一种基于加权非负矩阵分解的多维用户人格特质识别算法

王萌萌 左万利 王 英 王 鑫
(吉林大学计算机科学与技术学院 长春 130012)
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 随着社会媒体的普及,用户信息的爆炸式增长为深入理解在线用户行为提供了非常丰富的信息源.由于用户人格特质是用户行为的主要驱动力,人格特质的差异可能会对用户的在线行为产生一定的影响,因此,用户人格特质识别问题近年来受到了众多学者的关注.首先,基于用户网络结构信息和用户发布内容信息序列构建用户人格特质识别特征,并根据特征重要性为其分配权重.然后,以用户人格特质相关因子约束目标函数,从用户社会网络结构特征、语言学特征和情感特征三个维度利用非负矩阵分解方法识别社会网络中用户的五大人格特质.最后,在真实的数据集上验证了提出框架的有效性,并通过实验以更细的粒度进一步验证了用户人格特质之间相关性的存在,同时证明了特征权重和用户人格特质间的相关性在用户人格特质识别问题中的重要性.文中为社会网络中的多维用户人格特质识别问题提供了一种新思路.

关键词 多维用户人格特质识别;非负矩阵分解;用户人格特质相关因子;五大人格特质;社交网络
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2016.02562

A Multidimensional Personality Traits Recognition Model Based on Weighted Nonnegative Matrix Factorization

WANG Meng-Meng ZUO Wan-Li WANG Ying WANG Xin
(*Department of Computer Science and Technology, Jilin University, Changchun* 130012)
(*Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun* 130012)

Abstract With the pervasiveness of social media, the explosion of users' generated data provides a potentially very rich source of information for online researchers understanding user's behaviors deeply. Since user's personality traits are the driving force of user's behaviors and individual differences in user's personality traits may have an impact on user's online activities, as a consequence, user's personality traits recognition has attracted increasing attention in recent years. On the basis of user's network structure information and series of posts information, we first build user's personality traits recognition features, followed by distributing weights to features according to their different importance. And then, we utilize nonnegative matrix factorization to recognize user's Big Five personality traits from his/her network structure features dimension, linguistics features dimension and emotion features dimension by employing personality traits correlation factor to constrain objective function. Experiments on real-world Facebook dataset demonstrate the effectiveness of the proposed framework. Further experiments are conducted not

收稿日期:2015-10-10;在线出版日期:2016-03-02.本课题得到国家自然科学基金(61300148,61602057)、吉林省科技发展计划(20130206051GX)、吉林省科技计划(20130522112JH)、中国博士后基金(2012M510879)、吉林大学基本科研业务费科学前沿与交叉项目(201103129)资助.
王萌萌,女,1987 年生,博士研究生,主要研究方向为数据库、社会网络分析、数据挖掘、机器学习. E-mail: wmmwwlh@126.com. **左万利**,男,1957 年生,博士,教授,主要研究领域为数据库、社会网络分析、数据挖掘、机器学习. **王 英**(通信作者),女,1981 年生,博士,副教授,主要研究方向为数据库、社会网络分析、数据挖掘、机器学习. E-mail: wangying2010@jlu.edu.cn. **王 鑫**,男,1981 年生,博士,讲师,主要研究方向为数据库、社会网络分析、数据挖掘、机器学习.

only to validate the existence of the correlations between user's personality traits from a more fine-grained view, but also understand the importance of different feature's weight and the importance of the correlations between user's personality traits in recognizing user's personality traits. What's more, we provide a new train of thought for multidimensional personality traits recognition in social networks.

Keywords　multidimensional personality traits recognition; nonnegative matrix factorization; personality traits correlation factor; Big Five personality traits; social networks

1　引　言

作为一种新型的信息传播媒介，社会网络已经成为一种被人们广泛认可并使用的社交方式[1-3]. 尽管网络中的一些用户为了达到自我展示的目的，发布一些关于自己的非"真实"的照片和生活状态，社会网络中海量的用户内容仍然为用户行为的相关研究提供了宝贵的资源[4]. 心理学家认为，用户人格特质是用户行为的主要驱动力，人格特质的差异可能会对用户的在线行为产生一定的影响[5-6]. 因此，研究用户人格特质能够帮助人们更好地理解社会网络中的用户行为. 例如，可以利用用户人格特质预测用户对 Facebook 的接受情况[7]：具有责任型人格特质的用户在使用 Facebook 的过程中会有所保留；具有外向型人格特质的用户会经常使用 Facebook 并在其中结交很多朋友；具有神经质性人格特质的用户会进行高频率的交互活动. 此外，用户人格特质还有助于优化搜索结果[8]、分析社会影响力[9]、对人群中拥有共同特质的个体进行聚类[10]及预测客户的满意度和忠诚度[11]. 总之，用户人格特质识别在挖掘用户行为模式和获取用户潜在需求方面具有重要的理论意义及广阔的应用前景.

然而，Lee 等人[12]指出能够反映用户人格特质的特征逐渐趋于复杂化，因此就其本质而言，用户人格特质是难以识别的. 一般地，心理学家认为用户人格特质是一种长期的用户在思想上、情感上和行为上的表现出来的独特模式[13-14]，其主要反映在用户对待事物采取的态度和行为上[15]，故通过挖掘海量的用户发布状态中蕴含的用户语言学与情感特点以识别用户人格特质是一种可行的方法. 然而，通过深入分析可以发现，现有方法中大多没有考虑用户人格特质间的相关性对用户人格特质识别结果的影响，因此，本文提出了一个基于加权非负矩阵分解的用户人格特质识别模型（Weighted Nonnegative Matrix Factorization Model for Multidimensional Personality Traits Recognition，WNMF-MPTR），主要贡献如下：

（1）首次根据用户人格特质的不同级别对用户进行分组处理，并采用两种相关性度量方法验证了不同级别的用户人格特质之间弱相关性的存在.

（2）首次利用用户人格特质相关因子约束目标函数，将识别问题转化为求解社会网络特征、语言学特征和情感统计特征三个维度的加权非负矩阵分解最优解问题，有效地降低了时间复杂性并且使其能够对用户的多维人格特质进行准确识别.

本文第 2 节介绍相关工作和当前研究现状；第 3 节对用户人格特质识别特征进行了分析；第 4 节详细地阐述了提出的用户人格特质识别模型；第 5 节利用真实社会网络数据集验证提出方法的有效性；第 6 节给出结论与下一步工作.

2　相关工作

由于针对社会网络中用户行为的研究已经成为一个研究热点，因此，用户人格特质识别问题在理论和实践上均得到了广泛的关注.

现有用户人格特质识别算法大致可以分为两类：一类是基于用户语言学特征的方法；另一类是将用户语言学特征与用户社会网络特征相融合的综合方法.

第一批致力于此研究领域的是 Argamon 等人① 和 Mairesse 等人[16]. 基于文献[17]中提出的 essay 语料，Argamon 等人利用 SMO[18]对外向型人格特质和神经质型人格特质进行识别，并获得 57%～60% 的准确率. Mairesse 等人[16]将 LIWC 和 MRC 两个词典资源作为特征，分别通过 SVM[18]和 M5 模型[19]对用户人格特质分数和类别进行识别，实验结

① http://eprints.pascal-network.org/archive/00001492/01/argamon-etal-csna.pdf

果显示,利用提出的方法,用户人格特质的识别精确度范围在54%和62%之间. Oberlander和Nowson[20]将n元语法作为特征,分别利用SMO和朴素贝叶斯学习算法[18]识别用户的五大人格特质,并取得了较高的精确度(83%～93%,在责任型人格特质的识别中取得了最佳的精确度,但并未对开放型人格特质进行识别),同时通过实验指出特征选择过程对于提升算法性能的重要性. 然而,若将上述训练好的分类器应用于较大规模的数据集时,会发生过拟合现象,从而导致精确度下降到55%. Nguyen等人[21]首先抽取心理学特征和用户发布文本的情感倾向特征,然后利用SVM分类器识别社会链接,从而预测用户的影响力以及人格特质. 但该方法仅能判断用户有无影响力、用户人格特质的内向和外向,缺少对用户人格特质更细粒度的识别.

上述基于用户语言学特征的方法的精确度普遍不高,主要是由于一个数据集中不可能包含用户发布的所有信息,故其收集到的用户发布文本信息是有限的,其所反映的用户语言学特征也较为片面,因而准确识别用户人格特质存在一定的困难. 而用户的网络活动(如用户间建立链接)亦受用户人格特质的影响. 因此,一些学者将社会网络特征引入用户人格特质识别算法中,以提高算法的准确度.

Golbeck等人[22]基于结构化特征(链接)和语义特征,利用M5和Gaussian模型[23]计算特征集合与用户的五大人格特质之间的关联程度,从而识别279名Facebook用户的人格特质. 基于用户的朋友数和该用户最近发布的一条状态,Bai等人[24]使用C4.5算法[18]对人人网335名用户的人格特质进行识别,实验结果表明,通过融合用户网络结构特征和从用户发布状态中抽取的语言学特征,该方法的准确率能够达到69%～72%. Bai等人[25]基于从新浪微博中抽取的29种用户行为特征,分别提出一种多任务回归算法和一种增量回归算法以识别用户的五大人格特质,实验结果表明,通过用户在线微博的使用情况能够较为准确地对用户人格特质进行识别. Sun等人[26]通过实验表明,无需任何显著的增加或修改,认知框架可以作为一个通用的用户人格特质识别模型,同时还验证了结合用户人格特质模型与通用计算认知模型的可行性和有效性.

