ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1101－1114
DOI:10. 16337/j. 1004-9037. 2022. 05. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

异构社交平台中用户身份解析

刘俊岭，刘　颖，马晨旭，赵巧娜，孙焕良，许景科

(沈阳建筑大学计算机科学与工程学院,沈阳 110168)

摘　要：跨社交平台的用户身份解析是社交网络一个重要的研究方向,其可以有效集成不同平台的同一用户信息。现有的用户身份解析工作大多针对类型相似的社交平台,平台间的信息相对对称,通过用户在不同平台上的档案属性、空间位置、网络关系等信息的相似度来判别是否为同一用户。然而,在两个异构社交平台中用户信息是不对称的,难以直接获取到用于用户身份解析的相应属性信息。本文研究跨评论类与活动类平台间的用户身份解析方法。为了解决两类社交平台的用户信息属性不对称问题,把用户信息按档案属性、语义序列、特征词序列3类信息组织,从各自的社交平台中抽取相应的信息建立映射关系,提出了综合3类信息的集成匹配算法。考虑了用户活动的时间偏移现象,采用反向传播学习的方法获取时间偏移权重,提出了基于反向传播学习的语义序列与特征词序列相似性度量方法。同时,设计了总体相似度度用于用户身份解析。利用真实数据集进行了充分的实验,实验结果表明了所提出用户身份解析算法的有效性。

关键词：社会网络;用户身份解析;特征词序列;语义序列

中图分类号：TP311.13　　**文献标志码**：A

User Identity Resolution Across Heterogeneous Social Platforms

LIU Junling, LIU Ying, MA Chenxu, ZHAO Qiaona, SUN Huanliang, XU Jingke

(School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

Abstract: The identity resolution across social platforms is an important research aspect, which integrates the user's information from various platforms. Most of the existing user identity resolution work is aimed at social platforms with similar types. The information between platforms is relatively symmetrical. Whether the user is the same user is determined by the similarity of user's profile attributes, spatial location, network relations and other information on different platforms. However, in the two heterogeneous social platforms, the user information is asymmetric so that we cannot get the corresponding attribute information for user identity resolution. This paper discusses the method of user identity resolution across comment and activity platforms. To solve the problem of user information attribute asymmetry of across social platforms, the user information is organized according to three types of information: profile attribute, semantic sequence and feature word sequence. The corresponding information is extracted from their respective social platforms to establish mapping relationships, and an integrated matching algorithm integrating the

基金项目:国家自然科学基金(62073227);国家重点研发计划(2021YFF0306303);辽宁省自然科学基金(2019-MS-264);辽宁省教育厅项目(LJKZ0582);中国学位与研究生教育研究课题(2020MSA40)。
收稿日期:2021-08-23;**修订日期**:2022-01-21

three types of information is proposed. Considering the time offset phenomenon of user activities, the back propagation learning method is used to obtain the time offset weights, and a similarity measurement method between semantic sequence and feature word sequence based on back propagation learning is proposed. At the same time, an overall similarity is designed for user identity. Experimental results on real dataset show that the proposed method is effective on user identity resolution.

Key words: social network; user identity resolution; feature word sequence; semantic sequence

引　言

随着各类在线社交平台的快速发展,用户根据自身需要加入到提供不同服务的社交平台中。例如,人们在豆瓣上分享看过的电影、加入感兴趣的小组讨论分享主题相关体会,在微博上发表评论、抒发情感等,在微信、QQ上联络好友,分享动态等。根据社交平台的功能和类别可大体分为活动服务类(如豆瓣网)、评论类(如Twitter)、通信类(如Facebook)等。这些社交平台从不同角度记录了大量的用户活动信息,这些不同类型的信息反映了用户在互联网上的不同侧面,综合分析用户各类信息可以更好地实现好友推荐,观点发现等任务。

用户在加入多个社交平台时,通常使用不同的账户信息,并且处于多社交平台的同一用户关系不明确标记。将多个社交平台中同属一个真实用户的账号关联起来,称为异构社交平台用户身份解析,也称为同一用户匹配。社交平台中用户身份解析可以整合用户多维度的信息,有利于在社交平台中进行相关的分析任务[1-2]。

现有的用户身份解析工作大多针对类型相似的社交平台,平台间的信息相对对称[3-5]。文献[3]利用不同网络的社交信息、空间信息、时间信息、文本信息进行相似度计算,生成一个二分图,利用二分图匹配方法找出相应的用户。文献[4]在Twitter和Blogcatalog上根据两个网络中用户的好友在平台中发布文本的主题确定用户的核心兴趣,根据动态核心兴趣的演变识别用户。文献[5]从多个社交平台中收集用户共有的基本档案属性信息,将其转化为向量,通过计算多个网络间基本属性信息的相似性,判定是否是同一用户。

不同类型社交平台的通常是异构的,难以同时在两个异构平台获取到相对应的用户信息。例如,新浪微博中的信息大多是用户发表与转发的博文,重点体现用户观点与情绪的表达。豆瓣网则包含了用户参加的小组、看过的电影和书以及参加的同城活动等,体现了用户活动类信息。两类社交平台分别表达了用户思想活动与物理活动的两个不同方面信息,融合两类异质网络的社交行为数据会使用户的信息更加完整,对于发现用户行为模式、推荐等应用具有重要意义。

社交平台的异构性使得用户信息不对称和不完备,难以直接获取到用于用户身份解析的相应的属性信息。主要表现在以下方面:(1) 平台间的数据结构不对称,数据特征不完整,数据类型不同。例如,在豆瓣网中包含观看的电影、书籍等结构化信息,社会关系信息较弱,只有用户参加的小组信息。在微博网络中包含用户发表的评论与转发信息。(2) 同一网络中不同用户数据差别较大,存在数据稀疏性和缺失。这种差别由用户的个人习惯不同引起的,一些用户分享大量的信息,而一部分用户数据量很少。同时在同一平台中数据也存在差异,如豆瓣网中某些用户标记大量看过的电影和书信息,参加的小组却很少,而某些用户加入大量感兴趣的小组却鲜少标记看过的电影和书的信息。这使得同一平台中用户不同类型数据量不对称。由于以上挑战的存在,使得现有的用户身份解析方法难以直接应用解决不同类型的社交平台用户识别。

