基于兴趣偏好的微博用户性别推断研究

宋 巍,刘丽珍,王函石
(首都师范大学信息工程学院,北京 100048)

摘 要: 用户属性,如:性别、年龄等,是计算心理学、个性化搜索、社会化商业推广等研究和应用考察的核心因素.利用用户生成数据自动推断用户属性成为新兴的研究课题.本文提出基于用户兴趣偏好研究微博用户的性别推断问题.考察了用户内容偏好以及关注行为偏好对性别推断的作用.在新浪微博近万名用户的数据集上证明了用户偏好特征的有效性.与传统的语用特征相比,将用户内容偏好与关注偏好相结合能够显著提高推断准确率.关注偏好特征对推断非活跃用户的性别尤其有效.

关键词: 用户隐藏属性;用户性别推断;用户偏好建模;社交媒体
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)10-2522-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.10.034

User Interest Preferences for Gender Inference on Microblog

SONG Wei, LIU Li-zhen, WANG Han-shi
(*College of Information Engineering, Capital Normal University, Beijing* 100048, *China*)

Abstract: User demographic attributes, such as gender and age, are the core factors to be considered for research and applications in computational psychology, personalized search and social commerce marketing. Automatic user latent attribute inference based on user generated data becomes an emerging research topic. This paper proposes a methed for user gender inference on Microblog by exploiting user content preferences and following behaviour preferences. The experiments on a dataset collected from Sina Weibo that consists of nearly 10000 users demonstrate the effectiveness of user preferences features. Comparing with the traditional language usage features, combining user content preferences and user following preferences features can improve the inference accuracy largely. The user following preferences features are especially effective for inferring the gender of inactive users.

Key words: user latent attribute; user gender inference; user preference modeling; social media

1 引言

随着大规模用户生成的内容与行为数据被采集与保存,自动分析用户数据从而深入理解个人和群体的基本信息、挖掘社会心理和行为模式,成为多学科共同关注的重要课题.计算社会学[1]、计算心理学[2]等交叉研究领域应运而生.

在此背景下,对用户信息的深度理解成为其中核心问题.由于涉及隐私,个人用户的基本信息通常无法直接获取.用户隐藏属性推断,即自动推测用户没有显式公开的个人属性,如:性别、年龄等,具有重要意义并将在个性化搜索与推荐[3,4]、心理状况诊断[5]等方面发挥重要作用.

微博已成为人们记录生活,分享与获取信息和彼此互联的最主要平台,提供了丰富的用户语言、行为和社会关系等方面的公开数据.为研究用户隐藏属性推断问题提供了充分的数据准备.

用户隐藏属性推断的主流方法是分析用户文本中体现出的语言特征,如习惯使用的词语类别[6,7],用户使用词语的频次统计[8]等.语言是人类内在心理的外在表现,语言特征毫无疑问是推断用户属性的重要因素.然而,具有不同属性的用户的区别不仅仅体现在语言使用上的偏好,同时也体现在其兴趣爱好、品味等多个方面.这些特征很难通过简单的词类和词频统计精

收稿日期:2015-06-01;修回日期:2015-10-26;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61402304, No. 61303105);北京市自然科学基金(No. 4154065);教育部人文社会科学规划项目(No. 14YJAZH046);北京市教委科研支持项目(No. KM201610028015)

确描述,需要采用更为有效的用户建模方法. 此外,语言使用特征依赖于用户生成文本的规模. 已有研究主要针对活跃用户进行实验,但在社交媒体中存在大量的非活跃用户和新加入用户. 他们未发布足够的文本内容,但依然获取信息并且是潜在商业应用的消费者. 针对这些非活跃用户,能否利用文本数据之外的社交媒体上的关系信息作为补充,从新的维度描述用户特质和改进用户属性推断性能也是有价值的研究问题.

本文从用户兴趣偏好建模这一新角度研究用户属性推断问题并以性别推断为例进行验证. 本文的主要贡献包括:

(1)提出利用用户兴趣偏好建模推断用户隐藏属性的新思路. 将用户内容与关注行为相结合建立用户的内容偏好与关注偏好,构建性别推断的辨别特征. 实验表明用户兴趣偏好特征比传统的语用特征更为有效.