此外,2013年在ICWSM(The International AAAI Conference on Weblogs and Social Media)会议人格特质识别的专题研讨会中,大会基于Facebook用户人格特质标准数据集①中的文本和网络结构对不同的人格特质识别算法进行了系统地比较. Verhoeven等人[27]提出了一种元学习方法以识别用户的五大人格特质,实验结果表明其可以扩展为其他系统中的某些组件分类器,甚至是除英语外其他语言的分类器. Farnadi等人[28]分别利用SVM、k近邻[18]和朴素贝叶斯方法从用户的发布状态中自动识别用户人格特质,并通过实验证明,即使对于小规模的训练数据集,提出方法的性能仍高于大多数基线方法. 基于从Facebook中抽取的特征集合,Alam等人[29]对SVM、贝叶斯逻辑回归(Bayesian Logistic Regression, BLR)[30]和多项式贝叶斯(Multinomial Naïve Bayes, mNB)[31]三类分类算法的性能进行了比较,实验结果表明在Facebook用户人格特质识别问题中,mNB的性能要优于SVM和贝叶斯逻辑回归方法. Tomlinson等人[32]采用排序算法进行特征选择,并将逻辑回归模型(Logistic Regression, LR)[33]作为学习算法对Facebook用户人格特质进行识别,取得了较好的实验结果.

然而,现有用户人格特质识别工作仍然存在一些不足:(1)多数研究假设用户人格特质之间不存在或者存在较小的相关性[34],且在算法中并不考虑该因素对用户人格特质识别结果的影响,然而,正如"弱链接"在链接预测中发挥着重要的作用一样,人格特质间的这种"弱相关"亦在用户人格特质识别中扮演着不容忽视的角色[35];(2)尽管不同特征在用户人格特质识别中发挥着不同的作用[20,36-37],但仅有一小部分学者在识别用户人格特质时将特征与用户人格特质之间的相关性考虑在内[28]. 此外,目前并没有用户人格特质识别算法在量化不同特征重要性的同时又对用户人格特质之间的相关性加以考虑. 综上,如何在算法中刻画用户人格特质间的弱相关性以及如何构建用户人格特质识别模型以合理地融合多维特征都是非常具有挑战性的工作. 针对上述问题,本文提出了一个基于加权非负矩阵分解的识别模型以提高社会网络中多维用户人格特质识别的精度.

3 用户人格特质相关特征建模

3.1 用户人格特质识别特征定义

分析能够反映用户人格特质的特征因素是构建

① http://mypersonality.org/wiki/lib/exe/fetch.php?media=wiki:mypersonality_final.zip

识别模型的基础.因此,拟通过分析用户发布的状态对用户的语言学特征和情感统计特征进行抽取,并将其与 Facebook 用户人格特质标准数据集中直接提供的社会网络特征[38]共同作为用户人格特质识别特征.

3.1.1　社会网络特征

在 Facebook 用户人格特质标准数据集中,主要通过用户的拓扑结构反映用户的行为模式,包括 8 种社会网络特征,即用户的注册时间、网络规模、介数中心性、归一化的介数中心性、密度、中介性、归一化的中介性和传递性,之所以包含归一化的介数中心性和中介性是由于介数中心性、中介性与用户的网络规模相关,不同的用户其网络规模可能不同,因此,需要将其进行归一化才能够对不同用户的介数中心性、中介性进行比较.某用户 i 的网络 $Network_i$ 是由用户 i 的朋友以及其间的链接共同组成的,基于此,以用户 i 为例,其注册时间 $Create\_time_i$、网络规模 $Network\_size_i$、介数中心性 $Betweenness_i$、归一化的介数中心性 $NBetweenness_i$、密度 $Density_i$、中介性 $Brokerage_i$、归一化的中介性 $NBrokerage_i$ 和传递性 $Transitivity_i$ 的定义如表 1 所示.由于个体之间是相互联系并相互影响的[39],因此,与文献[40]中提出的假设类似,本文假设用户间的网络结构越相似,其越有可能具有相同类型的人格特质.

表 1　用户 i 的社会网络特征定义

社会网络特征	定义
$Create\_time_i$	表示用户 i 的账户创建时间
$Network\_size_i$	表示 $Network_i$ 中包括用户 i 在内的用户总数
$Betweenness_i$	表示 $Network_i$ 中经过用户 i 的最短路径总数,其计算公式如下:$Betweenness_i=\sum_{j,h\in Network_i} n_{jh}(i)/n_{jh}$,其中,$n_{jh}$ 表示用户 j 和用户 h 之间的最短路径总数;$n_{jh}(i)$ 表示用户 j 和用户 h 之间的最短路径经过用户 i 的总数
$NBetweenness_i$	$NBetweenness_i=2\times Betweenness_i/(Network\_size_i-1)((Network\_size_i-1)-1)$
$Density_i$	表示 $Network_i$ 中的用户间实际存在的链接总数与可能存在的最大链接数量的比值,其计算公式如下:$Density_i=Edge_i/Network\_size_i(Network\_size_i-1)$,其中,$Edge_i$ 表示 $Network_i$ 中的用户间实际存在的链接总数
$Brokerage_i$	表示 $Network_i$ 中非直接相连的结点对总数,其计算公式如下:$Brokerage_i=(Network\_size_i-1)((Network\_size_i-1)-1)-(Edge_i-(Network\_size_i-1))$
$NBrokerage_i$	$NBrokerage_i=2\times Brokerage_i/(Network\_size_i-1)((Network\_size_i-1)-1)$
$Transitivity_i$	表示 $Network_i$ 中拥有一个共同邻居的两个用户直接相连的平均概率,其计算公式如下:$Transitivity_i=\sum_{j\in Network_i} 2t_j \Big/ \sum_{j\in Network_i} k_j(k_j-1)$,其中 k_j 表示用户 j 的度数;t_j 表示围绕用户 j 的三角形数量,$t_j=\sum_{h,l\in Network_i} a_{jh}a_{jl}a_{hl}$,$a_{jh}$ 表示网络邻接矩阵的元素,如果用户 j 与用户 h 之间存在链接,则 $a_{jh}=1$;否则,$a_{jh}=0$

3.1.2　语言学特征

由于不同用户具有不同的表达方式,故一些学者认为用户的语言学特征与用户人格特质之间具有显著的联系[16-17],因此,语言学特征可以作为用户人格特质识别的一种新视角[41].

自然语言解析器主要用于解决句子的语法结构,如将词组合为"短语"并获取动词的主语或宾语.基于概率的解析器试图利用从现有的句法分析得到的语言知识中,获取新句子最准确的分析结果.Stanford Parser① 是斯坦福大学自然语言研究小组推出的一款基于概率的开源的自然语言语法解析工具[42].基于单独的 PCFG 短语结构和领域词汇,Stanford Parser 使用 A* 算法实现对自然语言的解析.此外,Stanford Parser 还提供了 GUI 界面,使用户可以将其简单地作为一个精确的、非词汇化的、随机的、上下文无关的语法解析工具,浏览其输出的短语结构树.

本文采用 Stanford Parser 从用户发布文本中抽取以下 35 种词性的词语使用频率作为语言学特征:连词、数词、限定词、方位词、外来词、情态助动词、物质名词、复数名词、专有名词、复数专有名词、前位限定词、所有格标记、人称代词、复数人称代词、基本形式的副词、比较副词、最高级副词、小品词、断句符、感叹词、基本形式的动词、过去时态的动词、动名词、过去分词形式的动词、现在时态的动词(非第三人称单数)、现在时态的动词(第三人称单数)、从属词、基本形式的形容词、比较形容词、最高级形容词、列表项标记、WH-限定词、WH-代词、WH-复数代词、WH-副词.在此基础上,本文增加了 5 个语言学特征:词语总数、逗号使用频率、句号使用频率、感

① http://nlp.stanford.edu/software/lex-parser.shtml#About

叹号使用频率和问号使用频率.由于用户发布文本中的超链接可能指向与用户人格特质识别无关的广告页[43],因此,在抽取语言学特征时并不考虑用户发布文本中包含的超链接.

3.1.3 情感统计特征

不同人格特质的用户在对待事物的态度上会反映出不同的情感表达方式.例如,一个具有神经质型人格特质的用户很容易产生不愉快的情绪,如愤怒、焦虑、抑郁和脆弱等.由此可见,用户的情感统计特征可以用于识别用户人格特质.因此,拟利用知网中英文情感分析用词语集①(其中包括 8945 个词汇和短语)中的正面、负面情感词列表和正面、负面观点词列表,基于上一节从用户发布状态中抽取的基本形式的形容词、比较级形容词和最高级形容词构建用户情感统计特征,主要包括用户发表状态中所含的正面情感词比例和负面情感词比例.用户 i 发表状态中的正面情感词比例 $PT(i)$ 和负面情感词比例 $NT(i)$ 计算如下:

$$PT(i)=pn_i/sum_i \tag{1}$$

$$NT(i)=nn_i/sum_i \tag{2}$$

其中,pn_i 和 nn_i 分别表示用户 i 发表的状态中包含在知网中英文情感分析用词语集中的正向情感词个数和负向情感词个数;sum_i 表示用户 i 发表状态中包含的词汇总数.