观察发现,社交平台用户的社交活动与评论在语义上在一段时间内具有稳定性,体现了该段时间

内的主题兴趣。同时,用户的语义特征会随时间发生变化,如一个用户一段时间会关注美食,而另一段时间则会关注运动,呈现出序列的特征。另外,在两个社交平台中用户语义特征可能存在偏移,根据在多个不同时间段内用户语义变化规律的相似性,来判定是否是同一用户,即用户语义序列相似性;另外,用户在活动网络中参加的活动名称对用户身份解析非常有价值,例如,一个用户的活动网络上看一部电影,电影名称为一个特征词,同时在评论类网络上发表了关于此电影的评论,在评论中出现了电影名称,则该电影名称可作为匹配同一用户的重要依据。

基于以上观察,本文将用户的语义信息与特征词信息组织成序列,通过对序列的匹配进行用户身份解析。语义序列与特征词序列的相似度计算是识别同一用户的关键。通常情况下,用户在多个社交平台上的活动或者评论不同步,例如在活动平台上看了电影,过一段时间在评论平台上才发表了评论。采用时间段一一对应的序列相似度计算难以反映这种时间上的偏移。本文考虑了用户活动在时间偏移现象,提出一种基于反向传播的序列相似度度量方法,采用反向传播学习的方法获取时间偏移权重。同时,设计了语义序列、特征词序列及用户档案属性的综合相似度计算方法,运用反向传播学习获重三者的权重,计算总体相似度后进行用户身份解析。爬取了两个社交平台中用户信息的真实数据,对所提出相似性度量进行实验。本文贡献包括:

(1) 提出了针对不对称异构社交平台的用户身份解析问题;

(2) 分析了用户在不同社交平台上的活动在时间上的偏移规律,提出了基于语义序列与特征词序列的用户身份解析方法;

(3) 采用反向传播学习方法解决了时间偏移对序列相似度计算影响;

(4) 利用真实数据集进行充分实验,验证了提出方法能够有效地识别社交平台用户身份。

1 相关工作

多社交平台用户身份解析相关工作可按所使用的信息进行划分,包括基于用户社交网络结构匹配的方法[1-2, 6-8],基于用户档案属性信息匹配的方法[9-14],基于用户行为信息匹配的方法等[4-5, 15-17],以及综合上述多种信息的集成匹配方法[3,18-20]。

(1)基于用户社交网络结构的用户识别方法。该方法将用户识别问题转化为网络间拓扑结构的链接预测问题。文献[6]依靠社交平台间的拓扑结构信息,实现跨网络用户身份识别,提出了一个分析社交平台隐私和匿名性的框架。文献[1]将用户识别转化为能量模型的求解过程。文献[2]研究发现同一用户可能在不同平台倾向建立部分相似的好友关系结构,提出基于朋友关系的用户识别算法。文献[8]提出利用多层图卷积网络进行用户链接预测。

基于网络拓扑结构的用户识别方法适用于网络关系结构明确且拓扑结构信息较易获得的情况。本文的研究问题是在活动网络中的社交关系不可用,因此此类方法不适用于本文提出的问题。

(2)基于用户档案属性信息匹配的方法。用户档案属性信息主要包括用户名、爱好等基本属性。现有的基于用户档案属性信息的匹配方法可以分为基于用户名和基于多个属性信息的用户识别两类方法。文献[9]使用n-gram模型来度量不同用户名间的独特性以及同一用户名间的相似性,并使用编辑距离计算了用户名间的相似程度。文献[11]提出利用用户名与显示名称进行身份解析,将用户档案中的属性表示为多维向量,并采用分类法、赋权值法进行用户身份解析。文献[12]利用Facebook和Myspace网络中用户的职业信息、学历情况等数据,其将这些属性的词整合为单词集合,通过计算属性信息单词的相似度来衡量两个账户的相似度。文献[13]收集了用户的地址、爱好、工作单位和经历等属性信息,利用了从基础的均等评价模型到具有复发性的训练混合模型等方法来计算属性信息相似度,用于识别用户身份。文献[14]提出了训练集中反面案例识别方法。

由于结构的不对称,使得网络间属性信息不一致,利用上述基于属性的识别方法难以有效处理。同时,传统对用户名属性的处理通常将其看作字符串,基于字符串间的相似性匹配不适用于用户使用不同的用户名情况。

(3)基于用户行为信息匹配的方法。用户的行为信息从侧面反映了用户的思想情感、兴趣和习惯等,可用于用户识别。文献[15]提出通过匹配用户行为直方图识别来自不同平台的相同用户。文献[16]研究了用户标签行为的不一致性,通过改进BM25算法,基于标签的语义相似性,在候选集中选择出同一用户,实现跨网络间的用户身份匹配。文献[4]考虑了用户兴趣的有限性及动态性,结合用户的社交平台结构和基于内容的主题分析,提出一种用户动态核心兴趣匹配算法。文献[17]从用户生成的位置数据中提取行为模式,将行为轨迹转换成语义文档,利用LDA模型表示的用户主题进行用户相似度计算。

用户的行为轨迹数据通常较稀疏且存在大量的噪声,利用单一的行为轨迹数据进行用户识别效果难以保证。

(4)综合多种信息的集成匹配方法。该方法是将前面3种维度信息综合起来用于用户识别。文献[3]提取用户社交、空间、时间和文本信息,综合考虑用户的多种维度信息,采用婚姻匹配算法进行用户匹配。文献[18]从用户的社交数据出发,包括用户社交关系、发布的文本数据、行为轨迹等信息,设计了异构行为模型用于度量用户行为相似性,进而进行用户身份解析。文献[19-20]均利用了档案属性信息和拓扑结构信息,其中文献[19]提出一种基于条件随机域的用户档案匹配方法,文献[20]建立了多维度的特征向量,通过特征向量相似度判定是否为同一用户。文献[21]决策树与贝叶斯网络实现了多源数据的网络空间中实体的分类与识别。

多维度信息的身份识别效果优于仅使用单一维度信息的算法,其优越性在于该类算法结合了更全面的用户特征信息来表示用户身份,使得匹配效果更好。然而,由于社交平台的复杂性,不同网络间的数据结构往往是不对称的,无法同时获得多维度信息用于匹配同一用户。本文研究跨评论类与活动类网络间的用户识别方法,以为了解决两类社交平台的用户信息属性不对称问题。