(2)深入分析特征对不同活跃程度用户的推断效果. 实验表明针对发布内容较少的非活跃用户,利用不依赖文本内容的用户关注偏好特征推断更为有效且健壮.

2　相关工作

2.1　用户隐藏属性推断

用户隐藏属性是指用户没有或者无法显式提供的属性,如:性别、年龄、教育层次、消费水平和人格特质(personal traits)等. 有研究表明,社交媒体上的用户不会为了隐藏自己的属性和心理特质蓄意地改变自己的信息和表达方式[9]. 因此,利用用户在社交媒体公开发布的文本和行为数据自动地推断用户的隐藏属性和特质是可行的.

国际上利用社交媒体数据推断多种典型的用户隐藏属性始于对博客(blog)用户的分析[10,11]. 随着微博兴起,使用微博数据预测用户隐藏属性成为热点[8,12,13]. 在 Facebook 等强关系社交网络上存在类似工作[14~16]. 研究者利用 Facebook 用户的好友、分享和群组等信息进行缺失属性补全[17,18]. 此外,有学者从政治立场[14,19]、性取向和宗教信仰[14]、人格[9,20~22]和是否有抑郁症倾向[5,23]等角度对用户进行分类. 近来,Jiwei Li 等将用户属性推断视为信息抽取问题,采取弱指导的方法,利用 Facebook 中的用户属性数据指导 Twitter 用户的属性抽取[24].

国内研究者在相关问题上的工作处于起步阶段. 中科院心理所根据英文的词类词典 LIWC(Linguistic Inquiry and Word Count)[25]构建了面向中文的词类词典 SCLIWC(Simplified Chinese LIWC)[26],并以此为基础进行心理诊断[27]. 部分工作着重挖掘文本中的性别倾向词识别[28]以及基于词汇特征的微博用户性别识别[29].

2.2　用户性别推断

性别是用户最主要的基本属性之一. 本文主要以性别推断作为主要研究对象. 用户性别推断的典型设置是将其视为有监督的二元分类问题. 在标注好用户性别的用户数据集上进行训练,学习得到分类模型用于推断[8,12,15,30~33]. 其关键在于有效特征的抽取. 下面简述用户性别推断的已有方法并分析它们的优缺点.

2.2.1　基于词类词典的方法

心理学上认为不同属性的人在用词、语气、风格等使用语言的方式上具有一定的差异. 通过对语言中不同类型词汇的统计信息推断用户属性是一种比较传统的方法. 在英文上,已有工作主要利用著名的心理语言分析工具 LIWC.

LIWC 是美国德克萨斯大学奥斯丁分校教授 James W Pennebaker 主导研究的一套语言分析工具,其核心为一部人工构建的词语词典[25]. 词典将词语划分到约 80 个词类中,涵盖了不同的语言维度. 基于 LIWC 的心理学研究分析出不同属性人群具有不同的语言风格和习惯用法. 相关研究发现男性更多使用冠词、介词,以及复杂、正式、具有专业性的语言,而女性更偏向于社会交往相关的语言,使用更多的代词等. 年龄大的人更多表达正面情感而较少表现出负面情感,较少地使用主观词以及否定词等.

2.2.2　基于词语统计特征的方法

计算机科学领域研究者更愿意采用直接的开放式语言特征,即通过对用户的文本信息进行处理,使用词或词组作为特征,构建统计分类模型进行推断[8,15]. 可作为文本信息的内容包括用户的昵称、自我描述以及发表的微博. 文本特征的选择通常基于传统的文本分类方法,选择具有高区分度的词和短语等. 用户在社交媒体中用文字表达思想时独有的、非正式的社会化口头表达方式,典型的如文字表情符、图形表情符和表示惊异的词语通常也会保留作为特征.

2.2.3　基于局部社交关系和交互特征的方法

Zamal[31]等利用社交网络具有同质性(homophily)的特点利用用户好友信息辅助属性预测. 然而类似研究主要使用简单的社会关系相关的统计特征,如:关注者和被关注者个数,以及交互统计特征,如:转发频率和发布频率等. 但这些特征在性别和年龄等属性上的分布并不具有明显的区分性[12].