3.2 用户人格特质识别特征权重分配

文献[44]指出对于与待识别结果间存在强关联的特征,对其进行约束能够提高算法的性能.由于在用户人格特质识别问题中不同特征具有不同的重要程度,故为用户人格特质识别特征合理地分配权重就显得尤为重要.文献[44]和文献[28]以皮尔森相关系数对各特征与五大人格特质间的相关性进行度量,但皮尔森相关系数存在两处局限,即必须假设数据是成对地从正态分布中取得的且数据至少在逻辑范围内是等距的,而肯德尔检验对原始变量的分布不作要求,是一种通过计算相关系数测试两个随机变量的统计依赖性的非参数假设检验,适用范围要广些.因此,拟利用肯德尔相关系数法度量各特征与五大人格特质间的相关性,从而计算各特征权重.随机变量 X 和 Y 之间的肯德尔相关系数的计算如式(3)所示:

$$\tau(X,Y)=\frac{C-D}{\sqrt{(N_3-N_1)\times(N_3-N_2)}} \tag{3}$$

其中,C 表示两个随机变量中拥有一致性的元素对数;D 表示两个随机变量中拥有不一致性的元素对数;N_1 表示随机变量 X 中重复元素的总数,其计算如式(4)所示:

$$N_1=\sum_{i=1}^{s}\frac{1}{2}U_i(U_i-1) \tag{4}$$

其中,s 表示变量中重复元素的组数;U_i 表示第 i 个元素拥有相同元素的数量;N_2 表示随机变量 Y 中重复元素的总数,其计算方法与 N_1 相同,N_3 表示合并序列的总数,其计算如式(5)所示:

$$N_3=\frac{1}{2}N(N-1) \tag{5}$$

其中,N 表示随机变量的维数.随机变量 X 和 Y 之间的肯德尔相关系数 $\tau(X,Y)\in[-1,1]$,当 $\tau(X,Y)$ 为 1 时,表示两个随机变量拥有一致的等级相关性;当 $\tau(X,Y)$ 为 -1 时,表示两个随机变量拥有完全相反的等级相关性;当 $\tau(X,Y)$ 为 0 时,表示两个随机变量是相互独立的.即两个随机变量间的肯德尔系数越高,则二者越相关,反之亦然.本文假设如果用户人格特质识别特征与识别结果较为相关,那么其在用户人格特质识别中具有较高的重要性.此外,由于用户人格特质识别中的特征值和用户人格特质分数均为数值类型的变量,因此,拟通过计算特征与用户人格特质之间的肯德尔相关系数以量化每一个特征的重要性,则第 i 个特征 f_i 的重要性其计算公式如下:

$$I(f_i)=\frac{\sum_{p_j\in P}\tau(f_i,p_j)}{5} \tag{6}$$

其中,$\tau(f_i,p_j)$ 表示 f_i 与第 j 维用户人格特质 p_j 之间的肯德尔相关系数;P 表示用户的五大人格特质集合,将在下一节对其进行详细定义.则基于 f_i 的重要性,f_i 的权重 $W(f_i)$ 计算公式如下:

$$W(f_i)=\frac{I(f_i)}{\sum_{f_j\in F}I(f_j)} \tag{7}$$

其中,F 表示用户人格特质识别特征集合.

4 基于加权非负矩阵分解的用户人格特质识别模型

4.1 问题定义

在心理学中,以五大因子模型 FFM(Five Factor Model)[45]为基础,用户的五大人格特质是以五个维度对用户人格特质进行定义的,分别为外向型(extraversion,简记为 EXT)、神经质型(neuroticism,

① http://www.keenage.com/download/sentiment.rar

简记为 NEU)、随和型(agreeableness,简记为 AGR)、责任型(conscientiousness,简记为 CON)和开放型(openness,简记为 OPN). 文献[46]又进一步将每一种人格特质分为两个级别:外向型分为外向级别和害羞级别;神经质型分为神经质级别和安全级别;随和型分为友善级别和不愿合作级别;责任型分为精确级别和粗心级别;开放型分为洞察力级别和缺乏想象力级别. 方便起见,拟将每一维人格特质的两个级别统一简记为正向级别和负向级别.

文献[47]中曾指出因为纯加性的和稀疏的描述能使其具有更好的解释性,便于数据可视化、减少计算量和传输存储,还因为相对稀疏性的表示方式能在一定程度上抑制由外界变化给特征提取带来的不利影响,所以非负矩阵分解方法已逐渐成为信号处理、生物医学工程、模式识别、计算机视觉和图像工程等研究领域中最受欢迎的多维数据处理工具之一. 由于能够反映用户人格特质的特征较少,而且用户人格特质矩阵是非负、低秩且稀疏的,因此本文首次将用户人格特质识别问题转化为求解非负矩阵分解的最优解问题. 令 $u=\{u_1,u_2,\cdots,u_m\}$ 表示用户集合,其中 m 表示用户数量;$\boldsymbol{R}\in\mathbb{R}^{m\times n}$ 表示用户-用户人格特质矩阵($\boldsymbol{R}$ 中作为测试集部分的用户的各维人格特质级别为缺省值),其中,n 表示用户人格特质维数,则用户 u_i 第 j 维人格特质识别结果分为以下 2 种情况:

情况 1. 正向级别,$R_{ij}=1$,即 u_i 具有正向级别的第 j 维人品性 p_j;

情况 2. 负向级别,$R_{ij}=0$,即 u_i 具有负向级别的第 j 维人品性 p_j;

则本文将社会网络中的用户人格特质识别问题定义为:首先,给定用户-用户人格特质矩阵 $\boldsymbol{R}$ 和用户-特征矩阵 $\boldsymbol{U}$,找到非负矩阵 $\boldsymbol{V}$,使其满足 $\boldsymbol{R}\approx\boldsymbol{UV}$,矩阵 $\boldsymbol{R}$ 中的缺省值,即待预测用户具有各维正向级别人格特质的可能性,可以通过分解后得到的矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$ 的乘积获取,从而实现用户人格特质的识别.

4.2　模型算法

首先,拟通过因式分解,将 $\boldsymbol{R}$ 分解为矩阵 $\boldsymbol{U}\in\mathbb{R}^{m\times d}$ 和矩阵 $\boldsymbol{V}\in\mathbb{R}^{d\times n}$,其中,$\boldsymbol{U}$ 为用户-特征矩阵,$d\ll m$ 为用户人格特质识别特征数量,$\boldsymbol{V}$ 为 $\boldsymbol{R}$ 与 $\boldsymbol{U}$ 低秩表示之间的关系. 拟通过式(8)最小化识别值与实际值之间的均方误差:

$$\min_{\boldsymbol{U},\boldsymbol{V}}\|\boldsymbol{R}-\boldsymbol{UV}\|_F^2 \tag{8}$$

其中,$\|\cdot\|_F$ 为 Frobenius 范数. 为避免发生过拟合,拟在式(8)的基础上,加入 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的正则化的 Frobenius 范数:

$$\min_{\boldsymbol{U},\boldsymbol{V}}\|\boldsymbol{R}-\boldsymbol{UV}\|_F^2+\lambda_1\|\boldsymbol{U}\|_F^2+\lambda_2\|\boldsymbol{V}\|_F^2 \tag{9}$$

其中 λ_1 和 λ_2 为正则化参数.

此外,尽管五大用户人格特质在界定时其间并不存在重叠[48],文献[25]和文献[49]均指出每个用户可能会同时具有两种或两种以上的人格特质,此外,文献[25]和文献[35]还指出用户的五大人格特质间存在较弱的相关性. 因此,拟在识别用户人格特质时将其间的弱相关性考虑在内以提升算法的精确度. 本文假设相关性较大的用户人格特质间的特征差异较小,因此,为约束不同维度用户人格特质特征间的差异,拟引入正则化的用户人格特质相关因子:

$$\sum_{j=1}^{n-1}\boldsymbol{PC}(j,j+1)\|\boldsymbol{V}_{*j}-\boldsymbol{V}_{*(j+1)}\|_F^2=Tr(\boldsymbol{V}^{\mathrm{T}}\mathcal{L}\boldsymbol{V}) \tag{10}$$

其中,$\boldsymbol{V}_{*j}$ 和 $\boldsymbol{V}_{*(j+1)}$ 分别表示 $\boldsymbol{V}$ 中用户人格特质 p_j 和 p_{j+1} 的特征向量;$Tr(\cdot)$ 为矩阵的迹;$\mathcal{L}=\boldsymbol{D}-\boldsymbol{PC}$ 为拉普拉斯矩阵,$\boldsymbol{D}$ 为对角矩阵,$\boldsymbol{D}$ 中的第 i 个元素 $\boldsymbol{D}(i,i)$ 等于 $\boldsymbol{PC}$ 中第 i 行元素之和;$\boldsymbol{PC}(j,j+1)$ 为用户人格特质 p_j 和 p_{j+1} 之间的用户人格特质相关因子,并且本文假设 $\boldsymbol{PC}(j,j+1)$ 越大,p_j 和 p_{j+1} 之间相似度越大,则其间 Frobenius 范数就越小. 用户人格特质 p_j 和 p_{j+1} 之间的用户人格特质相关因子其计算公式如下所示:

$$\boldsymbol{PC}(j,j+1)=\frac{1}{D_{\text{J-S}}(\boldsymbol{ps}_j,\boldsymbol{ps}_{j+1})} \tag{11}$$

其中,$\boldsymbol{ps}_j$ 和 $\boldsymbol{ps}_{j+1}$ 表示用户人格特质 p_j 和 p_{j+1} 的分数向量;$D_{\text{J-S}}(\boldsymbol{ps}_j,\boldsymbol{ps}_{j+1})$ 表示 $\boldsymbol{ps}_j$ 和 $\boldsymbol{ps}_{j+1}$ 之间的 J-S 散度(Jensen-Shannon divergence)[50],J-S 散度越小表示不同级别的用户人格特质间的一致性越高,反之亦然,其计算公式如下:

$$D_{\text{J-S}}(\boldsymbol{ps}_j,\boldsymbol{ps}_{j+1})=\frac{D_{\text{K-L}}(\boldsymbol{ps}_j\|\overline{ps})+D_{\text{K-L}}(\boldsymbol{ps}_{j+1}\|\overline{ps})}{2} \tag{12}$$

其中,$\overline{ps}$ 表示用户人格特质 $\boldsymbol{ps}_j$ 和 $\boldsymbol{ps}_{j+1}$ 分数的平均分布;$D_{\text{K-L}}(\boldsymbol{ps}_j\|\overline{ps})$ 表示 $\boldsymbol{ps}_j$ 与 $\overline{ps}$ 之间的 K-L 散度(Kullback-Leiblerdivergence)[51],其计算公式如下:

$$D_{\text{K-L}}(\boldsymbol{ps}_j\|\overline{ps})=\sum_m\boldsymbol{ps}_j(m)\log\frac{\boldsymbol{ps}_j(m)}{\overline{ps}(m)} \tag{13}$$

其中,$\boldsymbol{ps}_j(m)$ 表示第 m 个用户的 p_j 的分数;$\overline{ps}(m)$ 表示第 m 个用户 p_j 和 p_{j+1} 的平均分数. 同理可得 $D_{\text{K-L}}(\boldsymbol{ps}_{j+1}\|\overline{ps})$.

