



虚拟空间中在线同源用户行为相似性研究

马满福^{1,2}, 张凯旋^{1,2}, 李勇^{1,2}, 王常青³, 张强^{1,2}

(1. 西北师范大学 计算机科学与工程学院, 兰州 730070; 2. 甘肃省物联网工程研究中心, 兰州 730070;

3. 中国互联网络信息中心 互联网基础技术开放实验室, 北京 100190)

摘要: 虚拟空间中在线同源用户具有相似行为特征, 但现有相似性度量算法难以对其进行有效识别。提出一种基于序列对齐的在线同源用户识别算法, 根据在线用户行为日志提取点击流数据, 采用序列对齐方法计算在线用户的行为相似度, 将其用行为相似度矩阵表示并对用户进行层次聚类, 以识别虚拟空间中的在线同源用户, 同时分析不同维度的用户特征属性对用户行为相似性的影响程度。实验结果表明, 该算法能准确识别出在线同源用户, 用户行为相似性受性别、户籍和教育程度3种特征属性影响较大, 受年龄、社会阶层和收入水平的影响较小。

关键词: 行为特征; 在线同源用户; 序列对齐; 行为相似性; 特征属性

开放科学(资源服务)标志码(OSID):



中文引用格式: 马满福, 张凯旋, 李勇, 等. 虚拟空间中在线同源用户行为相似性研究[J]. 计算机工程, 2021, 47(5): 65-72.

英文引用格式: MA Manfu, ZHANG Kaixuan, LI Yong, et al. Research on behavioral similarity among online homologous users in virtual space[J]. Computer Engineering, 2021, 47(5): 65-72.

Research on Behavioral Similarity among Online Homologous Users in Virtual Space

MA Manfu^{1,2}, ZHANG Kaixuan^{1,2}, LI Yong^{1,2}, WANG Changqing³, ZHANG Qiang^{1,2}

(1. College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China;

2. Gansu Provincial IoT Engineering Research Center, Lanzhou 730070, China;

3. Domain Named System Laboratory, China Internet Network Information Center, Beijing 100190, China)

[Abstract] Online homologous users in virtual space have similar behavior characteristics, but the existing similarity measurement algorithms are difficult to effectively identify them. To address the problem, this paper proposes an online homologous user identification algorithm based on sequence alignment. The click stream data is extracted from online user behavior logs, and then processed by using the sequence alignment method to calculate the behavioral similarity of online users, which is represented as the behavioral similarity matrix. On this basis, the hierarchical clustering is carried out for users to verify the existence of online homologous users in virtual space. At the same time, the influence of different dimensions of user characteristics on their behavioral similarity is analyzed. Experimental results show that the proposed algorithm can identify online homologous users accurately, and the similarity of user behavior is principally influenced by gender, registered residence and education level, and is less affected by age, social class and income level.

[Key words] behavior characteristic; online homologous users; sequence alignment; behavioral similarity; feature attribute

DOI: 10.19678/j.issn.1000-3428.0058042

0 概述

由于生物遗传和变异, 自然界中存在大量性状

相同的物种, 其在进化上或个体发育上因具有共同来源呈现出的相似性称为同源性。这种同源性被广泛应用于医疗健康、生物制药和遗传研究等诸多领域

基金项目: 国家自然科学基金(71764025, 61863032, 61662070); 甘肃省高等学校科学研究项目(2018A-001); 甘肃省教育科学规划课题项目(GS[2018]GHBBKZ021, GS[2018]GHBBKW007)。

作者简介: 马满福(1968—), 教授、博士, 主研方向为大数据、机器学习; 张凯旋, 硕士研究生; 李勇(通信作者), 副教授、博士; 王常青, 高级工程师、博士; 张强, 教授、博士。

收稿日期: 2020-04-13 **修回日期:** 2020-06-01 **E-mail:** 384379992@qq.com

域^[1-2],也为虚拟空间中在线用户行为相似性研究提供了新思路。在虚拟空间中,不同用户群体的行为特征通常存在不同程度的差异性^[3-4]和相似性^[5-6]。文献[7]基于用户主题感知和行为相似性分析动态用户的相关性,指出同种社区类型的用户具有强相关性,不同社区类型的用户具有弱相关性。文献[8]对16个国家微博用户行为的差异性和相似性进行研究,发现在人口少且凝聚力强的国家,用户更关注微博的社会功能,而在人口较多的国家,用户仅将微博作为新闻传播平台。

在证实虚拟空间中用户行为特征具有差异性和相似性的基础上,研究人员结合在线用户的自身特征属性给出部分应用场景^[9-11]。文献[12]提出一种基于同义词组的用户行为汇聚方法,利用汇聚结果对用户进行性别预测,证明不同性别群体的兴趣具有差异性,该方法能根据用户性别进行有效的个性化系统推荐。文献[13]通过调查欧洲60 000多名工人的收入、教育程度、职业类型、自治水平、时间压力和社会互动6个维度的信息,提出双变量有序概率计量经济模型以衡量互联网对工人工作满意度的影响,该研究对提高企业管理水平具有重要意义。