2.2.4　已有方法的局限性

基于词典的方法具有以下局限性:(1)词典具有语言相关的特点,英文之外其他语言资源的建设相对落后,此类方法不易于快速扩展到其他语言. (2)词典中词语覆盖范围较小,社交媒体上大量涌现的新词及社会化语言用法无法被有效覆盖,影响了基于词典的方

法的适用范围.

根据用户发表内容分析用户的语言使用特征推断用户属性是目前已有工作中最为有效的方法. 然而,此类方法主要面向具有丰富内容资源的活跃用户(如要求评测用户至少发表千条以上微博[8]). 在微博等社会媒体平台中,存在大量非活跃用户,他们仅具有有限的内容数据,从而面临数据稀疏问题. 因此,有必要对用户的内容进行进一步的抽象,缓解数据稀疏. 此外,已有工作没有充分利用社交媒体的交互特征. 以关注行为为例,已有方法仅利用基本的关注对象数目作为特征,而没有深入分析关注对象群体的特点和联系.

3 基于用户兴趣偏好建模的方法

用户兴趣建模是个性化搜索与推荐的核心内容. 个性化搜索与信息过滤主要针对用户的查询、文档以及上下文信息使用关键词、分类、潜在主题或子空间对用户进行建模[34~36]. 心理学有研究表明心理特质影响人们在兴趣和态度上的选择[37]. 受此启发,本文尝试结合用户兴趣建模技术构建有效特征支持用户隐藏属性推断.

应用概率主题模型 Latent Dirichlet Allocation (LDA)[38]于大规模无标注的微博用户数据,分别训练内容主题模型(Content Topic Models,CTM)和关注主题模型(Followee Topic Models,FTM). 以此为基础,对用户的主题兴趣偏好与关注兴趣偏好进行建模作为特征,改进用户性别推断.

接下来首先简要介绍 LDA 模型,而后分别介绍使用 LDA 对用户微博内容与关注行为进行建模并应用于用户性别推断.

3.1 LDA 模型

LDA 模型可视为层次贝叶斯模型. 假设一篇文档是由多个潜在主题混合组成,每个主题为词汇表上的多项式分布. LDA 的图模型表示如图 1 所示. 每一篇文档 d 表示为 N 个词的序列 $\boldsymbol{w}=\{w_1,w_2,\cdots,w_N\}$,则包含 M 篇文档的集合 D 表示为 $D=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_M\}$,文档集合 D 由主题数为 T 的 LDA 模型生成的过程可描述为:

(1)对每个主题 k,根据狄利柯雷(Dirichlet)分布生成该主题在词汇表 V 上多项式分布:$\varphi_k \sim Dir(\beta)$;

(2)对每篇文档 $d\in D$ 根据狄利克雷分布生成其在主题上的多项式分布 $\theta_d \sim Dir(\alpha)$;

(3) 对文档 d 中的每一个词:

ⅰ. 根据分布 θ_d 生成主题 $z\sim Multi(\theta_d)$;

ⅱ. 根据分布 φ_z 生成 $w\sim Multi(\varphi_z)$.

其中 α 和 β 为狄利克雷分布的超参数. 生成过程描述了如何由模型生成数据. 模型的参数学习则可视为生成过程的逆过程,即根据真实数据的分布学习参数模型.

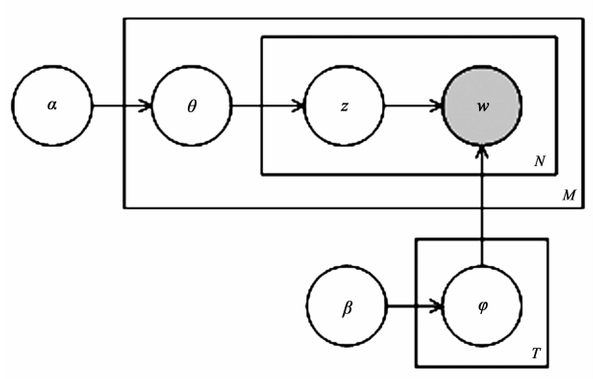

图1 主题模型的概率图表示

模型参数可采用吉布斯抽样等方法习得[39]. 训练好的主题模型可对新的文档样本进行推断,得到文档的主题概率分布.