通过在式(9)中引入上述正则化的用户人格特质相关性因子,得到如下目标函数:

$$\min_{\boldsymbol{U},\boldsymbol{V}} F_1 = \|\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V}\|_F^2+\lambda_1\|\boldsymbol{U}\|_F^2+\lambda_2\|\boldsymbol{V}\|_F^2+\lambda_3 Tr(\boldsymbol{V}^{\mathrm{T}}\mathcal{L}\boldsymbol{V}) \tag{14}$$

其中，若 $\boldsymbol{V}$ 固定，则 F_1 是一个关于 $\boldsymbol{U}$ 的凸函数；若 $\boldsymbol{U}$ 固定，则 F_1 是一个关于 $\boldsymbol{V}$ 的凸函数；若 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 均不固定，则 F_1 是一个非凸函数，因此，较难形式化 F_1 的全局最优解. 然而，F_1 的局部最优解可以通过乘性迭代方法求得[52]. 为计算 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的更新规则，将式(14)目标函数中的常数去掉后，其拉格朗日函数如下所示：

$$\mathcal{L}_{F_1} = Tr((\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V})(\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V})^{\mathrm{T}})+\lambda_1 Tr(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})+\lambda_2 Tr(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}})+\lambda_3 Tr(\boldsymbol{V}^{\mathrm{T}}\mathcal{L}\boldsymbol{V})-Tr(\psi\boldsymbol{U})-Tr(\varphi\boldsymbol{V}) \tag{15}$$

其中，ψ 和 φ 分别是 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的非负的拉格朗日乘子. 然后，分别计算式(15)中关于 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的梯度，并设其为 0：

$$\begin{cases}\dfrac{\partial\mathcal{L}_{F_1}}{\partial\boldsymbol{U}} = -2\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\lambda_1\boldsymbol{U}-\psi=0\\ \dfrac{\partial\mathcal{L}_{F_1}}{\partial\boldsymbol{V}} = -2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\lambda_2\boldsymbol{V}+\lambda_3\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{P}\boldsymbol{C})-\varphi=0\end{cases} \tag{16}$$

在式(16)中关于 $\boldsymbol{U}$ 的梯度等式两边同时乘以 $\boldsymbol{U}$，关于 $\boldsymbol{V}$ 的梯度等式两边同时乘以 $\boldsymbol{V}$：

$$\begin{cases}-2\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}+\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}+\lambda_1\boldsymbol{U}\boldsymbol{U}-\psi\boldsymbol{U}=0\\ -2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{V}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}+\lambda_2\boldsymbol{V}\boldsymbol{V}+\lambda_3\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{P}\boldsymbol{C})\boldsymbol{V}-\varphi\boldsymbol{V}=0\end{cases} \tag{17}$$

根据 KKT (Karush-Kuhn-Tueker)条件：$\psi\boldsymbol{U}=0$ 且 $\varphi\boldsymbol{V}=0$，可以得到式(18)：

$$\begin{cases}-2\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}+\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}+\lambda_1\boldsymbol{U}\boldsymbol{U}=0\\ -2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{V}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}+\lambda_2\boldsymbol{V}\boldsymbol{V}+\lambda_3\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{P}\boldsymbol{C})\boldsymbol{V}=0\end{cases} \tag{18}$$

其中 $\boldsymbol{R}$，$\boldsymbol{D}$ 和 $\boldsymbol{P}\boldsymbol{C}$ 中的每一个元素均为非负，λ_1，λ_2 和 λ_3 也是非负的. 此外，$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 中每一个初始值都是非负的，因此，$\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}$，$\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$，$\lambda_1\boldsymbol{U}$，$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}$，$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}$，$\lambda_2\boldsymbol{V}$，$\lambda_3\boldsymbol{V}^{\mathrm{T}}\boldsymbol{D}$ 和 $\lambda_3\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}$ 中的每一个元素亦是非负的，因此，式(18)可以写作：

$$\begin{cases}(\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\lambda_1\boldsymbol{U})\boldsymbol{U}=2\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}\\ (\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\lambda_2\boldsymbol{V}+\lambda_3\boldsymbol{V}^{\mathrm{T}}\boldsymbol{D})\boldsymbol{V}=(2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}+\lambda_3\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C})\boldsymbol{V}\end{cases} \tag{19}$$

则 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的更新规则计算如下：

$$\begin{cases}\boldsymbol{U}\leftarrow\boldsymbol{U}\dfrac{2\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}}{\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\lambda_1\boldsymbol{U}}\\ \boldsymbol{V}\leftarrow\boldsymbol{V}\dfrac{2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}+\lambda_3\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}}{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\lambda_2\boldsymbol{V}+\lambda_3\boldsymbol{V}^{\mathrm{T}}\boldsymbol{D}}\end{cases} \tag{20}$$

基于以上分析，基于加权非负矩阵分解的用户人格特质识别模型构建如算法 1 所示.

算法 1. 基于加权非负矩阵分解的用户人格特质识别模型构建.

输入：用户-用户人格特质矩阵 $\boldsymbol{R}$；用户-特征矩阵 $\boldsymbol{U}$；用户人格特质相关因子矩阵 $\boldsymbol{P}\boldsymbol{C}$；正则化参数 $\lambda_1,\lambda_2,\lambda_3$

输出：用户人格特质识别矩阵 $\boldsymbol{R}'$

1. FOR $\boldsymbol{U}$ 中的每一列 g DO
2. 根据式(7)计算其所表示的特征的权重 $W(f_g)$
3. 以 $W(f_g)$ 乘以列 g 中所有元素的原始值，得到带有权重的特征值
4. END FOR
5. 初始化 $\boldsymbol{V}\leftarrow(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U})^{-1}\boldsymbol{R}$，并将 $\boldsymbol{V}$ 中所有小于 0 的元素设置为 0
6. REPEAT
7. 根据式(20)更新 $\boldsymbol{U}$ 和 $\boldsymbol{V}$
8. UNTIL 式(14)中的 F_1 收敛
9. RETURN $\boldsymbol{R}'\leftarrow\boldsymbol{U}\boldsymbol{V}$

4.3 时间复杂度分析

假设数据集规模为 m，特征数量为 d，迭代次数为 T. 基于加权非负矩阵分解的用户人格特质识别模型的时间复杂度计算包括两个阶段——特征预处理阶段和模型构建阶段. 在特征预处理阶段(即步骤 1～步骤 4)，首先，分别计算各用户人格特质识别特征与结果之间的肯德尔相关系数，以及用户人格特质识别特征之间的肯德尔相关系数，其时间复杂度为 $O(m^2)$；然后，根据式(7)为每个特征分配权重，其时间复杂度为 $O(d)$；最后，对训练集合中每个实例的每个特征赋予相应的权重，其时间复杂度为 $O(dm)$，由于实际训练中特征数量 d 远远小于数据集规模 m，所以特征预处理阶段总的时间复杂度为 $O(m^2)$. 在模型构建阶段(即步骤 5～步骤 9)，步骤 6～步骤 9 中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的更新规则在提出算法的时间复杂度中占较高比重；由于 $\boldsymbol{R}$ 和 $\boldsymbol{P}\boldsymbol{C}$ 都是稀疏的，$\boldsymbol{D}$ 为对角矩阵，故 $\boldsymbol{R}\boldsymbol{V}^{\mathrm{T}}$、$\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$、$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}$、$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}$ 和 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{D}$ 的时间复杂度分别为 $O(nmd)$、$O(\max(n,m)d^2)$、$O(nmd)$、$O(\max(n,m)d^2)$ 和 $O(nd)$，由于 $d\ll\max(n,m)$，因此，模型构建阶段的时间复杂度为 $T\times O(nmd)$. 综上，基于加权非负矩阵分解的链接识别模型的时间复杂度为 $O(m^2)+T\times O(nmd)$.

5 实验及结果分析

5.1 数据集及实验设置

myPersonality① 是一个较为流行的 Facebook

① http://mypersonality.org

应用，其通过 336 道心理测试题测量出能够反映出用户潜在五大人格特质的 30 个方面. 虽然参与该应用的用户来自不同年龄段、不同背景和不同的文化程度，但是每一个用户都是以很认真诚实地态度回答测试中的每一道问题. 在该应用中用户的心理学特征和用户概要是经过该用户所在的朋友圈“验证”过的，此外，用户是在没有压力的情况下进行测试并且无需分享其用户概要信息，故 myPersonality 避免了蓄意伪造数据情况的出现. 因此，该应用中用户的心理学特征和 Facebook 用户概要可以为研究所用. 目前，该项应用的数据库已经包含超过 6 000 000 的测试结果和超过 4 000 000 个用户的用户概要.

Moore 等人[53]通过对 219 名大学生的调查问卷和日志数据分析 Facebook 用户人格特质，指出在 Facebook 中挖掘用户的五大人格特质是可行的. 因此，为研究社会网络中用户人格特质识别的问题，本文选用 2013 年 ICWSM 会议人格特质识别的专题讨会中提供的 Facebook 用户人格特质标准数据集验证提出方法的有效性，该数据集是 myPersonality 数据集的一个子集，其收集了 Facebook 用户的人格特质(包括通过采用 100 项加长版的 IPIP① 人格问卷进行自我评估而得到的用户人格特质分数和基于黄金标准划分的用户人格特质标签)、社会网络结构特征(包括网络规模、中介中心性、密度、中介性、传递性等)和其发布状态，其统计数据如表 2 所示.

表 2　Facebook 用户人格特质标准数据集统计数据

用户数	用户发布状态数
250	9900

本文的实验设置如下：随机将数据集分为两部分：A 和 B. A 为训练集合，占数据集的 90%；余下的 10%记作 B，作为测试集合. 为确保实验结果的可靠性，本文采用 10 折交叉验证对实验结果进行评估.