用户点击路径反映出用户在一段时间内点击的页面和驻留时间^[14],分析用户的点击行为是研究用户行为相似性的有效方法^[15]。目前关于用户行为相似性缺乏统一量化标准,对虚拟空间中在线同源用户(根据同源理论,即点击序列相似度超过30%的在线用户)是否存在也未有验证。此外,关于不同特征属性对在线同源用户行为相似性影响程度的研究也较少。因此,本文提出一种虚拟空间中在线同源用户识别算法。从在线用户行为数据集中提取点击流数据,采用序列对齐方法处理点击流数据以度量在线用户的行为相似性。同时从数据集的人口统计信息中获取在线用户不同维度的特征属性,研究各种特征属性对在线同源用户行为相似性的影响程度。

1 数据描述

用户的在线行为主要通过点击流数据来体现。本文采用中国互联网信息中心(China Internet Network Information Center, CNNIC)提供的在线行为日志作为数据集(以下称为CNNIC数据集)进行研究,其中数据要素包含每个用户的点击路径以及每个路径对应的点击时刻,点击时刻采用标准时间格式记录。CNNIC数据集中某用户的部分点击流数据如表1所示。

表1 原始点击流数据

Table 1 Raw click stream data

点击时刻	点击路径
08-01 20:29:57	['explorer.exe']
08-01 20:29:59	['AliIM.exe']
08-01 20:30:05	['SohuNews.exe']
08-01 20:30:23	['360SE.exe']
08-01 20:30:27	['AliIM.exe']
08-01 20:30:31	['360SE.exe']
08-01 20:30:59	['AliIM.exe']

本文主要研究在线同源用户的识别及特征属性对其行为的影响程度,因此用户特征属性提取是关键。利用上述数据集中的人口统计信息提取用户的年龄、社会阶层、教育程度、性别、户籍和收入水平6个维度的特征属性,部分用户的人口统计信息如表2所示。对每个特征属性进一步分类,结果如表3所示。

表2 部分用户的人口统计信息

Table 2 Demographic information of partial users

用户	人口统计信息
User1	{男,青年,高等教育,社会中等,中等收入,城市}
User2	{女,青年,高等教育,社会中等,中等收入,城市}
User3	{男,青年,高等教育,社会中等,最高收入,乡村}
User4	{男,少年,中等教育,学生,低收入,乡村}

表3 特征属性分类

Table 3 Classification of feature attributes

特征属性	属性类别
年龄	{少年,青年,中年,老年}
社会阶层	{社会下层,社会中层,社会高层,学生}
教育程度	{初等教育水平,中等教育水平,高等水平}
性别	{男,女}
户籍	{乡村,城郊,城市}
收入水平	{低水平收入,中等收入,较高收入,最高收入}

2 研究方法

本文提出基于序列对齐的在线同源用户识别(Sequence Alignment-based Online Homologous User Recognition, SA-OHUR)算法,其主要包括以下步骤:1)处理点击行为数据;2)基于序列对齐思想计算在线用户的行为相似度,并对其以相似度矩阵形式进行量化;3)根据行为相似度矩阵对用户进行聚类验证并识别在线同源用户。此外,采用基于特征属性的方法计算聚类结果的熵值和纯度,并由此分析在线用户特征属性对其行为的影响程度。

2.1 在线同源用户识别算法

在线用户行为由一系列点击路径及其对应的点击时刻构成,若将每个点击路径看作用户点击序列中一个字符串,则在点击流数据中点击路径和对应路径花费的时间可反映用户的点击行为,其用包含

时间的字符串序列表示。例如,表1中点击流数据对应的该用户点击序列 $S_p = \{(['explorer.exe'], 08-01 20: 29: 57), (['AliiM.exe'], 08-01 20: 29: 59), (['SohuNews.exe'], 08-01 20: 30: 05), (['360SE.exe'], 08-01 20: 30: 23), (['AliiM.exe'], 08-01 20: 30: 27), (['360SE.exe'], 08-01 20: 30: 31), (['AliiM.exe'], 08-01 20: 30: 59)\}$ 。用户行为相似度计算问题可转换为编辑距离的问题。

2.1.1 序列对齐方法

序列对齐也称编辑距离,主要通过对齐的方法来度量两个序列的相似性^[16],其核心思想是利用一个序列转换为另一个序列所花费的最小代价衡量两个序列的相似性。序列 Q 和序列 C 之间的编辑距离和相似度分别定义为:

$$d_{SAM(Q,C)} = (\omega_d D + \omega_i I) + \mu R \quad (1)$$

$$S_{SAM(Q,C)} = 1 - \frac{\mu_{SAM(Q,C)}}{|Q| + |C|} \quad (2)$$

其中, $d_{SAM(Q,C)}$ 为序列 Q 和序列 C 之间的编辑距离, $S_{SAM(Q,C)}$ 为序列 Q 和序列 C 之间的相似度, D , I 和 R 分别为转换过程中删除、插入和重排的次数, $|Q|$ 和 $|C|$ 分别为序列 Q 和序列 C 的长度, ω_d , ω_i 和 μ 分别为序列 Q 转换为序列 C 过程中删除、插入和重排操作的代价,且均为用户给定的正常数。

