise-enhanced sequential modeling for student performance prediction [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 2435-2443

[29] Agrawal R, Golshan B, Terzi E. Grouping students in educational settings [C] //Proc of the 20th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1017-1026

[30] Yao Huaxiu, Nie Min, Su Han, et al. Predicting academic performance via semi-supervised learning with constructed campus social network [G] //LNCS 10178: Proc of the Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2017: 27-30

[31] Ashay T, Shajith I, Bikram S, et al. Predicting student risks through longitudinal analysis [C] //Proc of the 20th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1544-1552

[32] Cetintas S, Si Luo, Xin Yanping, et al. Probabilistic latent class models for predicting student performance [C] //Proc of the 22nd ACM Int Conf on Information and Knowledge Management. New York: ACM, 2013: 1513-1516

[33] Chen Yuying, Liu Qi, Huang Zhenya, et al. Tracking knowledge proficiency of students with educational priors [C] //Proc of the 2017 ACM Int Conf on Information and Knowledge Management. New York: ACM, 2017: 989-998

Wang Xin'ao, born in 1996. Master candidate. His main research interests include data mining and knowledge engineering.

Duan Lei, born in 1981. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data mining, health-informatics and evolutionary computation.

Cui Dingshan, born in 1995. Master candidate. Her main research interests include data mining and knowledge engineering.

Lu Li, born in 1972. PhD, professor. Her main research interests include computational medicine, environmental medicine, biometrics.

Dun Yijie, born in 1975. Master, associate professor. Member of CCF. Her main research interests include educational data mining and knowledge engineering.

Qin Ruiqi, born in 1995. Master in the School of Computer Science, Sichuan University. Her main research interests include bioinformatics and data mining.