2 问题定义

给定一个活动类社交平台G^{A}和一个评论类社交平台G^{R},将社交平台表示为无向图,活动类网络$G^{\mathrm{A}}=(V^{\mathrm{A}}, E^{\mathrm{A}})$包含不同类型的结点和关系,$V^{\mathrm{A}}$是活动类网络中的结点集合,包括4种类型的结点信息$V^{\mathrm{A}}=U^{\mathrm{A}}\cup G\cup I\cup Z$,其中$U^{\mathrm{A}}=\{u_1^{\mathrm{A}}, u_2^{\mathrm{A}},\cdots, u_n^{\mathrm{A}}\}$为一组用户账号集合,代表真实用户在活动类网络上的社交账号,$G=\{g_1, g_2,\cdots, g_G\}$为用户参加过的小组集合,$I=\{<c_1, x_1>,<c_2, x_2>,\cdots,<c_m, x_m>\}$为用户在活动类网络中标记的项目集合,如用户看过的书、电影等项目,且每个项目包含项目名称c_i和描述标签x_i。$Z=\{t_1, t_2,\cdots,t_z\}$表示为用户在网络中标记项目的时间段集合。$E^{\mathrm{A}}\subset V^{\mathrm{A}}\times V^{\mathrm{A}}$为异构社交平台$G^{\mathrm{A}}$中不同类型的边集,其中$E_u^{\mathrm{A}}\subset E^{\mathrm{A}}$为用户间的关注关系。

评论类社交平台$G^{\mathrm{R}}=(V^{\mathrm{R}}, E^{\mathrm{R}})$,其中结点集$V^{\mathrm{R}}=U^{\mathrm{R}}\cup C\cup Z$为3类结点信息,$U^{\mathrm{R}}$与$Z$分别为用户在评论类网络中的社交账号集合与在该网络中发表评论的时间段集合,与活动类网络相类似。$C=\{x_1, x_2,\cdots,x_c\}$表示为用户在评论类网络中发表的评论集,$x_i$为将用户发表的文本内容分词之后的第$i$个词集合。

问题1 跨活动类与评论类社交平台用户身份解析。给定两个社交平台G^{A}和G^{R},用户身份解析判断来自两个网络的用户u_i^{A}与u_j^{R}是否为同一用户,其中$u_i^{\mathrm{A}}\in G^{\mathrm{A}}$,$u_j^{\mathrm{R}}\in G^{\mathrm{R}}$。如果两个账号是同一用户则目标输出值为1,否则为0。

本文定义的用户身份解析是发现在两个异构社交平台间的一对一用户匹配关系,与现有工作如文

献[3]的区别是本文中的异构社交平台是指两个网络间结构不同,属性不对称的社交平台,针对社交平台关系不可用情况。

3 用户身份解析模型

本文提出了面向异构社交平台的用户身份解析模型。图1给出了模型结构,首先将不对称用户信息进行抽取和表示,按照用户档案属性、语义序列和特征词序列计算用户间的相似度,结合反向传播学习得到多个相似度权值,通过集成的用户相似度识别出同一用户。由于采用反向传播学习可以发现隐含的用户习惯,模型中计算用户间的语义序列和特征词序列相似度时均采用反向传播学习来确定时间偏移权重。同时,3类相似度重要性确定也采用反向传播学习实现。学习模型的输入层为用户一个序列及相似度,输出层为是否同一用户。

为了实现同一用户的有效匹配,用户信息按3种类别组织,第一类用户档案属性信息;第二类用户表达语义信息,按时间序列组织;第三类为特征词信息,按时间序列组织。对于档案属性信息,采用文本语义相似度度量用户的相似性,可表示为f_a。对于用户表达语义信息,设计语义序列相似度度量计算用户的相似性,表示为f_s,采用反向传播学习的方法确定相邻时间段内的相似度权重;对于特征词特征序列,采用特征词匹配方法进行度量,表示为f_k,采用反向传播学习的方法确定不同时间隔的匹配权重。两个用户的整体相似度计算表示为$F(f_a, f_s, f_k)$,3类信息相似度所占权重采用反向传播学习的方法获得。

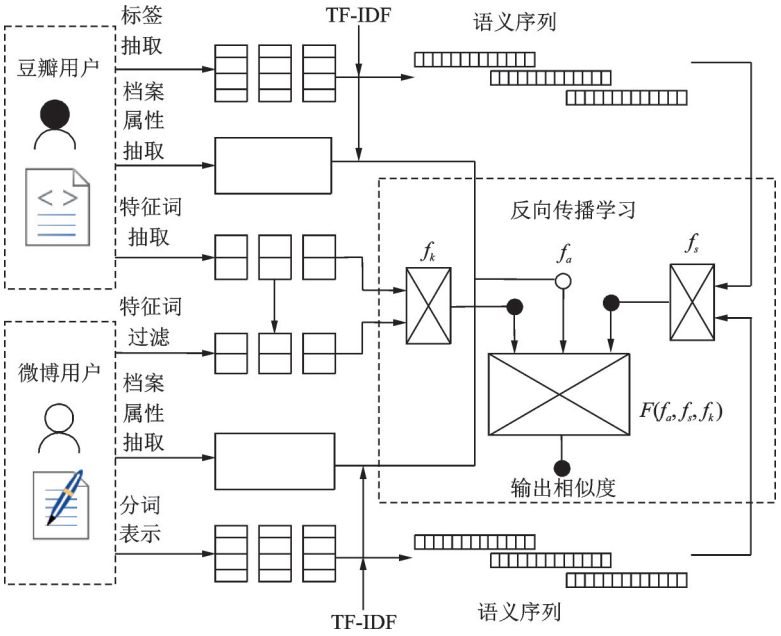


图1 异构社交平台中用户身份解析模型

Fig.1 User identity resolution model in heterogeneous social platforms

3.1 特征数据抽取

异构社交平台中属性信息不对称,需要将这些不对称的信息集成映射到可比较的形式。

(1) 档案属性抽取。活动类与评论类用户的账户信息分别表示为P_u^{A}与P_u^{R},包括用户账号、简介、位置、生日、加入小组的名称和类别等属性信息,表示为一段文本描述。将文本描述根据中文语句的句

法、语义信息，结合句子的上下文，对句子进行单字切分、组合等操作形成词集合。由于此类信息存在缺失值，且文本长度差异较大，如活动类小组名称较多，对于长文本描述先利用主题模型进行降维，生成长度差异不大的文本进行比较。

(2) 语义序列抽取。用户的语义序列表示为 $X_s=\{(x_1, t_1), (x_2, t_2), \cdots, (x_n, t_n)\}$，其中设 $X_{si}=(x_i, t_i)$ 表示时间段 t_i 的词集合为 x_i。活动类用户语义序列表示为 X_s^{A}，通过抽取用户参加的活动名称与标签、看过的电影和书名称与标签，进行分词，按时间段加入相应的词集合。评论类用户语义序列表示为 X_s^{R}，通过抽取用户所有评论文本，对文本进行分词表示，去停用词，并按时间段序列组织成词集。