采用 LDA 对用户进行建模的原因如下:(1)用户兴趣很难用固定的类别本体进行描述. 概率主题模型是一种无监督的数据挖掘方法可自动发现数据中的隐藏结构. 以往工作证明,LDA 是对社交媒体用户进行兴趣建模的有效手段[36]. (2)本文利用用户兴趣模型作为特征推断用户隐藏属性,需要将训练样本与测试样本映射到同一特征空间. LDA 是一种完全的生成模型,能够对新文本进行有效的推断有利于对新增用户进行处理. (3)微博用户的文本内容信息和社交关系信息均可使用主题模型进行建模.

3.2 训练主题模型

在微博数据上训练内容主题模型与关注主题模型. LDA 模型是无监督的数据挖掘算法,因此训练 LDA 模型只需从微博平台获取的一定规模的用户的微博及其关注信息. 关注是微博用户独特的一种行为. 用户可以关注任何感兴趣的其他用户. 用户 A 关注用户 B,称用户 A 为关注者(follower),称用户 B 为关注对象(followee). 整个微博平台形成一个非对称社交网络. 显然,用户关注行为也表现出用户的兴趣偏好. 然而,这一行为并没有被充分利用进行用户建模以及性别推断.

将训练主题模型使用微博数据表示为 $C=\{C_1,C_2,\ldots,C_U\}$,U 为数据集中用户的数目. 用户 u_i 的数据表示为 $C_i=(S_i,F_i)$,$C_i\in C$,其中 S_i 表示用户 u_i 发表过的所有微博拼接得到的伪文档,F_i 表示其关注对象的列表. 关注对象集合可表示为 $E=\cup_{i=1}^{U}F_i$.

利用微博文本内容和关注对象列表数据可分别训练内容主题模型和关注主题模型. 训练主题模型并不需要知道数据中用户的属性(如:性别)的取值. 同时,训练主题模型的数据不必包括待推断的用户数据,新用户的主题分布可由训练好的主题模型推断得到.

3.2.1　训练内容主题模型

内容主题模型 CTM 用于挖掘大规模微博文本中涵盖的主题. 将所有用户的微博伪文档聚合形成伪文档集合 $S=\{S_1,\ldots,S_U\}$. 假设每一篇文档由 T 个主题生成,使用 LDA 模型在 S 上训练主题模型. 训练得到的主题模型包括 T 个语言模型,每个语言模型为词汇表 V 上的多项式分布.

3.2.2　训练关注主题模型

期望将微博平台上关注对象集合 E 划分为若干个不同类型的群体,从而能够描述不同用户的关注对象在分布上的异同. 将所有用户的关注对象列表聚合到一起形成关注对象列表集合 $F=\{F_1,\ldots,F_U\}$,将每一个关注列表 F_i 视为一篇文档,将其中每一个关注对象 $e\in E$ 类比为一个词. 假设每个用户的关注列表由 G 个类型的关注对象构成,则可在 F 上训练得到包括 G 个主题的关注主题模型 FTM. FTM 由 G 个语言模型构成,每个语言模型是在关注对象集合 E 上的多项式分布.

3.3　用户的兴趣偏好表示

基于已训练好的内容主题模型 CTM 和关注主题模型 FTM,可以对任一用户 u 的兴趣进行表示. 设 S_u 为用户 u 发布的微博拼接而成的伪文档,则可利用 CTM 对 S_u 进行推断,获得 S_u 在 T 个主题上的概率分布向量 θ_u. 将 θ_u 作为用户的内容兴趣偏好表示. 类似地,设 F_u 为用户的关注对象列表,使用 FTM 对其进行推断,可获得 F_u 在 G 个被关注对象主题上的概率分布向量 δ_u,将 δ_u 作为用户的关注兴趣偏好表示. θ_u 和 δ_u 分别表达了用户对不同主题的内容及不同类型关注对象群体的偏好.