5.2　评估方法

由于研究者们基于不同的数据集进行用户人格特质识别工作，并采用不同的实验评估方法，因此很难对其性能进行充分地评价. 然而，在 2013 年ICWSM 会议中用户人格特质识别的专题讨会中，基于相同的标准数据集，研究者们可以按照其意愿随意分割训练集合和测试集合，并统一使用精确度、召回率和 $F1$ 值对实验结果进行评估. 因此，为便于与上述专题研讨会中的工作进行比较，本文亦采用精确度、召回率和 $F1$ 值对实验结果进行评估，则关于正向级别 p_j 的精确度、召回率和 $F1$ 值，其定义分别如式(21)、(22)和(23)所示.

$$Precision_j^+=\frac{tp_j^+}{tp_j^++fp_j^+} \tag{21}$$

$$Recall_j^+=\frac{tp_j^+}{tp_j^++fn_j^+} \tag{22}$$

$$F1_j^+=\frac{2\times Precision_j^+\times Recall_j^+}{Precision_j^++Recall_j^+} \tag{23}$$

其中，tp_j^+ 表示 p_j 分类正确的具有正向级别的实例数目；fp_j^+ 表示 p_j 分类错误的具有正向级别的实例数目；fn_j^+ 表示 p_j 分类错误的具有负向级别的实例数目. 同理可得关于负向级别 p_j 的精确度、召回率和 $F1$ 值：$Precision_j^-$、$Recall_j^-$ 和 $F1_j^-$，则 p_j 的平均精确度、召回率和 $F1$ 值其定义分别如式(24)、(25)和(26)所示.

$$Precision_j=\frac{Precision_j^++Precision_j^-}{2} \tag{24}$$

$$Recall_j=\frac{Recall_j^++Recall_j^-}{2} \tag{25}$$

$$F1_j=\frac{F1_j^++F1_j^-}{2} \tag{26}$$

5.3　实验结果分析

本文主要针对以下几个问题展开对比实验以验证提出方法的有效性：

(1) 用户人格特质之间是否存在相关性？

(2) 提出的用户人格特质识别模型，WNMF-MPTR，效率是否优于其他的用户人格特质识别算法？

(3) 特征权重和用户人格特质相关因子对 WNMF-MPTR 的性能是否有影响？

5.3.1　用户人格特质间的相关性验证

文献[46]将每一种人格特质分为两个级别，而不同级别的用户人格特质间可能具有不同的相关性，因此，拟根据第 4.1 节中定义的用户人格特质的正向级别和负向级别对用户采用分组处理：例如，若某一用户具有正向级别的随和型人格特质，另一用户也具有正向级别的随和型人格特质，则这两个用户将会被分到同一组内；若某一用户具有正向级别的随和型人格特质，而另一用户具有负向级别的随和型人格特质，则这两个用户将会被分到不同的组内.

由于仅根据一种衡量标准并不能准确分析不同用户人格特质间的相关性，文献[25]中提出采用皮

① http://ipip.ori.org/newFinding_Labeling_IPIP_Scales.htm

尔森相关系数度量新浪微博中用户人格特质间的相关性,但其未根据用户人格特质的不同级别对用户进行分组度量.此外,如第 3.2 节所述,皮尔森相关系数存在一定的局限性,相比于皮尔森相关系数,肯德尔相关系数的适用范围要广些.因此,拟基于分组后的用户人格特质分数序列,在采用式(12)计算不同级别用户人格特质之间的 J-S 散度的同时,再利用式(3)中定义的肯德尔相关系数以对不同级别用户人格特质之间的相关性进行验证.表 3 为不同级别用户人格特质间的 J-S 散度实验结果,表 4 为不同级别用户人格特质间的肯德尔相关系数实验结果.

表 3 不同级别用户人格特质间的 J-S 散度

		EXT		NEU		AGR		CON		OPN	
		p	n	p	n	p	n	p	n	p	n
EXT	p			0.58	14.79	0.44	2.19	0.70	2.36	0.73	0.84
	n			4.79	4.49	5.21	2.36	5.61	1.91	10.43	2.11
NEU	p					0.80	2.12	1.10	2.46	2.98	0.40
	n					19.78	2.94	16.97	3.05	24.73	3.61
AGR	p							0.77	3.29	0.82	0.91
	n							3.40	0.86	4.63	1.25
CON	p									1.15	1.17
	n									6.16	1.40
OPN	p										
	n										

表 4 不同级别用户人格特质间的肯德尔相关系数

		EXT		NEU		AGR		CON		OPN	
		p	n	p	n	p	n	p	n	p	n
EXT	p			0.22	−0.03	0.09	−0.15	−0.06	−0.13	0.07	0.01
	n			−0.23	−0.18	0.04	0.04	0.04	0.18	0.07	0.22
NEU	p					−0.09	−0.17	0.10	−0.21	−0.03	0.27
	n					−0.14	−0.13	0.02	−0.01	−0.06	−0.10
AGR	p							−0.07	0.07	0.16	0.29
	n							−0.08	0.16	0.06	0.02
CON	p									−0.01	−0.10
	n									0.07	−0.07
OPN	p										
	n										

表 3 和表 4 中,p 和 n 分别表示用户人格特质的正向级别和负向级别.J-S 散度越小表示不同级别的用户人格特质间的一致性越高;肯德尔相关系数绝对值越大,不同级别用户人格特质之间的相关性越高.

通过表 3 和表 4 可知,负向级别的外向型人格特质与正向级别的神经质型人格特质间的一致性相对较低且呈负相关关系,换言之,若某一用户外向型人格特质的分数较高,则该用户神经质型人格特质分数不可能很高.正向级别的外向型人格特质与正向级别的神经质型人格特质、正向级别的随和型人格特质间具有相对较高的一致性且呈正相关关系.负向级别的神经质型人格特质与正向级别的随和型人格特质、正向级别的责任型人格特质、正向级别的开放型人格特质间具有相对较低的一致性,且与正向级别的随和型人格特质和正向级别的开放型人格特质间呈负相关关系,与正向级别的责任型人格特质呈正相关关系.从实验结果可以看出,不同级别用户人格特质之间存在一定的弱相关性,与文献[35]中得出的结论一致,从而回答了本节开始提出的第一个问题.

此外,文献[54]通过实验证明了五大人格特质的收敛性及其区别联系.文献[55]指出 BFI 的每一个维度与相应的 NEO PI-R 中的维度间存在较强的收敛性.文献[41]以自我报告的形式,从领域和维度视角分别验证了用户人格特质的收敛性.总之,虽然使用不同的五大人格特质度量方法,用户人格特质之间的相关性同样存在于 myPersonality 之外的其他数据集中,进而证明了 myPersonality 数据集中关于五大人格特质的标注是有效的,因此,用户人格特质之间的相关性有助于提升用户人格特质识别算法的性能,这为本文提出的方法提供了较好的理论依据.

5.3.2 WNMF-MPTR 与其他用户人格特质识别方法的比较

首先,拟将基于不同用户人格特质识别特征集

合的 WNMF-MPTR 方法分别定义为：

(1) sWNMF-MPTR：仅基于社会网络特征；

(2) lWNMF-MPTR：仅基于语言学特征；

(3) eWNMF-MPTR：仅基于情感统计特征；

(4) WNMF-MPTR：同时基于社会网络特征、语言学特征和情感统计特征.

然后，通过 10 折交叉验证对其性能进行对比，实验结果如图 1 所示.

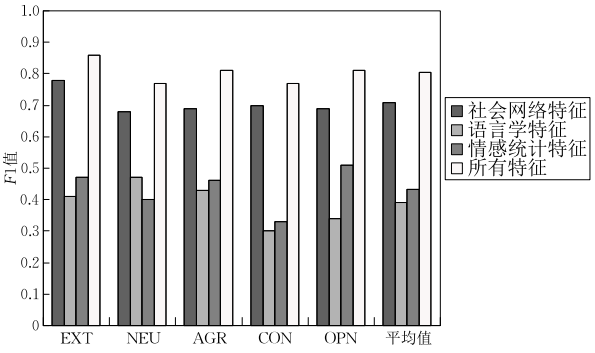


图 1　WNMF-MPTR 在不同用户人格特质识别特征集合上的性能比较

通过图 1 可知，情感特征的性能略高于语言学特征，这是由于尽管用户发表的言论中所表达出的情感能够较为真实地反映出其某一方面的人格特质，但情感特征仅针对形容词及其变型对语言学特征进行了进一步分析，因此，其二者的性能差别不大. 社会网络特征对于外向型人格特质的分类性能最佳，与文献[28]中的结论一致，并且相比于仅基于其他特征的方法而言，仅基于社会网络特征的方法具有较高的分类精度. 一方面是因为用户网络结构是一项能够较好反映用户人格特质的参数[44]，例如，具有正向级别 EXT 人格特质的用户和具有正向级别 CON 人格特质的用户一般拥有较多的朋友，但其朋友之间很少存在联系，即形成了一种特殊的用户网络结构——规模庞大但稀疏，NEU 的识别主要与用户网络的规模和传递性具有较强的关联，AGR 的识别主要与用户网络的传递性具有较强的关联；另一方面是由于社会网络特征是从用户注册账户开始统计的一组特征，而数据集中的用户语料并不是从用户注册账户开始收集的，故从中抽取的情感特征和语言学特征仅能反映最近一段时间内的用户习惯及状态. 综上，虽然基于社会网络特征能够较为准确地识别用户的人格特质，但是通过合理地融合社会网络特征、语言学特征和情感统计特征，能够弥补传统用户人格特质识别算法的不足，并达到更高的分类精度.

此外，为验证用户人格特质识别模型 WNMF-MPTR 的有效性，基于相同的 Facebook 标准数据集，拟将 WNMF-MPTR 与 2013 年 ICWSM 会议中用户人格特质识别的专题讨会中的相关工作进行比较，表 5、表 6 和表 7 中为每一种方法关于每一维用户人格特质的精确度、召回率和 $F1$ 值，最大值以粗体标示.