(3) 特征词序列抽取。特征词序列可表示为 $X_k=\{(c_1, t_1), (c_2, t_2), \cdots, (c_n, t_n)\}$，其中 c_i 指在 t_i 时刻出现的关键字。活动类网络用户的特征词序列表示为 X_k^{A}，通过抽取用户参加过的活动名、看过的电影和书名称等，以整体名称作为一个词，并取时间戳标记词的时刻。评论类网络用户的特征词序列表示为 X_k^{R}。其获取方法为：首先将活动类网络所有用户的特征词集合合并为一个词集，除去重复词，得到该网络特征词集合；然后将此集合与评论类网络中各用户的评论文本进行字符串匹配，得到评论类网络用户的特征词集合，并取时间戳标记词的时刻。

将活动类的特征词与评论类中文本进行简单字符串匹配，并不能保证匹配出的词为特征词。例如，电影名称为“伞”，在评论类文本可能匹配了“雨伞”“送伞”“带伞”中的“伞”。再如，电影名称为“墙”，会与文本的“穿墙”“修墙”“山墙”等词匹配。这些匹配出的词并不是特征词。另外，活动类网络中的特征词会存在词的包含关系，例如，存在3个电影名称“春天”“过春天”“四个春天”，当利用“春天”去匹配微博中的词可能将其他两个电影名称匹配成“春天”。

匹配出的特征词存在以下特点：(1) 活动类中的词是生活中常见词；(2) 词比较短。设计了如下策略进行特征词匹配：

(1) 当不同特征词存在包含关系时，优先考虑长的特征词；

(2) 分析词的前后语义关系，所匹配上的词与上下文可组织实际语义，则不看作特征词。

(3) 如所匹配的词前后存在分隔字符，如《》、#、“”等确定为特征词；

(4) 如评论类网络中的词前后存在其他类词，如昵称、##、@等词情况，不确定为特征词。

3.2 相似性度量

本节介绍各类信息的相似度计算方法，包括语义序列相似度、特征词序列相似度、以及整体相似度的计算方法。

3类信息中除了特征词，其他两类均需要对关键词进行统一的表示，采用TF-IDF (Term frequency-inverse document frequency)进行相似性计算。对于用户档案属性信息，分别将来自两个网络的用户档案属性信息集合作为文档，一个用户形成一个文档，再进行TF-IDF表示。对语义序列根据时间段划分，每一个时间段作为一个文档，进行TF-IDF表示。

以用户基本属性信息为例说明TF-IDF表示方法，对于活动类网络用户 u 其文档为 P_u^{A} (评论类网络用户为 P_u^{R})，利用TF-IDF可表示成向量为 $\boldsymbol{W}=\{w_1, w_2, \cdots, w_D\}$，其中 w_q 表示文档 P_u^{A} 中第 q 个词的权重，D 是向量化词的数量。在TF-IDF模型中，w_q 为

$$w_q=f_q\times f_{qd} \tag{1}$$

式中 f_q 为第 q 个词在文档 P_u^{A} 中出现的频率，可表示为

$$f_q=\frac{N_q}{N} \tag{2}$$

式中：N_q 为第 q 个词在文档 P_u^{A} 中出现的次数，N 为文档 P_u^{A} 的总词数。

f_{qd} 为词在所有文档中的逆文档频率，表达式为

$$f_{qd}=\lg\frac{|P|}{f_d+1} \tag{3}$$

式中:P为活动类与评论类所有文档之和;f_d表示出现第q个单词的文档个数。

对用户的基本属性信息与语义序列信息用TF-IDF表示后,利用余弦相似度计算语义相似程度。

3.2.1 语义序列相似度

语义序列相似度反映了两个用户的行为的规律的相似度,为了解决用户在不同社交平台上的活动与评论语义上的偏移,本文提出一种基于反向传播的语义序列相似度度量方法。该方法的基本思想是将一个序列的每个时间段内的语义分别与另一个序列前后时间段进行比较,通过反向传播计算偏移时间段内的相似度权重,累计所有时间段与另一个序列的语义相似度作为两个序列的相似度。反向传播计算偏移时间段内的相似度权重方法如下。

给定两个语义序列X_s^{A}与X_s^{R},对于每个时间段t_i的语义x_i分别与X_s^{R}序列中时间t_i前后共p个时间段语义进行计算,两个序列的相似度为

$$f_s\left(X_s^{\mathrm{A}},X_s^{\mathrm{R}}\right)=\frac{1}{L}\sum_{i=1}^{L}\frac{1}{p}\sum_{j=i-p/2}^{i+p/2}w_jS_{ij} \tag{4}$$

式中:L为两个序列总长度;S_{ij}为序列X_s^{A}与X_s^{R}在时间段i与j上语义序列X_{si}^{A}、X_{sj}^{R}的余弦相似度。

本文通过最小化目标函数E_s来评估不同的相邻时间段对语义序列识别同一用户的重要性,如式(5)所示。其中,t为目标值,如果是同一用户其值为1,否则为0。

$$\min_{w_j}E_s=\frac{1}{2}(t-f_s)^2 \tag{5}$$

为了得到最优的权重值,计算每个权重目标函数的导数,有

$$\frac{\partial E_s}{\partial w_j}=-\left(t-f_s\right)\frac{\partial f_s}{\partial w_j}=-\left(t-w_jS_{ij}\right)S_{ij} \tag{6}$$

在每次迭代过程中,从权重中减去该权重目标函数的导数值,会使得权重变化较大,导致过度纠正。因此通过将每个导数乘以学习率η来更新权重系数,学习率η可以控制模型的学习进度,缓解参数更新时的过度纠正,本文将学习率η设置为0.5。第$(r+1)$次迭代为

$$w_j^{r+1}=w_j^r-\eta\frac{\partial E_s}{\partial w_j}\quad\forall j=1,\cdots,p \tag{7}$$

与不同用户的语义序列相比,来自同一用户的不同网络上的两个语义序列在整体上应该是相似的,两个语义序列大部分采样点也应该是相似的。因此,反向传播学习相邻时间段的权重时,将来自两个网络的同一用户两个语义序列中一部分采样点作为正例,本文选取平均相似度前50%的采样点作为正例,其目标输出值为1。将不同用户的语义序列中平均相似度后50%的采样点作为反例,其目标输出值为0。经过反向传播更新每个相似度的权值,使得输出值无限接近目标值。