2.1.2 数据预处理

本文基于序列对齐思想处理持续点击流数据,具体步骤如下:

1) 计算在线用户在每个点击路径的持续时间,当前点击路径的持续时间即为当前点击时刻与前一个点击时刻之差,若某一个点击路径的持续时间超过 30 min,则默认为用户已经下线,并将该点击路径及其持续时间从用户点击序列中去除,即会话时间间隔阈值定义为 30 min^[17],处理后的持续点击流数据如表4所示。

表4 持续点击流数据

Table 4 Continuous click stream data

持续时间/s	点击路径
2	['explorer.exe']
6	['AliiM.exe']
18	['SohuNews.exe']
4	['360SE.exe']
4	['AliiM.exe']
28	['360SE.exe']
14	['AliiM.exe']

2) 记录用户一个月内的点击路径并计算其对应的持续时间,处理后的累计点击流数据如表5所示,用户累计点击序列 $S_U = \{(['explorer.exe'], 2), (['AliiM.exe'], 24), (['SohuNews.exe'], 18), (['360SE.exe'], 32)\}$ 。

表5 累计点击流数据

Table 5 Cumulative click stream data

累计时间/s	累计点击路径
2	['explorer.exe']
24	['AliiM.exe']
18	['SohuNews.exe']
32	['360SE.exe']

2.1.3 在线用户行为相似度算法

本文提出的 SA-OHUR 算法是利用基于序列对齐的在线用户行为相似度算法获得用户间相似度。由于该算法所用累计点击序列的时间为累计时间,因此不考虑点击路径的先后顺序,即转换过程中重排操作代价为 0。同时,若两个用户点击路径相同但对应路径的累计时间不同,则可能造成点击行为的差异,因此,增加两个在线用户点击的相同路径所对应累计时间差值的绝对值作为补偿操作。设在线用户 U_i 的点击序列 $S_{U_i} = \{(a_{i1}, T_{i1}), (a_{i2}, T_{i2}), \dots, (a_{im}, T_{im})\}$, 在线用户 U_j 的点击序列 $S_{U_j} = \{(a_{j1}, T_{j1}), (a_{j2}, T_{j2}), \dots, (a_{jm}, T_{jm})\}$ 。其中, $(a_{i1}, a_{i2}, \dots, a_{im})$ 与 $(a_{j1}, a_{j2}, \dots, a_{jm})$ 分别为在线用户 U_i 和 U_j 的点击路径集 A_i 和 A_j 。 $(T_{i1}, T_{i2}, \dots, T_{im})$ 与 $(T_{j1}, T_{j2}, \dots, T_{jm})$ 分别为在线用户 U_i 和 U_j 的累计时间集 T_i 和 T_j 。在线用户 U_i 和 U_j 基于序列对齐的编辑距离定义为:

$$d_{SAM(U_i, U_j)} = \sum_{p=1}^n (T_{ip} \times \text{sgn}(a_{ip} \notin A_j)) + \sum_{q=1}^m (T_{jq} \times \text{sgn}(a_{jq} \notin A_i)) + \sum_{p=1, q=1}^{n, m} (|T_{ip} - T_{jq}| \times \text{sgn}(a_{ip} = a_{jq})) \quad (3)$$

其中,删除和插入的代价分别为删除和插入路径所对应的累计时间, $|T_{ip} - T_{jq}|$ 为补偿操作的代价。

两个用户基于序列对齐的行为相似度计算公式为:

$$S_{SAM(U_i, U_j)} = 1 - \frac{d_{SAM(U_i, U_j)}}{\sum_{p=1}^n T_{ip} + \sum_{q=1}^m T_{jq}} \quad (4)$$

其中,当用户点击序列(点击路径及其对应的累计时间)完全相同时,用户的相似度为 1,当点击序列完全不同时,相似度为 0。在线用户 U_i 和 U_j 的行为相似度计算如算法1所示。

算法1 基于序列对齐的用户行为相似度算法

输入 n 个用户的累计点击序列 $S_{U_1}, S_{U_2}, \dots, S_{U_n}$

输出 在线用户点击行为相似度矩阵 $A_{n \times n}$

/*循环遍历 n 个用户的是点击序列,计算每一个用户与其他用户的行为相似度*/

```

1. i=0, j=0, T=0, S={}
2. For i in range(0, n):
3. For j in range(i, n):
4. If  $a_{ip}$  not in  $A_j$ :
5.  $T=T+T_{ip}$ 
6. End If