表 5　WNMF-MPTR 与用户人格特质识别方法的精确度比较

相关工作	方法	EXT	NEU	AGR	CON	OPN	平均值
WNMF-MPTR	weighted NMF	**0.95**	**0.87**	**0.93**	**0.88**	**0.89**	**0.904**
文献[27]	SVM	0.79	0.71	0.67	0.72	0.87	0.752
文献[28]	SVM, kNN, NB	0.58	0.54	0.50	0.55	0.60	0.554
文献[29]	SVM, BLR, mNB	0.58	0.59	0.59	0.59	0.60	0.590
文献[32]	LR	NA	NA	NA	NA	NA	NA

表 6　WNMF-MPTR 与用户人格特质识别方法的召回率比较

相关工作	方法	EXT	NEU	AGR	CON	OPN	平均值
WNMF-MPTR	weighted NMF	**0.79**	0.69	**0.72**	0.68	0.75	0.726
文献[27]	SVM	**0.79**	**0.72**	0.68	**0.72**	**0.87**	**0.756**
文献[28]	SVM, kNN, NB	0.61	0.53	0.50	0.54	0.70	0.576
文献[29]	SVM, BLR, mNB	0.58	0.58	0.59	0.59	0.60	0.588
文献[32]	LR	NA	NA	NA	NA	NA	NA

表 7　WNMF-MPTR 与用户人格特质识别方法的 $F1$ 值比较

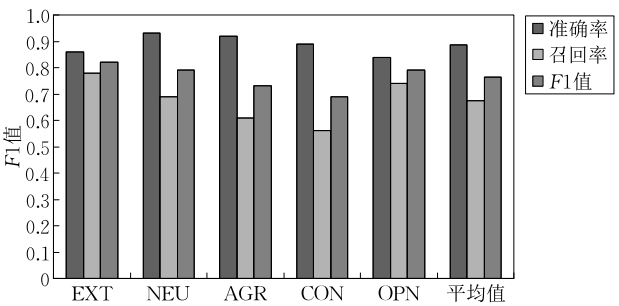
相关工作	方法	EXT	NEU	AGR	CON	OPN	平均值
WNMF-MPTR	weighted NMF	**0.86**	**0.77**	**0.81**	**0.77**	0.81	**0.804**
文献[27]	SVM	0.79	0.70	0.67	0.72	**0.86**	0.748
文献[28]	SVM, kNN, NB	0.56	0.52	0.50	0.54	0.61	0.546
文献[29]	SVM, BLR, mNB	0.58	0.58	0.58	0.59	0.60	0.586
文献[32]	LR	NA	NA	NA	NA	NA	0.630

从表5、表6和表7中的实验结果可知，Verhoeven等人[27]同时基于Facebook用户人格特质标准数据集和Eassys用户人格特质标准数据集构建一个集成的SVM模型，而本文中仅基于Facebook用户人格特质标准数据集构建用户人格特质识别模型，故当仅在Facebook用户人格特质标准数据集上进行测试时，WNMF-MPTR的性能要低于文献[27]中提出的方法.由于本文采用非负矩阵分解方法的同时对用户人格特质之间的相关性以及特征权重这两个因素进行考虑，因此，当在Facebook用户人格特质标准数据集上进行训练和测试时，相比于文献[28-29]和文献[32]中提出的方法，WNMF-MPTR具有较优越的性能.

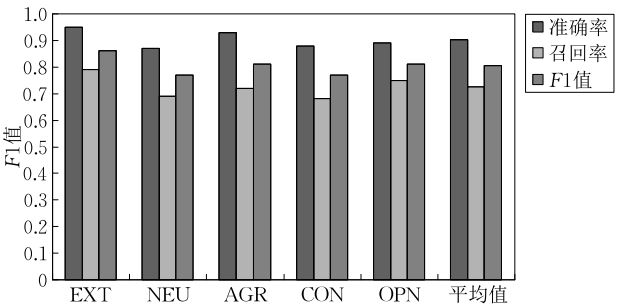
上述实验结果表明，相比于其他方法，本文利用基于加权非负矩阵分解方法识别社会网络中用户的五大人格特质，使得用户人格特质识别算法的综合性能得到了较大提升，从而回答了本节开始提出的第二个问题.

5.3.3 特征权重对WNMF-MPTR性能的影响

首先，为验证特征权重对WNMF-MPTR性能的影响，拟分别在带有权重的数据集上和不带有权重的数据集上通过10折交叉验证比较WNMF-MPTR在Facebook用户人格特质标准数据集上的平均准确率、召回率和$F1$值，实验结果如图2所示.



(a) 基于不带有权重的特征集合



(b) 基于带有权重的特征集合

图2 特征权重对WNMF-MPTR性能的影响

图2中的实验结果表明，WNMF-MPTR在带有权重的数据集上的性能优于其在不带有权重的数据集上的性能.当采用不带有权重的数据集时，即认为所有特征具有相同的重要性，会导致用户个人品性预测结果被大量重要性较低的特征支配，而根据特征重要性为其分配权重，有助于缩短对应于不太重要特征的坐标轴，拉长对应于更重要特征的坐标轴.这种伸展坐标轴的优化过程，抑制了重要性较低的特征的影响.综上，在WNMF-MPTR中考虑特征权重有助于提升算法性能.

此外，除肯德尔相关系数法外，拟采用熵权法为特征赋予权重，从而衡量以肯德尔相关系数法确定的特征权重的优劣.

熵权法中假设特征的信息熵越小，该特征值的变异程度越大，其提供的信息量也就越大，那么，其在待预测问题中所起作用越大，权重就越高，则利用熵权法计算特征权重的计算公式如下：

$$W''(f_i)=\frac{1-H(f_i)}{\|F\|-\sum_{f_j\in F}H(f_i)} \tag{27}$$

其中，$W''(f_i)$表示特征f_i的权重值并且$0\leq W''(f_i)\leq 1$，$\sum_{f_j\in F}W''(f_i)=1$；$F$表示用户人格特质识别特征集合；$\|F\|$表示用户人格特质识别特征集合中包含的元素数量；$H(f_i)$表示$f_i$的熵权，其计算公式如下：

$$H(f_i)=-\frac{1}{\ln m}\sum_{j=1}^{m}z(j,f_i)\ln z(j,f_i) \tag{28}$$

其中，m表示用户数量，$z(j,f_i)$表示第j个用户的特征f_i的标准化值，其计算公式如下：

$$z(j,f_i)=\frac{y(j,f_i)}{\sum_{j=1}^{m}y(j,f_i)} \tag{29}$$

其中，$y(j,f_i)$表示第j个用户的特征f_i的取值.表8中的权重值1和权重值2分别为由肯德尔相关系数法和熵权法计算得到的不同用户人格特质识别特征的权重值.

由表8可知，基于肯德尔相关系数法和熵权法计算得到的各类特征的平均权重大小关系均为：社会网络特征＞情感统计特征＞语言学特征，而5.3.2节图1中WNMF-MPTR在不同用户人格特质识别特征集合上的实验结果大小关系亦为sWNMF-MPTR＞eWNMF-MPTR＞lWNMF-MPTR，此外，文献[44]和文献[28]以皮尔森相关系数对各特征与五大人格特质间的相关性进行度量，实验结

果表明，社会网络特征与五大人格特质间的皮尔森相关系数绝对值均较高，即拥有较强的相关性. 由此可见，采用肯德尔相关系数法和熵权法对不同用户人格特质识别特征赋予的权重值均是合理的. 然而，如表 8 中所示，基于肯德尔相关系数法计算得到的特征权重值与基于熵权法计算得到的特征权重值并不是完全一致的. 因此，为进一步比较两种特征权重计算方法，拟分别基于由肯德尔相关系数法和熵权法计算得到的带权重的特征集合对用户的人格特质进行识别，通过 10 折交叉验证对其性能进行对比.

表 8　不同用户人格特质识别特征权重值

特征	权重值 1	权重值 2	特征	权重值 1	权重值 2
注册时间	0.023	0.029	小品词	0.018	0.024
网络规模	0.021	0.026	断句符	0.018	0.014
介数中心性	0.025	0.027	感叹词	0.021	0.027
归一化的介数中心性	0.032	0.036	基本形式的动词	0.016	0.014
密度	0.031	0.032	过去时态的动词	0.017	0.013
中介性	0.027	0.029	动名词	0.016	0.015
归一化的中介性	0.032	0.032	过去分词形式的动词	0.016	0.013
传递性	0.022	0.027	现在时态的动词(非第三人称单数)	0.016	0.015
连词	0.017	0.015	现在时态的动词(第三人称单数)	0.017	0.014
数词	0.016	0.010	从属词	0.017	0.018
限定词	0.016	0.014	基本形式的形容词	0.021	0.026
方位词	0.016	0.009	比较形容词	0.022	0.028
外来词	0.020	0.024	最高级形容词	0.022	0.027
情态助动词	0.017	0.012	列表项标记	0.019	0.014
物质名词	0.017	0.011	WH-限定词	0.017	0.015
复数名词	0.016	0.013	WH-代词	0.019	0.019
专有名词	0.021	0.022	WH-复数代词	0.017	0.016
复数专有名词	0.022	0.020	WH-副词	0.018	0.021
前位限定词	0.022	0.016	词语总数	0.016	0.017
所有格标记	0.030	0.019	逗号使用频率	0.016	0.012
人称代词	0.017	0.026	句号使用频率	0.016	0.010
复数人称代词	0.017	0.023	感叹号使用频率	0.020	0.018
基本形式的副词	0.021	0.022	问号使用频率	0.018	0.019
比较副词	0.020	0.024	正面情感词比例	0.023	0.024
最高级副词	0.021	0.022	负面情感词比例	0.025	0.027

实验结果表明，WNMF-MPTR 的平均 $F1$ 值在基于由肯德尔相关系数法计算得到的带权重的特征集合上比在基于熵权法计算得到的带权重的特征集合上能够提升 12.7%. 由于熵权法在计算特征权重时均是根据特征值本身的变异程度衡量其重要性，而肯德尔相关系数法在计算特征权重时主要侧重考虑特征值与用户人格特质分数间的联系，使其能够更为确切地反映不同特征对用户人格特质识别的影响，因此，肯德尔相关系数法在计算用户人格特质识别特征的权重时要优于熵权法.