3.2.2 特征词序列相似度

特征词序列的相似度取决于两个方面,一方面是两个序列匹配上词的个数,即匹配的词数量越多说明两个序列越相似;另一方面匹配上的词的重要程度,由匹配的时间间隔与该词出现的频率决定。直观上匹配上的词越相近,则为同一用户发布的可能性越大,同时,匹配上的词很小出现,则指向同一个用户可能性也越大。例如,一个用户看了一部很少人看的电影,并且在同一天在另一个网络上出现该电影的评论,则此事件为小概率事件,小概率事件发生了往往指向一个事实,即由同一个人在两个不同的网络发布的信息,即为同一用户。基于以上考虑,本文设计的特征词序列相似度计算方法如下。

给定两个特征词序列X_k^{A}与X_k^{R},对X_k^{A}序列中的每个特征词c与X_k^{R}序列比较,如果在序列X_k^{R}中存

在相同的词,且这两个序列的时间间隔在某一范围内,用m个时间段表示,则产生一个匹配记为$\langle c, T\rangle$,其中T为词c在两个序列中的时间间隔,存在$T<m*\Delta t$,Δt为单位时间间隔。

通过累记m个时间段内的匹配数量可作为两个序列的相似度,一次匹配$\langle c, T\rangle$的重要程序取决于c出现的频数N_c及时间间隔T的大小。c出现的频数为1时最重要,频数越大其重要程度越小,采用指数函数表示c的频数N_c与相似度值的关系,有

$$f_k(\langle c, T\rangle)=\frac{1}{e^{w_a(N_c-1)}} \tag{8}$$

式中w_a用于调整函数的曲率。当$N_c=1$时,函数取最大值1,N_c越大表明匹配相似度越小。

时间间隔T与相似度成反比,采用同样的指出函数表示T与相似度的关系,有

$$f_k(\langle c, T\rangle)=\frac{1}{e^{w_bT}} \tag{9}$$

式中w_b用于调整函数的曲率。当$T=0$时,即为同一时间段内的匹配上此匹配贡献的相似度最大,随着T的增加此匹配的相似度贡献值越小。

综合考虑以上两个因素,对于一次特征词匹配的相似度值,有

$$f_k(\langle c, T\rangle)=\frac{1}{e^{w_a(N_c-1)+w_bT}} \tag{10}$$

特征词的出现频数N_c和匹配时间段T对识别同一用户的重要性由参数w_a与w_b决定,本文采用反向传播学习来确定这两个参数。通过调整w_a与w_b,最小化目标函数E_k,如式(11)所示。其中,t为目标值,如果是同一用户其值为1,否则为0。

$$\min_{w_a,w_b} E_k=\frac{1}{2}(t-f_k)^2 \tag{11}$$

然后,计算每个权重的目标函数导数,对于参数w_a,有

$$\frac{\partial E_k}{\partial w_a}=-(t-f_k)\frac{\partial f_k}{\partial w_a}=\left(t-\frac{1}{e^{w_a(N_c-1)+w_bT}}\right)\frac{N_c-1}{e^{w_a(N_c-1)+w_bT}} \tag{12}$$

对于参数w_b,有

$$\frac{\partial E_k}{\partial w_b}=-(t-f_k)\frac{\partial f_k}{\partial w_b}=\left(t-\frac{1}{e^{w_a(N_c-1)+w_bT}}\right)\frac{T}{e^{w_a(N_c-1)+w_bT}} \tag{13}$$

在每次迭代过程中,通过使用学习率η更新权重系数,第$(r+1)$次迭代为

$$w_k^{r+1}=w_k^r-\eta\frac{\partial E_k}{\partial w_k}\qquad k=a,b \tag{14}$$

在反向传播学习w_a与w_b时,将来自两个网络的同一用户已匹配上的特征词作为正例,其目标输出值为1,将不同用户已匹配上的特征词作为反例,其目标输出值为0。经过反向传播更新每个相似度的权值,使得输出值无限接近目标值。在计算用户身份解析过程中,根据学习所得的权重累加不同网络用户间相同特征词匹配f_k的值,根据相似度排序确定是否为同一用户。

为了与其他相似度集成,将特征词序列相似度归一化,即统计特征词序列的相似度最大值,然后将每一对序列的相似度除以最大值。两个特征词序列的相似度计算为

$$f_k(X_k^A, X_k^R)=\sum_{i=1}^{n} f_{k_i}(\langle c, T\rangle) \tag{15}$$

式中n为已匹配上特征词的个数。

根据特征词匹配,得出两个特征词序列匹配相似度如式(16)所示,两个序列中所有特征词匹配累加和得到

$$F_k(X_k^{\mathrm{A}},X_k^{\mathrm{R}})=\frac{f_k(X_k^{\mathrm{A}},X_k^{\mathrm{R}})}{\max f_k(X_k^{\mathrm{A}},X_k^{\mathrm{R}})} \tag{16}$$

3.2.3 相似度集成

综合考虑用户的基本属性信息、语义序列以及特征词序列,设计一个整体相似度函数

$$F=w_1f_a+w_2f_k+w_3f_s=\sum_{l=1}^{3}w_lf_l \tag{17}$$

式中:f_a为基本属性相似度;f_k为特征词序列相似度;f_s为语义序列相似度。参数w_1、w_2、w_3用于决定3类信息的重要性,同样采用反向传播学习的方法来确定。通过最小化目标函数E_F来评估不同方法对识别同一用户的重要性,有

$$\min_{w_l}E_F=\frac{1}{2}(t-F)^2 \tag{18}$$

然后,利用式(19)目标函数的导数,有

$$\frac{\partial E_F}{\partial w_l}=-(t-F)\frac{\partial F}{\partial w_l}=-\left(t-w_lf_l\right)f_l \tag{19}$$

对于每次迭代,通过使用学习率η更新权重。第$(r+1)$次迭代为

$$w_l^{r+1}=w_l^r-\eta\frac{\partial E_F}{\partial w_l}\quad\forall l=1,\cdots,3 \tag{20}$$

在集成匹配同一用户的过程中,分别将来自不同网络中两个账户的属性信息、特征词序列信息以及特征词序列信所得到函数值作为输入,输出为目标值t,经过反向传播计算3种信息在识别同一用户过程中所占的权值,使得输出值无限接近目标值。其中,同一用户中的3类信息作为正例,即目标输出为1,不是同一用户三类信息为反例,目标输出为0。