```



```

7.If  $a_{jq}$  not in  $A_i$ :
8. $T=T+T_{jq}$ 
9.End If
10.If  $a_{ip} = a_{jq}$ :
11. $T=T+|T_{ip} - T_{jq}|$ 
12.End If
13. $d_{SAM}(U_i, U_j) = T$ 
14. $S_{SAM}(U_i, U_j) = 1 - \frac{d_{SAM}(U_i, U_j)}{\sum_{p=1}^n T_{ip} + \sum_{q=1}^m T_{jq}}$ 
15.S.append( $S_{SAM}(U_i, U_j)$ )
16.End For
17.End For
18.Return S
/*将相似度转换成  $n \times n$  的相似度矩阵, 矩阵中第  $i$  行第  $j$  列
表示第  $i$  个用户与第  $j$  个用户的行为相似度*/
19.SAM_matrix=numpy.zeros(n, len(S))
20.For x in S:
21. $A_{n \times n}[:, x]=1$ 
22.End For
23.Return  $A_{n \times n}$ 

```

上述算法在用户行为相似度计算过程中, 主要利用用户累计点击流数据, 且无需考虑点击顺序。在处理点击流数据时, 将点击序列按照点击路径进行扫描, 可得到用户之间的行为相似度。由于在数据处理阶段已去除冗余点击路径, 因此与传统的序列对齐算法相比, 算法1复杂度大幅降低。

2.1.4 基于行为相似度矩阵的层次聚类

SA-OHUR算法最后一步是根据相似度矩阵对在线用户进行聚类, 以验证在线同源用户的存在。为更直观地区分出用户在线行为并识别同源用户群, 该算法采用基于行为相似度矩阵的层次聚类。由于传统层次聚类HC算法每进行一次簇间合并均需更新相似度矩阵, 造成算法步骤重复, 因此为避免该问题, SA-OHUR算法将相似度矩阵中在线用户之间相似度值和用户编号采用数组的形式按照相似度值进行降序排列, 根据相似度值在数组中的位置从大到小合并用户, 即引入优先级队列。

SA-OHUR算法将在线用户按照点击行为划分为不同类别, 具体流程如下: 1) 初始化每个用户作为单独的簇; 2) 根据相似度矩阵将相似度值及其对应的用户存入已定义的数组并按照降序排列; 3) 合并数组中第1个相似度值, 将最大相似度值对应的两个用户作为一个簇; 4) 从第二轮合并开始, 若相似度值对应的两个用户均未合并到某个簇中, 则这两个用户合并为一个簇; 若其中一个用户已合并到另外一个簇中, 则将另一个用户也合并到该簇中; 若两个用户分别合并到不同簇中, 则这两个用户所在的两个簇合并; 5) 按顺序取数组 N 的相似度值, 且在每轮合并时簇的个数减少1; 6) 重复步骤4和步骤5直到生成 K 个簇。

给定在线用户集 $U=\{u_1, u_2, \dots, u_n\}$, 将其根据点

击行为相似性划分 K 个类 C_1, C_2, \dots, C_K , 要求每个类别不能为空且类与类之间用户不相同, 主要步骤如算法2所示。

算法2 基于行为相似度矩阵的层次聚类算法

输入 在线用户行为相似度矩阵 $A_{n \times n}$

输出 聚类结果 C_1, C_2, \dots, C_K

1.For $S_{SAM}(U_i, U_j)$ in $A_{n \times n}$:

2.N.append($i, j, S_{SAM}(U_i, U_j)$)/其中元素分别为行号, 列号,

相似度值/

3.End For

4.Max($N=(a, b, S_{SAM}(U_a, U_b))$)/ $S_{SAM}(U_a, U_b)$ 为最大相似度值/

5. $C_{ab}=(U_a, U_b)$ /将相似度值最大的两个用户 a, b 合并到一个簇中/

/*基于行为相似度矩阵的层次聚类算法流程将用户进行聚类*/

6.count=n

7.while(count>K):

8.For item in N:

9.if U_p in C_p and U_q in C_q :

10. $C_{pq}=(C_p, C_q)$

11.count= count-1

12.if U_p not in C_p and U_q in C_q

13. $C_{pq}=(U_p, C_q)$

14.count= count-1

15.if U_p in C_p and U_q not in C_q

16. $C_{pq}=(C_p, U_q)$

17.count= count-1

18.if U_p not in C_p and U_q not in C_q

19. $C_{pq}=(U_p, U_q)$

20.count= count-1

21.End For

22.Return C_1, C_2, \dots, C_K

在算法2中, 先对 n 个用户的 $n \times (n-1)/2$ 个相似度进行快速降序排列, 排序的时间复杂度为 $O(n^2 \times \lg n)^{[18]}$, 再对 n 个用户根据相似度进行聚类, 该过程中聚类循环的时间复杂度为 $O(n)$ 。因此, 相较传统层次聚类的时间复杂度 $O(n^3)$, 算法2的时间复杂度降低为 $O(n^2 \times \lg n)$, 算法运行效率更高。

SA-OHUR算法的关键是计算簇间相似度和簇内相似度, 进而识别出在线同源用户群。由于每个簇即为在线用户集合, 因此本文采用簇内在线用户与另一个簇内在线用户的平均相似度来表示。例如, 给定聚类簇 C_i 和 C_j , 则两个簇间的相似度定义为:

$$S_{avg}(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{z \in C_j} S_{SAM}(x, z) \quad (5)$$