5.3.4　用户人格特质相关因子对 WNMF-MPTR 性能的影响

拟通过参数 λ_3 以验证用户人格特质相关因子对 WNMF-MPTR 性能的影响. 实验设置如下：λ_3 分别取值 0、0.2、0.5、0.7、1，为避免由于训练集合规模导致实验结果出现偏差，本文分别以 A 中 50%、60%、80%和 100%的数据作为训练集合，在其上进行 10 折交叉验证，比较其平均 $F1$ 值，实验结果如图 3 所示.

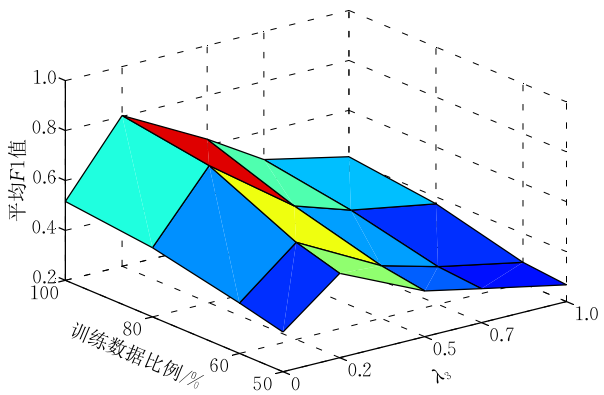


图 3　用户人格特质相关因子对 WNMF-MPTR 性能的影响

通过比较不同的 λ_3 值可知：当 $\lambda_3=0$ 时，即不考虑用户人格特质相关因子对用户人格特质识别问题的影响，此时的 $F1$ 值比峰值低很多，并且随着 λ_3 的增加，$F1$ 值先不断增加，在其到达峰值之后又迅速下降. 当 λ_3 较大时，用户人格特质识别学习过程主要受用户人格特质相关因子控制，此时通过学习得到的矩阵 $\boldsymbol{V}$ 会出现失真，从而不能够得到精确的用户人格特质识别结果. 由此可见，将用户人格特质间

的相关性融入用户人格特质识别问题中可以有效地提高识别算法的性能.

为进一步验证每一种人格特质与其他人格特质之间的弱相关性对 WNMF-MPTR 性能的影响,拟利用弃一法,每次将一种用户人格特质独立出来,忽略该人格特质与其他人格特质之间的弱相关性,即在第 4.2 节式(10)中提及的用户人格特质相关因子中仅考虑余下 4 种人格特质之间的相关性,并对其进行交叉验证,实验结果如图 4 所示.

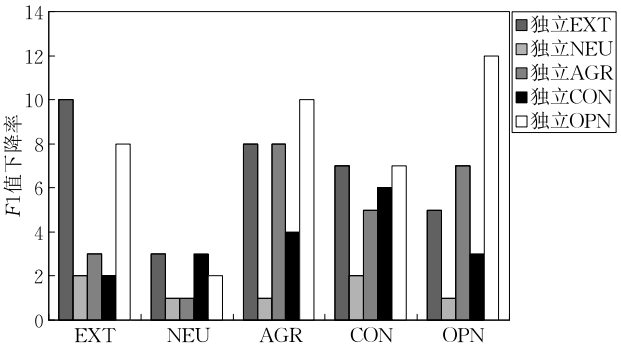


图 4 不同用户人格特质间的弱相关性对 WNMF-MPTR 性能的影响

从图 4 可知,当分别将随和型人格特质、责任型人格特质、外向型人格特质、开放型人格特质独立出来时,$F1$ 值会出现大幅度的下降,其平均 $F1$ 值下降值分别为 5%、4%、7%、8%;当将神经质型人格特质独立出来时,$F1$ 值下降并不明显,其平均 $F1$ 值下降值为 1%.由此可见,不同用户人格特质与其他人格特质之间的弱相关性对 WNMF-MPTR 性能的影响有所不同:随和型人格特质、责任型人格特质、外向型人格特质和开放型人格特质与其他人格特质之间的弱相关性对用户人格特质识别贡献较大,而神经质型人格特质与其他用户人格特质之间的弱相关性对用户人格特质识别贡献相对较小.

此外,为探究不同相关性度量方法对 WNMF-MPTR 性能的影响,拟分别采用 J-S 散度和肯德尔相关系数度量用户人格特质间的弱相关性,其中,以肯德尔相关系数度量用户人格特质间的弱相关性时,由于肯德尔相关系数绝对值越大,不同级别用户人格特质间的相关性越高,故其与用户人格特质间的弱相关性成正比,并且肯德尔相关系数的取值范围为[−1,1],为满足条件:非负矩阵中的所有元素均为非负,需要将其取值映射到[0,1]区间上,则式(11)修改为

$$\boldsymbol{PC}(j,j+1)=\left|\tau(\boldsymbol{ps}_j,\boldsymbol{ps}_{j+1})\right| \tag{30}$$

对上述两种相关性度量方法进行 10 折交叉验证,实验结果如图 5 所示.

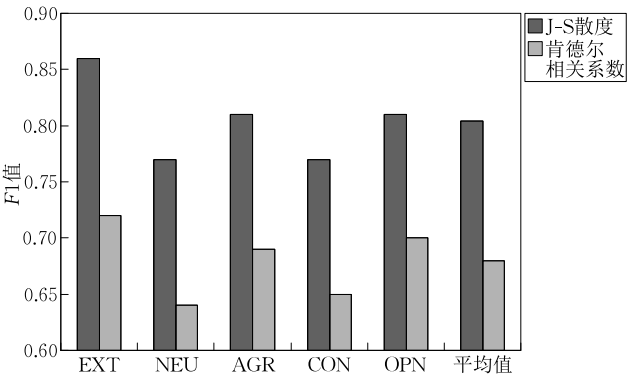


图 5 不同相关性度量方法对 WNMF-MPTR 性能的影响

肯德尔相关系数和 J-S 散度的取值范围分别为[−1,1]和[0,1],当肯德尔相关系数以绝对值的形式被映射到[0,1]区间上时,弱化了用户人格特质间正相关关系和负相关关系之间的差别,而 J-S 散度实际的取值范围是符合前述非负矩阵分解方法对矩阵中元素的要求的.因此,如图 5 所示,相比于肯德尔相关系数,以 J-S 散度度量用户人格特质间的弱相关性时,各维用户人格特质的 $F1$ 值均较高,由此可见,就非负矩阵分解方法而言,J-S 散度更适于度量用户人格特质间的弱相关性.

综上,5.3.3 节和 5.3.4 节中的实验结果进一步证明了特征权重和用户人格特质相关因子在用户人格特质识别问题中的重要性,从而回答了本节开始提出的第三个问题.

6 结 论

针对传统用户人格特质识别方法的不足,本文提出一种基于加权非负矩阵分解的用户人格特质识别算法.基于用户发布内容信息序列,该算法首先构建用户语言学特征和用户情感统计特征;然后,通过对用户网络结构特征、用户语言学特征和用户情感统计特征进行相关性分析,根据特征的重要性赋予其权重以得到带有权重的训练集合;最后,以用户人格特质相关性因子约束目标函数,构建基于加权非负矩阵分解的用户人格特质识别模型,从而实现用户五大人格特质的识别.在真实数据集上的实验结果表明,提出的算法能够有效地提高用户人格特质识别模型的性能.后续研究中,将群体智慧技术引入用户人格特质识别算法以更准确可靠地识别用户的五大人格特质将成为主要目标.

致　谢　在此，我们向对本文工作给予支持和建议的评审老师表示感谢！

参　考　文　献

[1] Zhang D Y, Guo G. A comparison of online social networks and real-life aocialnetworks: A study of sinamicroblogging. Mathematical Problems in Engineering, 2014: article ID578713

[2] Zhou Ya-Dong, Liu Xiao-Ming, Du You-Tian, Guan Xiao-Hong, Liu Ji. A method for identifying the evolutionary focuses of online social topics. Chinese Journal of Computers, 2015, 38(2): 261-271(in Chinese)
(周亚东，刘晓明，杜友田，管晓宏，刘霁. 一种网络话题的内容焦点迁移识别方法. 计算机学报，2015，38(2)：261-271)

[3] Cao Jiu-Xin, Dong Dan, Xu Shun, Zheng Xiao, Liu Bo, Luo Jun-Zhou. A k-core based algorithm for influence maximization in social networks. Chinese Journal of Computers, 2015, 38(2): 238-248(in Chinese)
(曹玖新，董丹，徐顺，郑啸，刘波，罗军舟. 一种基于 k-核的社会网络影响最大化算法. 计算机学报，2015，38(2)：238-248)

[4] Oussalah M, Bhat F, Challis K, Schnier T. A software architecture for Twitter collection, search and geolocation services. Knowledge-Based Systems, 2013, 37: 105-120

[5] McCrae R R, Costa P T. Validation of the 5-factor model of personality across instruments and observers. Journal of Personality and Social Psychology, 1987, 52(1): 81-90

[6] Saucier G, Goldberg L R. The structure of personality attributes//Barrick M R, Ryan A M eds. Personality and Work: Reconsidering the Role of Personality in Organizations. San Francisco, CA: Jossey-Bass, 2003: 1-29

[7] Caci B, Cardaci M, Tabacchi M E, Scrima F. Personality variables as predictors of Facebook usage. Psychological Reports, 2014, 114(2): 528-539

[8] Kosinski M, Bachrach Y, Kohli P, et al. Manifestations of user personality in website choice and behaviour on online social networks. Machine Learning, 2014, 95(3): 357-380

[9] Cojocaru M G, Thille H, Thommes E, et al. Social influence and dynamic demand for new products. Environmental Modelling & Software, 2013, 50(12): 169-185