4 实验结果与分析

本节利用两个真实的社交平台数据集对提出方法的有效性进行测试,算法使用Python语言实现。

4.1 数据准备

数据来自目前较为流行豆瓣和微博两个异构网络的真实数据集,豆瓣网是活动类社交平台,微博网是评论类社交平台。在两个网络平台各抓取了部分用户数据,这些用户已在豆瓣网标记了其微博账号或链接,这些用户数据作为同一用户事实,用于在实验中计算准确率。

对于豆瓣网中的用户数据,抓取了这些用户2018年1月1日到2019年1月1日一年的数据,数据中包含用户参加的6 531个小组与活动、观看的35 894个电影、书,抽取了这些小组及电影等的名称及标签。对于微博网中的用户,抓取了用户的基本属性信息,同时收集相应时间内用户发表的微博文本数据。另外,收集了与豆瓣网用户没有对应关系的部分微博用户数据,作为反例。数据集如表1所示。

对于微博用户发布的文本信息数据小于5条,豆瓣活动数据少于10条的用户进行删除,并对于具有少量缺失值的用户信息数据进行相应的数据缺失值填充。对于微博及豆瓣网络中的某些文本数据不能直接用于用户身份解析,对于个人简介信息不能直接用字符串进行匹配,需要进行向量维度转化,所用到的向量维度转化的方法正如前文所提到的TF-IDF和Word2vec等方法,向

表1 实验数据集

Table 1 Experimental dataset

社交平台	微博网	豆瓣网
用户数	3 425	4 749
Tip数	214 796	357 423
时间范围	2018—2019年	2018—2019年
真实用户数	2 861	

量表示之后在通过余弦相似度计算语句间的相似度值。表2给出了实验相关参数设置。

表2 实验参数设置

Table 2 Experimental parameter setting

变量	值	说明
w	5	时间窗口大小
d	300	向量维度
η	0.5	学习率
iter	1 000	迭代次数

4.2 比较方法

为了评估本文提出算法的有效性，本文采用身份解析中应用较为广泛的评价指标：Acc@k，对于测试集中的一个实例，如果识别的同一用户出现在Top-k建议的相似用户集中，则Acc@k为1，否则为0。本文分别选取k={1, 5, 10, 15, 20}来说明Acc@k的不同结果。实验中使用5折交叉验证。采用准确率、召回率及调和平均数作为评价指标，由Precision、Recall和F_1-score表示。

由于本文所解决的为不同类型的社交平台间的用户身份解析，假设用户名与社交关系不可用的情况。因此，将现有文献的利用语义匹配的方法作为比较方法如文献[3-4,12,16-17]。比较的方法归纳如下：

(1)整体语义匹配方法(WS)。将两类网络用户的所有信息转换为词集，采用比较整体语义相似度的方法进行匹配。语义信息包括：用户名、简介、位置和加入小组的名称及标签，参加活动的名称及标签，发表的评论名称与内容等。此方法为文献[3,12,16]提出的整体语义匹配方法。

(2)语义序列匹配方法(TS)。将用户参加活动的名称与标注、发表的评论组织成词序列，计算语义序列相似性进行用户匹配。此方法与文献[4,17]提出的主题演化类似，但本文应用反向学习学习了最优的演化偏移时间，同时考虑了特征词的序列匹配。

(3)特征词序列匹配方法(FS)。将用户参加活动名称组织成特征词序列，在评论网中过滤产生另一类特征词序列，计算两个序列的相似性进行用户匹配。

(4)集成方法(MUL)。利用属性信息，特征词信息，语义序列信息3类信息，结合语义序列匹配、特征词序列匹配方法进行用户识别。

4.3 实验结果

本节将给出不同识别算法的实验结果及各参数对算法的影响。

(1)不同识别算法的对比

为了对比本文所提算法的有效性，分别对4个算法做了实验，其中图2是豆瓣网作为源网络，微博网作为目标网络，图3是豆瓣作为目标网络，微博作为源网络的实验结果图。从图2和3可以看出，无论是将微博设置为源网络还是将豆瓣设置为源网络，集成方法(MUL)效果最好。实验时各

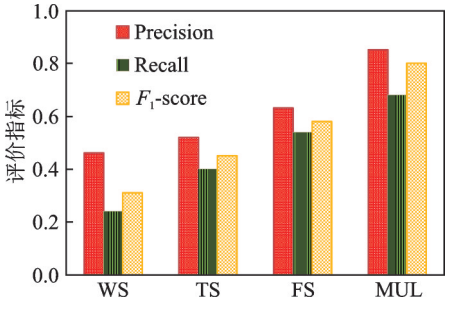


图2 豆瓣作为源网络的识别结果

Fig.2 Identification results when Douban is a source network

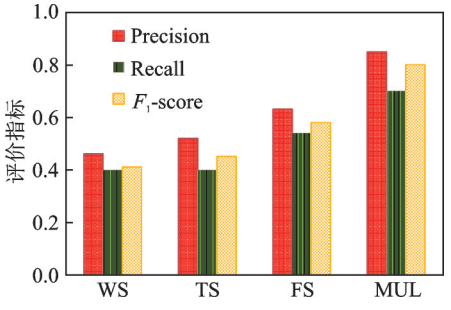


图3 微博为源网络的识别结果

Fig.3 Identification results when Weibo is a source network

个模型参数均采用默认参数,集成方法在Precision、Recall和F_1-score上都有了明显的提高。通过TS算法和WS算法可以看出:用户在参加活动时的语义序列特征可以有效提升用户识别的准确率。

(2)权重学习效果对比

图4显示了在集成算法中3种方法的权重分配方式对识别率的影响,分别对比了两种权重分配的方法:一是采用反向传播学习到每种方法在用户身份解析中所占的权重值w_l的方法(MUL),二是采用自然平均分配权重的方法(MUL_no)。显然,根据反向传播学习到的权重值在用户身份解析中的效果更好,反向传播方法可以学习到不同属性信息对于用户识别的重要性,从而提高了识别率。

(3)语义序列中时间间隔对用户识别的影响

图5显示了在语义序列中,两个序列匹配时的前后匹配不同间隔的时间段对算法性能的影响,可以看出评估指标在一定的时间间隔下会取到最优值,整体上呈先下降后增大趋势。当时间间隔T为2时存在一个异常,与真实数据集的分布有关。通过语义序列匹配同一用户的方法可以有效识别出两个用户在一段时间内行为规律的相似度,解决了用户在一段时间内不同社交平台上的活动与评论在语义上的偏移,同时也避免短期兴趣的干扰。