其中, $S_{avg}(C_i, C_j)$ 为聚类簇 C_i 和 C_j 的相似度, x 为属于聚类簇 C_i 的用户, z 为属于聚类簇 C_j 的用户, $S_{SAM}(x, z)$ 为在线用户 x 和 z 的行为相似度, $|C_i|$ 和 $|C_j|$ 分别为聚类簇 C_i 和 C_j 的在线用户个数。

2.2 基于特征属性的熵值和纯度算法

采用SA-OHUR算法对在线用户进行聚类时, 为更好地将在线用户特征属性与其行为相似性进行结

合,可用熵值和纯度来衡量用户特征属性对其行为相似性的影响程度。熵值和纯度可用来衡量某个指标的混乱度^[19],本文分别计算基于表3中6个不同特征属性下点击行为相似度聚类结果的熵值和纯度,由此判定特征属性对用户行为相似性的影响程度。

给定 n 个在线用户,根据在线用户行为相似度将其分为 K 个簇,其中每个簇分别包含 n_1, n_2, \dots, n_K 个用户。假设某个特征属性有 M 个类别,如教育程度分为初等教育水平、中等教育水平、高等教育水平3个类别,则在该特征属性下聚类簇 i 的熵值计算公式为:

$$e_i = - \sum_{j=1}^M \frac{n_{ij}}{n_i} \lg \frac{n_{ij}}{n_i} \quad (6)$$

在该特征属性下聚类的整体熵值计算公式为:

$$e = \sum_{i=1}^K \frac{n_i}{n} e_i \quad (7)$$

在该特征属性下聚类簇 i 的纯度计算公式为:

$$p_i = \max \left(\frac{n_{ij}}{n_i} \right) \quad (8)$$

在该特征属性下聚类的整体纯度计算公式为:

$$p = \sum_{i=1}^K \frac{n_i}{n} p_i \quad (9)$$

其中, n_{ij} 表示聚类簇 i 中用户属于类别 j 的个数, n_i 为聚类簇 i 中所有用户个数, n 为参加聚类的所有用户个数。基于特征属性的熵值和纯度计算如算法3所示。

算法3 基于特征属性的熵值和纯度算法

输入 某一个特征属性的 M 个类别包含用户集 m_1, m_2, \dots, m_M ; K 个聚类结果簇 C_1, C_2, \dots, C_K

输出 该特征属性下的熵值 e 和纯度 p

/*统计聚类簇分属每个类别的用户数*/

1. For i range(0, K):

2. For j range(0, M):

3. For a range(0, n):

4. If U_a in C_i and m_j :

5. $n_{ij} = n_{ij} + 1$

6. If U_a in C_i :

7. $n_i = n_i + 1$

8. End For

/*计算聚类结果每个簇的熵值和纯度*/

9. For i range(0, K):

10. For j range(0, M):

11. $e_i = e_i + \frac{n_{ij}}{n_i} \lg \frac{n_{ij}}{n_i}$

12. If $p_i < \frac{n_{ij}}{n_i}$:

13. $p_i = \frac{n_{ij}}{n_i}$

14. Return e_i, p_i

15. End For

/*计算聚类结果整体的熵值和纯度*/

16. For i range(0, K):

17. $e = e + \frac{n_i}{n} e_i$

18. $p = p + \frac{n_i}{n} p_i$

19. Return e, p

算法3是通过聚类结果的熵值和纯度衡量特征属性对行为相似性的影响程度。若基于某一个特征属性计算得到的聚类结果熵值越小,混乱程度越低,该特征属性下类别分散程度越小,则基于该属性聚类结果的综合评价越好,即特征属性对同源用户行为相似性的影响程度越大。而纯度相反,若基于某一个特征属性计算得到的聚类结果纯度越大,混乱程度越低,该特征属性下的类别分散程度就越小,则基于该属性聚类结果的综合评价越好,即特征属性对用户行为相似性的影响程度越大。

3 实验与结果分析

本文抽取848名用户一个月内约1.2亿条点击流数据进行分析,实验采用Windows 8操作系统和8 GB运行内存并通过Python3.6实现。

3.1 结果分析

按照SA-OHUR算法流程,本文将点击流数据进行处理后得到在线用户累计点击行为序列。例如,在线用户 U_a 的累计点击行为序列 $S_{U_a} = \{(['explorer.exe'], 2), (['AliIM.exe'], 24), (['SohuNews.exe'], 18), (['360SE.exe'], 32)\}$, 在线用户 U_b 的累计点击行为序列 $S_{U_b} = \{(['explorer.exe'], 2), (['AliIM.exe'], 34), (['xmp.exe'], 5)\}$, 并由式(4)计算得到用户 U_a 与 U_b 的相似度如下:

$$S_{(U_a, U_b)} = 1 - \frac{18+32+5+|24-34|}{2+24+18+32+2+34+5} = 0.44 \quad (10)$$