[10] Durupinar F, Pelechano N, Allbeck J, et al. The impact of the OCEAN personality model on the perception of crowds. IEEE Computer Graphics and Applications, 2011, 31(3): 22-31

[11] Feng X. Study on customer personality characteristics and relationship outcomes using SEM analysis//Proceedings of the International Conference on Mechatronics and Information Technology. Guilin, China, 2013: 841-844

[12] Lee J G, Moon S, Salamatian K. Modeling and predicting the popularity of online contents with Cox proportional hazard regression model. Neurocomputing, 2012, 76(1): 134-145

[13] Cobb-Clark D A, Schurer S. The stability of Big-Five personality traits. Economics Letters, 2012, 115(1): 11-15

[14] Mischel W. Toward an integrative science of the person. Annual Review of Psychology, 2004, 55(1): 1-22

[15] Stoughton J W, Thompson L F, Meade A W. Big five personality traits reflected in job applicants' social media postings. Cyber Psychology, Behavior and Social Networking, 2013, 16(11): 800-805

[16] Mairesse F, Walker M A, Mehl M R, Moore R K. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research, 2007, 30(1): 457-500

[17] Pennebaker J W, King L A. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 1999, 77(6): 1296-1312

[18] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. The 3rd Edition. Waltham, Mass: The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, 2011

[19] Quinlan J R. Learning with continuous classes//Proceedings of the 5th Australian Joint Conference on Artificial Intelligence. Adelaide, Australia, 1992: 343-348

[20] Oberlander J, Nowson S. Whose thumb is it anyway? classifying author personality from weblog text//Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, 2006: 627-634

[21] Nguyen T, Phung D Q, Adams B, Venkatesh S. Towards discovery of influence and personality traits through social link prediction//Proceedings of the International Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 566-569

[22] Golbeck J, Robles C, Turner K. Predicting personality with social media//Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems. Vancouver, Canada, 2011: 253-262

[23] Sam T R, Zoubin G. A unifying review of linear Gaussian models. Neural Computation, 1999, 11(2): 305-345

[24] Bai S, Zhu T, Cheng L. Big-five personality prediction based on user behaviors at social network sites. PLOS Neglected Tropical Diseases, 2012, 8(2): e2682-e2682

[25] Bai S T, Hao B B, Li A, et al. Predicting Big Five personality traits of microblog users//Proceedings of the 12th IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology. Atlanta, USA, 2013: 501-508

[26] Sun R, Wilson N. A model of personality should be a cognitive architecture itself. Cognitive Systems Research, 2014, 29-30(1): 1-30

[27] Verhoeven B, Daelemans W, Smedt T D. Ensemble methods for personality recognition//Proceedings of the Workshop on Computational Personality Recognition. Boston, USA, 2013: 35-38

[28] Farnadi G, Zoghbi S, Moens M F, Cock M D. Recognising personality traits using Facebook status updates//Proceedings of the Workshop on Computational Personality Recognition. Boston, USA, 2013: 14-18

[29] Alam F, Stepanov E A, Riccardi G. Personality traits recognition on social network-Facebook//Proceedings of the Workshop on Computational Personality Recognition. Boston, USA, 2013: 6-9

[30] Genkin A, Lewis D D, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics, 2007, 49(3): 291-304

[31] McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification//Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. Madison, Wisconsin, 1998: 137-142

[32] Tomlinson M T, Hinote D, Bracewell D B. Predicting conscientiousness through semantic analysis of Facebook posts//Proceedings of the Workshop on Computational Personality Recognition. Boston, USA, 2013: 31-34

[33] Ruczinski I, Kooperberg C, Leblanc M. Logic regression. Journal of Computational and Graphical Statistics, 2003, 12(3): 475-511

[34] Ortigosa A, Carro R M, Quiroga J I. Predicting user personality by mining social interactions. Journal of Computer and System Sciences, 2014, 80(1): 57-71

[35] John O P, Naumann L P, Soto C J. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues//John O P, Robins R W, Pervin L A eds. Handbook of personality: Theory and research (The 3rd Edition). New York: Guilford Press, 2008: 114-158

[36] Kosti M V, Feldt R, Angelis L. Personality, emotional intelligence and work preferences in software engineering: An empirical study. Information and Software Technology, 2014, 56(8): 973-990

[37] Schwartz H A, Eichstaedt J C, Kern M L, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS One, 2013, 8(9): e73791

[38] Kosinski M, Stillwell D J, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 2013, 110(15): 5802-5805

[39] Dong L Y, Li Y L, Yin H, et al. The algorithm of link prediction on social network. Mathematical Problems in Engineering, 2013: article ID125123

[40] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neurocomputing, 2003, 15(6): 1373-1396

[41] Park G, Schwartz H A, Eichstaedt J C, et al. Automatic personality assessment through social media language. Journal of Personality & Social Psychology, 2014, 108(6): 934-952

[42] Socher R, Bauer J, Manning C D, Andrew Y N. Parsing with compositional vector grammars//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 455-465

[43] Kazama K, Imada M, Kashiwagi K. Characteristics of information diffusion in blogs, in relation to information source type. Neurocomputing, 2012, 76(1): 84-92

[44] Markovikj D, Gievska S, Kosinski M, Stillwell D. Mining Facebook data for predictive personality modeling//Proceedings of the International AAAI Conference on Web and Social Media. Massachusetts, USA, 2013: 23-26

[45] Costa P T, McCrae R R, Revised N E O. Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)// Professional Manual, Odessa, Florida: Psychological Assessment Resources, 1992: 2-14

[46] Atkinson R L, Richard C A, Edward E S, et al. Hilgard's Introduction to Psychology (13 Edition). Orlando, Florida: Harcourt College Publishers, 2000: 437

[47] Li Le, Zhang Yu-Jin. A survey on algorithms of non-negative matrix factorization. Acta Electronica Sinica, 2008, 36(4): 737-743(in Chinese)
(李乐, 章毓晋. 非负矩阵分解算法综述. 电子学报, 2008, 36(4): 737-743)

[48] Schacter D S, Gilbert D T, Wegner D M. Psychology. 2nd Edition. New York: Worth, 2011: 474-475

[49] Lima A C E S, Castro L N D. Multi-label semi-supervised classification applied to personality Prediction in tweets// Proceedings of the BRICS Congress on Computational Intelligence & 11th Brazilian Congress on Computational Intelligence. Ipojuca, Brazil, 2013: 195-203

[50] Endres D M, Schindelin J E. A new metric for probability distributions. IEEE Transactions on Information Theory, 2003, 49(7): 1858-1860

[51] Kullback S, Leibler R A. On information and sufficiency. The Annals of Mathematical Statistics, 1951, 22(1): 79-86

[52] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788-791

[53] Moore K, McElroy J C. The influence of personality on Facebook usage, wall postings, and regret. Computers in Human Behavior, 2012, 28(1): 267-274

[54] Soto C J, John O P, Gosling S D, Potter J. Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology, 2011, 100(2): 330-348

[55] Soto C J, John O P. Ten facet scales for the Big Five inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. Journal of Research in Personality, 2009, 43(1): 84-90

WANG Meng-Meng, born in 1987, Ph. D. candidate. Her main research interests include database, social network analysis, data mining and machine learning.

ZUO Wan-Li, born in 1957, Ph. D., professor, Ph. D. supervisor. His main research interests include database, social network analysis, data mining and machine learning.

WANG Ying, born in 1981, Ph. D., lecturer. Her main research interests include database, social network analysis, data mining and machine learning.

Wang Xin, born in 1981, Ph. D., lecturer. His main research interests include database, social network analysis, data mining and machine learning.

Background

As a new medium for information dissemination, social networks have become a widely-used and popular facilitator for social interactions. Hence, user's individual behaviors have gradually turned into the key factors in social networks analysis. Besides, although some users posted their desirable images and livesonto social media to achieve self-presentation which reflected some sort of "untrue" self, users' contributions and activities, which can be instantly made available to entire social network, still provide a valuable insight into individual behaviors.

Psychologists believed that user's personality traits are the driving force of user's behaviors, individual differences in personality traits may have an impact on user's online activities, so a better understanding of user's personality traits can bring us a deep understanding of online social networks. For instance, user's personality traits can be used to predict early adoption about Facebook: a person with conscientiousness has sparing use of Facebook, a person with extraversion has long sessions and abundant friendships, and a person with neuroticism has high frequency of sessions. Moreover, user's personality traits may help optimize search results, manifest social influence, distinguish individuals who have some common properties in the crowd. It also plays an important role on relationship outcomes (customer trust, satisfaction and loyalty). To sum up, user's personality traits recognition has an important theoretical significance to mine user's behavior patterns and get user's potential needs under different contexts. Hence, it can be employed as a novel factor in a variety of social researches, analyzing and forecasting user's personality traits through mining data on online social networking sites has become a research focus in the current. Our work on recognizing user's personality traits is motivated by its broad application prospect.

However, some drawbacks can be pointed out in previous work on user's personality traits recognition: (1) most previous works have made an assumption that there had little or no correlations between user's personality traits, however, as a matter of fact, there were significant inter-correlations between user's personality traits which cannot be ignored. (2) Although different features played different roles in recognizing user's personality traits, only a few researchers considered correlations between features and personality traits. Therefore, we propose a multidimensional personality traits recognition model based on weighted nonnegative matrix factorization. Our main contributions are summarized next.

(1) Demonstrate the existence of interdependencies between user's personality traits from a more fine-grained view;

(2) Employing personality traits correlation factor to constrain objective function, we cast the recognizing problem into solutions of nonnegative matrix factorization from network structure, linguistics and emotion dimensions respectively, so as to reduce complexity effectively.

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61300148, 61602057; the Scientific and Technological Break-Through Program of Jilin Province under Grant No. 20130206051GX; the Science and Technology Development Program of Jilin Province under Grant No. 20130522112JH; the Science Foundation for China Postdoctor under Grant No. 2012M510879; the Basic Scientific Research Foundation for the Interdisciplinary Research and Innovation Project of Jilin University under Grant No. 201103129.