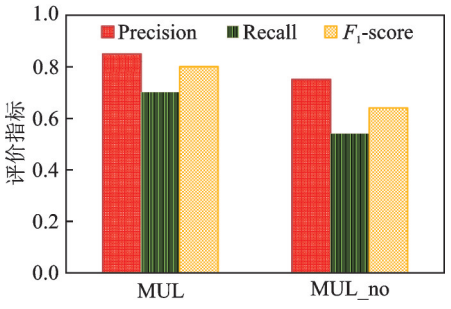


图4 权值分配对识别结果的影响

Fig.4 Effect of weight distribution on identification rate

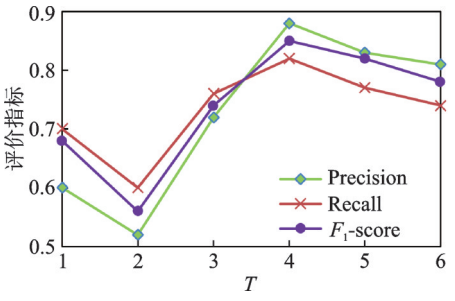


图5 不同时间间隔对识别结果的影响

Fig.5 Effect of different time intervals on results

(4)特征词序列中时间间隔对应的权重值

图6显示了在特征词匹配过程中,两个序列中相同特征词出现的时间间隔对于用户身份解析的影响。观察发现,对于不同的时间间隔通过反向传播学习到的权重值有所不同。在活动类网络上的特征词匹配评论类网络中同一特征词时,不同的时间间隔的权重值,随着间隔的天数增加,权重值先上升后下降。这表明用户在网络中标记特征词时,不同的社交平台间通常不会在同一天标记,用户在豆瓣网络中标记该特征词,会在一段时间之后在微博网中发表关于此特征词相关的评论信息,因此权重值的大小可以反映出在特征词匹配时的时间间隔规律,有利于提高用户识别的准确率。

(5)语义序列中时间间隔权重值

图7展示了在语义序列中通过反向传播学习的方法所得到的对应时间间隔的权重值,观察发现,其值在一个时间间隔会取到最大值。两个序列间的用户动态语义信息与在一定的时间段内具有相似性,超过某个时间点其相似性降低,甚至不相似,这就使得对应的时间段内所分配的权重值降低,可见权重值的大小在一定程度上反映出语义序列匹配时的语义偏移规律。

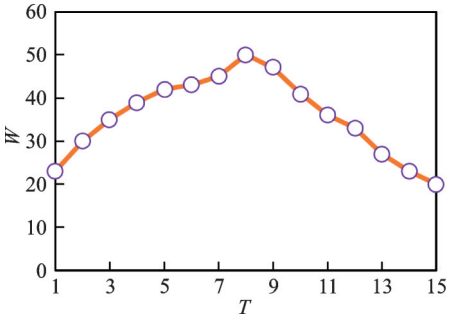


图6 特征词序列中不同时间间隔对权重影响

Fig.6 Effect of different time intervals on weights in feature word sequences

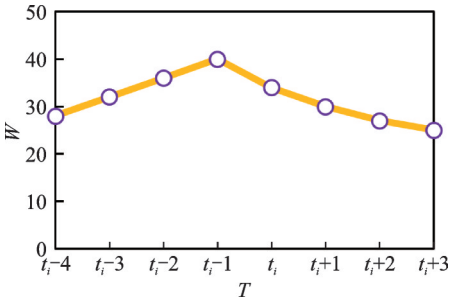


图7 语义序列中不同时间间隔对权重影响

Fig.7 Effect of different time intervals on weights in semantic sequences

(6)特征词序列中词频和时间间隔影响

图8显示了在特征词匹配中,不同词的词频c在用户识别过程中的影响。词频比较低时,其识别的准确率更高,随着词频数量的增大,准确率呈现缓慢下降状态。这也证实了特征词匹配的特性,词频越小的词即为小概率事件,小概率事件发生了往往指向一个事实,即由同一个人在两个不同的网络发布的信息,即为同一用户。

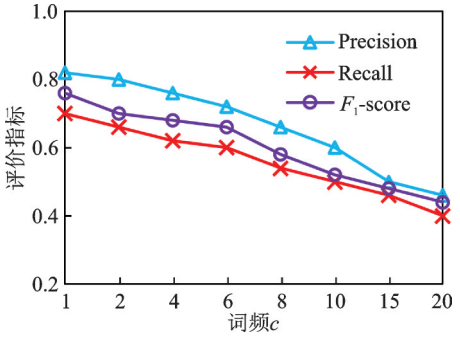


图8 词频对识别结果的影响

Fig.8 Effect of word frequency on results

图9显示了在特定词序列匹配过程中,相同词在两个序列中出现时所时间间隔T对算法识别同一用户的影响。可以看出特征词匹配随着时间间隔的增大,识别率先逐渐增大后趋于下降,在两个特征词匹配上的间隔时间间隔越短则其识别准确率并不是越高,而是在一定的时间间隔内,其识别的准确率较高。可见同一用户在两个网络中发布相同特定词时往往不会间隔太长的时间,给定特定词时间间隔的限定,避免冗余信息干扰,可以显著提高用户识别的准确率。

(7)主题个数设置对结果的影响

在主题模型中,本文分别选择了LDA和LSI两个目前较为流行的主题模型,实验发现设置不同的主题个数所获得的准确率不同。图10显示随着主题模型设置个数的增多,准确率逐渐增长,当达到某