采用算法1得到848名用户间相似度并将结果以相似度矩阵 A 输出,表达式如下:

$$A = \begin{bmatrix} 1.00 & 0.63 & 0.01 & \dots & 0.58 & 0.65 \\ 0.63 & 1.00 & 0.28 & \dots & 0.05 & 0.38 \\ 0.01 & 0.28 & 1.00 & \dots & 0.86 & 0.31 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.58 & 0.05 & 0.86 & \dots & 1.00 & 0.22 \\ 0.65 & 0.38 & 0.31 & \dots & 0.22 & 1.00 \end{bmatrix} \quad (11)$$

相似度矩阵 A 是一个 848×848 对称矩阵,其中第 i 行第 j 列的数值表示第 i 个在线用户和第 j 个在线用户的点击行为相似度,对角线元素表示每个在线用户与自身行为的相似度,相似度值均为1,在该矩阵中相似度取值分布范围为 $0 \sim 1$ 。

由相似度矩阵 A 得到在线用户不同相似度区间数量统计如图1所示。其中, x 轴为相似度值, y 轴为投影在该区间相似度值的个数。图1中相似度值主要分布在 $(0.00, 0.60)$ 区间内,表明虚拟空间中存在行为相似度超过30%的在线同源用户,SA-OHUR算法能有效验证在线同源用户的存在。

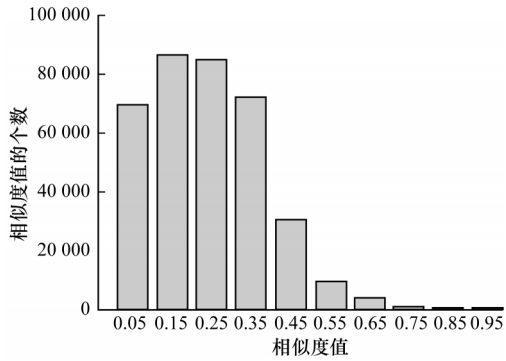


图1 不同区间相似度值统计情况

Fig.1 Statistics of similarity values in different intervals

为进一步识别在线同源用户群,采用SA-OHUR算法基于相似度矩阵 A 和算法2对用户进行聚类。不同聚类簇个数下的簇间相似度值和簇内相似度值的对比如图2所示。可以看出,随着聚类簇个数的增加,簇内相似度值逐步上升并最终稳定在(0.4,0.5)区间,而簇间相似度值虽然呈现上升趋势但始终低于簇内相似度值,且最大值不超过0.3。这表明属于同一个簇的在线用户即为在线同源用户且其点击行为相似度超过40%,而属于不同簇的在线用户即为

在线非同源用户,采用SA-OHUR算法能有效识别在线同源用户群。识别出在线同源用户后,可根据表3中用户特征属性类别,采用SA-OHUR算法将用户分为2个簇、3个簇和4个簇,并利用算法3研究特征属性对在线同源用户行为相似性的影响程度。不同特征属性下各个簇及聚类结果整体的熵值和纯度如图3所示。

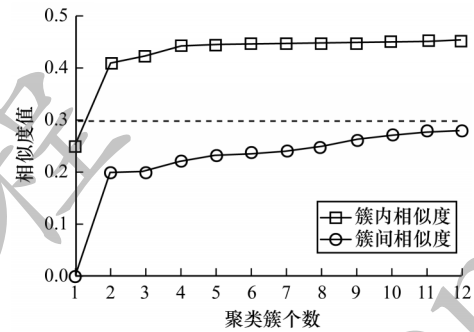


图2 簇间相似度值和簇内相似度值的对比

Fig.2 Comparison of similarity values between clusters and similarity values within clusters