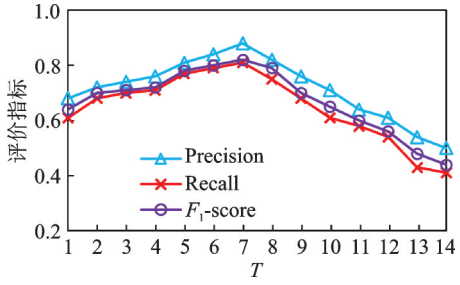


图9 时间间隔T对识别结果的影响

Fig.9 Effect of time period T on results

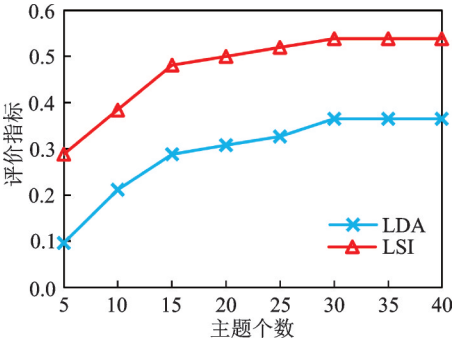


图10 主题模型个数设置对准确率影响

Fig.10 Effect of the number of theme models on accuracy

一值时,随着主题个数的增加则准确率不再增长,而是趋于平缓状态。因此,通过实验可以得到最佳的主题个数设置值。

5 结束语

本文提出了结合语义序列与特征词序列匹配识别同一用户的方法,分别从整体语义、时间序列语义偏移以及特征词匹配等多个角度识分析用户,采用反向传播学习方法获取相应的权重值,提高用户识别的准确率。利用真实异构社交平台数据集进行实验,结果表明该方法可以有效地解决在异构社交平台中的用户身份解析问题。

参考文献:

[1] ZHANG Yutao, TANG Jie, YANG Zhilin, et al. Cosnet: Connecting heterogeneous social networks with local and global consistency[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2015: 1485-1494.

[2] ZHOU Xiaoping, LIANG Xun, ZHANG Haiyan, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(2): 411-424.

[3] KONG Xiangnan, ZHANG Jiawei, PHILIP S Y. Inferring anchor links across multiple heterogeneous social networks[C]//Proceedings of the 22nd ACM Conference of Information and Knowledge Management. New York: ACM Press, 2013: 179-188.

[4] NIE Yuanping, JIA Yan, LI Shudong, et al. Identifying users across social networks based on dynamic core interests[J]. Neurocomputing, 2016, 210: 107-115.

[5] MU Xin, ZHU Feida, EEPENG L, et al. User identity linkage by latent user space modelling[C]//Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1775-1784.

[6] NARAYANAN A, SHMATIKOV V. De-anonymizing social networks[C]//Proceedings of the 30th IEEE Symposium on Security and Privacy. [S.l.]: IEEE Computer Society, 2009: 173-187.

[7] 汪潜, 申德荣, 冯朔, 等. 全视角特征结合众包的跨社交网络用户识别[J]. 软件学报, 2018, 29(3): 811-823.
WANG Qian, SHEN Derong, FENG Shuo, et al. Identifying users across social networks based on global view features with crowdsourcing[J]. Journal of Software, 2018, 29(3): 811-823.

[8] CHEN Hongxu, YIN Hongzhi, SUN Xiangguo, et al. Multi-level graph convolutional networks for cross-platform anchor link prediction[C]//Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2020: 1503-1511.

[9] PERITO D, CASTELLUCCIA C, KAAFAR M, et al. How unique and traceable are usernames?[C]//Proceedings of the 11th International Symposium, Privacy Enhancing Technologies. Waterloo: [s.n.], 2011: 1-17.

[10] 刘东, 吴泉源, 韩伟红, 等. 基于用户名特征的用户身份同一性判定方法[J]. 计算机学报, 2015, 38(10): 2028-2040.
LIU Dong, WU Quanyuan, HAN Weihong, et al. User identification across multiple websites based on username features[J]. Chinese Journal of Computers, 2015, 38(10): 2028-2040.

[11] LI Yongjun, PENG You, ZHANG Zhen, et al. Matching user accounts across social networks based on username and display name[J]. World Wide Web, 2019, 22(3): 1075-1097.

[12] MOTOYAMA M A, VARGHESE G . I seek you: Searching and matching individuals in social networks[C]//Proceedings of the 11th International Workshop on Web Information & Data Management. New York: ACM Press, 2009: 67-75.

[13] ZAMANI K, PALIOURAS G, VOGIATZIS D. Similarity-based user identification across social networks[M]//Similarity-Based Pattern Recognition. [S.l.]: Springer International Publishing, 2015.

[14] ESFANDYARI A, ZIGNANI M, GAITO S, et al. User identification across online social networks in practice: Pitfalls and solutions[J]. Journal of Information Science, 2016, 44(1): 58-69.

[15] NAINI F M, UNNIKRISHNAN J, THIRAN P, et al. Where you are is who you are: User identification by matching statistics [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(2): 358-372.

[16] ZHAO Dongsheng, ZHENG Ning, XU Ming, et al. An improved user identification method across social networks via tagging behaviors[C]//Proceedings of the 30th International Conference on Tools with Artificial Intelligence (ICTAI). Piscataway, NJ: IEEE, 2018.

[17] HAN Xiaohui, WANG Lianhai, XU Shujiang, et al. Linking social network accounts by modeling user spatiotemporal habits [C]//Proceedings of the 15th IEEE International Conference on Intelligence and Security Informatics. Piscataway, NJ: IEEE, 2017: 19-24.

[18] LIU Siyuan, WANG Shuhui, ZHU Feida, et al. HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling[C]//Proceedings of the 33rd ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2014:51-62.

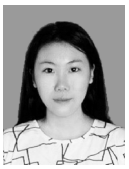
[19] BARTUNOV S, KORSHUNOV A, PARK S T, et al. Joint link-attribute user identity resolution in online social networks [C]//Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. [S.l.]: ACM, 2012.

[20] PELED O, FIRE M, ROKACH L, et al. Entity matching in online social networks[C]//Proceedings of 2013 International Conference on Social Computing. Washington DC: [s.n.], 2013: 339-344.

[21] 马旸,仲思超,蔡冰,等. 一种基于多源数据的网络实体推断方法[J]. 南京航空航天大学学报,2019,51(6): 870-878.
MA Yang, ZHONG Sichao, CAI Bing, et al. A network entity inference method based on multi-source data[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2019, 51(6): 870-878.

作者简介:

刘俊岭(1972-),女,博士,副教授,研究方向:数据挖掘和时空数据查询,E-mail: liujl@sjzu.edu.cn。



刘颖(1998-),女,硕士研究生,研究方向:数据挖掘。

马晨旭(1996-),男,硕士研究生,研究方向:数据挖掘。

赵巧娜(1994-),女,硕士研究生,研究方向:数据挖掘。

孙焕良(1969-),**通信作者**,男,博士,教授,博士生导师,研究方向:时空数据管理、数据挖掘、机器学习等,E-mail:sunhl@sjzu.edu.cn。

许景科(1976-),男,博士,教授,研究方向:空间数据管理、数据挖掘。

(编辑:夏道家)