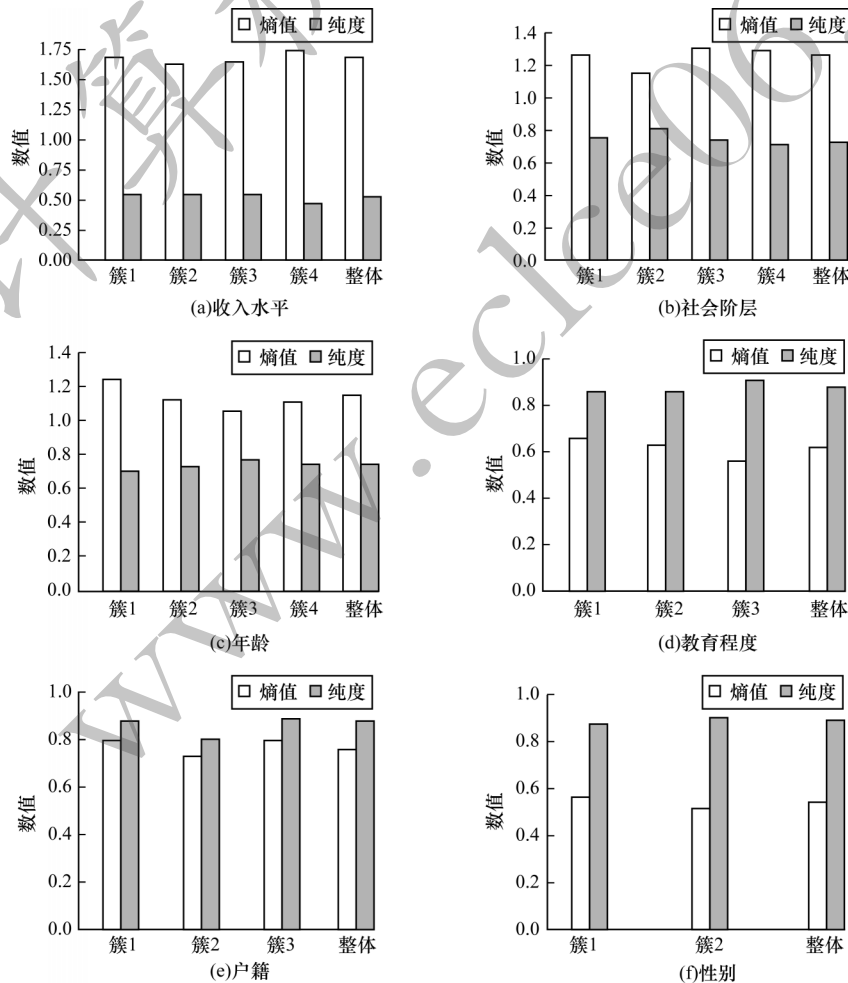


图3 不同特征属性下聚类结果的熵值和纯度

Fig.3 Entropy values and purity of clustering results under different feature attributes

由图3可以看出,基于性别、户籍和教育程度的聚类结果整体熵值分别为0.541、0.754和0.622,其结果低于基于收入水平、社会阶层和年龄的聚类结果(熵值分别为1.689、1.259和1.144),表明基于性别、户籍和教育程度的聚类结果分散程度低且综合评价较好,即该3类特征属性对在线同源用户行为相似性影响较大。基于性别、户籍和教育程度的聚类结果整体纯度分别为0.890、0.872和0.878,其结果高于基于收入水平、社会阶层和年龄的聚类结果(纯度分别为0.517、0.740和0.732),表明基于性别、户籍和教育程度的聚类结果纯度较高且混乱度较低,这3类特征属性对在线同源用户行为相似性影响更大。综上可知,性别、户籍和教育程度3种特征属性对在线同源用户行为相似性的影响程度大于收入水平、社会阶层和年龄的影响程度,其中影响最高的特征属性为性别,影响最低的特征属性为收入水平。

3.2 对比实验

为验证SA-OHUR算法的时间复杂度,本文采用时间序列相似性度量(DTW)^[18]和莱文斯坦相似性度量(Leven)^[20]两种经典的相似性度量算法,分别计算点击流数据中前200名、400名、600名和800名在线用户的相似度,3种算法运行时间如表6所示。

表6 3种算法的运行时间对比

Table 6 Running time comparison of three algorithms

算法	用户数为 200	用户数为 400	用户数为 600	用户数为 800
SA-OHUR算法	117.2	462.7	806.1	1 438.8
DTW算法	109.3	463.3	983.0	1 745.9
Leven算法	182.4	182.4	1 640.0	2 675.0

由表6可知,SA-OHUR算法在一定程度上减少程序运行时间,提升了程序运行效率,在处理大批量数据时该算法有明显优势。这是因为SA-OHUR算法采用累计点击数据流进行计算,无需考虑累计点击数据流中序列的顺序性,同时去除冗余序列,降低了算法复杂度。

相较传统层次聚类HC算法,SA-OHUR算法降低了时间复杂度,提高了运行效率,但其聚类效果还未知。因此,本文将采用传统层次聚类HC算法和SA-OHUR算法所得聚类结果的熵值和纯度进行对比,结果分别如表7和表8所示。

表7 2种算法不同特征属性的熵值对比

Table 7 Comparison of entropy values of different feature attributes of two algorithms

算法	性别	教育程度	户籍	年龄	社会阶层	收入水平
SA-OHUR算法	0.54	0.62	0.75	1.14	1.26	1.69
HC算法	0.61	0.70	0.85	1.23	1.12	1.86

表8 2种算法不同特征属性的纯度对比

Table 8 Comparison of purity of different feature attributes of two algorithms

算法	性别	教育程度	户籍	年龄	社会阶层	收入水平
SA-OHUR算法	0.89	0.88	0.87	0.74	0.73	0.52
HC算法	0.83	0.78	0.69	0.73	0.76	0.44

由表7和表8可知,SA-OHUR算法得到的聚类结果整体熵值较低且纯度较大,其中在社会阶层属性中较反常。从整体来看,年龄对行为相似性影响程度低于性别、教育程度、户籍3种属性,对结果影响不大。因此,在分析特征属性对在线同源用户行为相似性影响程度时,基于相似度矩阵的层次聚类整体效果更好。

4 结束语

利用海量的互联网信息找出在线用户行为的主要影响因素,并据此对不同用户群体进行分类具有重要意义。本文基于序列对齐技术提出一种在线同源用户识别算法,提取在线用户点击流数据和特征属性,采用序列对齐方法计算用户行为相似度,识别具有相似行为的在线同源用户,并分析不同特征属性对用户行为相似性的影响程度。实验结果表明,该算法能有效区分在线同源用户,用户行为相似性受性别、户籍和教育程度3种特征属性影响较大。本文主要研究独立的特征属性,未考虑不同特征属性组合对用户行为的影响,后续将从用户行为权值较大的部分特征属性入手,进一步研究包含该部分属性不同组合的用户行为。

参考文献

- [1] QUIGNOT C, REY J, YU J C, et al. InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs [J]. *Nucleic Acids Research*, 2018, 46(1):408-416.
- [2] RAJAN D R, CHANDRAMOHAN B. Annotation transfer by homology among closely related genomes helps to identify protein function in plasmodium species [J]. *International Journal of Bioinformatics Research and Applications*, 2017, 13(1):22-39.
- [3] NGUYEN T H, TRAN D Q, DAM G M, et al. Estimating the similarity of social network users based on behaviors[J]. *Vietnam Journal of Computer Science*, 2018, 5(2):165-175.
- [4] HUANG Hua, PENG Rong, FENG Zaiwen. A time-aware method to process behavioral similarity calculation[C]// *Proceedings of 2016 IEEE International Conference on Services Computing*. Washington D. C., USA: IEEE Press, 2016:31-37.
- [5] CHEN Jian, ZHOU Xiaokang, JIN Qun. Recommendation of optimized information seeking process based on the

- similarity of user access behavior patterns[J]. *Personal and Ubiquitous Computing*, 2013, 17(8): 1671-1681.
- [6] GAO Maoting, WANG Ji. Topic model recommendation algorithm combining user social relationships and time factors[J]. *Computer Engineering*, 2020, 46(3): 66-72. (in Chinese)
高茂庭, 王吉. 融合社交关系与时间因素的主题模型推荐算法[J]. *计算机工程*, 2020, 46(3): 66-72.
- [7] ZHOU Xiaokang, WU Bo, JIN Qun, et al. Analysis of user network and correlation for community discovery based on topic-aware similarity and behavioral influence[J]. *IEEE Transactions on Human-Machine Systems*, 2018, 48(6): 559-571.
- [8] CHEN L, NUGENT C D. Human activity recognition and behaviour analysis: for cyber-physical systems in smart environments[M]. Berlin, Germany: Springer, 2019.
- [9] CASO A, ROSSI S. Users ranking in online social networks to support POI selection in small groups[EB/OL]. [2020-03-29]. https://www.researchgate.net/publication/263779579_Users_Ranking_in_Online_Social_Networks_to_Support_POI_Selection_in_Small_Groups.
- [10] SI Jianfeng, LI Qing, QIAN Tieyun, et al. Users' interest grouping from online reviews based on topic frequency and order[J]. *World Wide Web*, 2014, 17(6): 1321-1342.
- [11] GAO Jian, ZHOU Tao. Evaluating user reputation in online rating systems via an iterative group-based ranking method[J]. *Physica A: Statistical Mechanics and Its Applications*, 2017, 473(5): 546-560.
- [12] FENG Tingting, GUO Yuchun, CHEN Yishuai. A novel user behavioral aggregation method based on synonym groups in online video systems[J]. *Science China Information Sciences*, 2016, 59(2): 1-3.
- [13] CASTELLACCI F, VINASBARDOLET C. Internet use and job satisfaction[J]. *Computers in Human Behavior*, 2019, 90(1): 141-152.
- [14] ZHANG Yanchun, XU Guandong. On Web communities mining and recommendation[J]. *Concurrency and Computation: Practice & Experience*, 2009, 21(5): 561-582.
- [15] ZHAO Jie. Web usage mining based on granular computing[D]. Guangzhou: South China University of Technology, 2010. (in Chinese)
赵洁. 基于粒计算的Web使用挖掘研究[D]. 广州: 华南理工大学, 2010.
- [16] BOGUSZ M. Evolutionary approaches to sequence alignment[EB/OL]. [2020-03-29]. <https://dblp.uni-trier.de/db/>.
- [17] KUMAR R, TOMKINS A. A characterization of online browsing behavior[C]//Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM Press, 2010: 561-570.
- [18] MISHRA S, SHAFI Z, PATHAK S, et al. Time series event correlation with DTW and hierarchical clustering methods[EB/OL]. [2020-03-29]. <https://doi.org/10.7287/peerj.preprints.27959v1>.
- [19] UDDIN J, GHAZALI R, DERIS M M. Does number of clusters effect the purity and entropy of clustering?[C]//Proceedings of 2016 International Conference on Soft Computing and Data Mining. Berlin, Germany: Springer, 2016: 355-365.
- [20] PAVEL B, ILYA M, PANOS P, et al. Using levenshtein distance for typical user actions and search engine switching detection[EB/OL]. [2020-03-29]. https://www.onacademic.com/detail/journal_1000039473966110_f25d.html.

编辑 宋 圆