3.4　性别推断

将性别推断视为有监督的二元分类问题. 图 2 给出了系统的处理流程. CTM 与 FTM 模型需要预先在大规模无标注的微博用户数据上训练完成. 在标注好用户性别的训练数据集上进行训练学习到性别推断的分类模型.

特征抽取阶段为每名用户建立一个特征向量. 该特征向量包含多种类型的特征. 用户兴趣偏好特征的构建需要 CTM 与 FTM 模型. 使用训练好的 CTM 和 FTM 模型推断出用户的兴趣偏好. 将用户内容兴趣偏好表示向量 θ_u 和关注兴趣偏好表示向量 δ_u 拼接在一起形成维度为 $T+G$ 的向量. 该向量与其它类型特征的特征向量进一步拼接,形成完整的表征用户的特征向量.

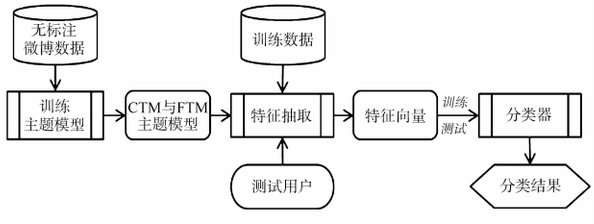

图2　系统的基本处理流程

对于待推断性别的测试用户,使用相同的 CTM 和 FTM 模型推断其兴趣偏好表示构建用户的特征向量. 使用在训练数据上学习得到的分类器将该特征向量作为输入,输出性别推断的结果.

4　实验设置与结果分析

4.1　研究问题

本文希望回答以下研究问题:

(1)通过用户兴趣建模得到的用户兴趣偏好特征能否提高性别推断的性能?

(2)针对活跃程度不同的用户,用户兴趣偏好特征与已有特征相比是否具有更好的健壮性?

4.2　实验数据采集与评价

从中国最大的微博平台新浪微博采集实验数据. 为避免垃圾用户对实验的影响,以经过官方认证的微博账号作为种子,按照关注关系进行了 2 层扩展,即首先获取种子账户的关注对象的数据,并进一步获取新扩展的账户的关注对象. 最终采集约 5000 万微博账户,每个账户获取的信息包括:最近发布的 1000 条微博以及其关注对象列表. 从该数据集中随机选取了 10 万名用户作为实验数据,其中 9 万名用户用于训练内容主题模型 CTM 和关注主题模型 FTM. 对 1 万名用户进行性别标注,供训练分类器和测试实验效果使用. 两名标注者根据用户名称、描述、微博内容以及关注对象列表进行性别标注. 根据 Kappa 值[40]度量,标注的一致性为 92%.

需要指出的是,新浪微博实际上要求用户在注册时添加性别信息,因此获取的用户数据中已经包含性别取值. 要求标注者对数据进一步进行标注一方面原因在于部分用户可能为了缩短注册花费的时间而随意填写信息,另外一方面也考察人类直接观察进行判断的准确率,反映这一问题的难度. 从一致性结果来看,微博用户性别推断并不是一项容易的工作. 对于有分歧的标注,标注者讨论决定最终的标注结果. 无法达成共识的用户将被移除. 最终获得了 9076 名具有性别标注的用户. 数据的基本统计列在表 1 中.

表 1　测试数据集的基本统计

统计项目	数量
男性	4287
女性	4789
平均发表微博数	310
平均关注对象数	197

从表 1 可见,测试数据中男女比例约为 9:10,女性用户略多. 尽管试图获取每个用户的最近 1000 条微博,

但实际中大比例用户发布的微博数都未能达到1000.每个用户平均的发布微博数大于平均的关注对象数目.

采取准确率(Accuracy)来衡量自动性别推断系统的表现,其计算方法为正确判断的样本数量与全部样本数量的比值.测试过程中,将9076名用户组成的数据集上采取5折交叉验证的方法进行测试,采用准确率的平均值来评价系统的性能.

4.3 对比系统

(1)词类特征(Word Category):词类特征依赖于词类词典.采取简体中文LIWC(SCLIWC)词类词典[26],该词典根据英文版LIWC针对中文进行了翻译与扩充.统计用户发布内容中被包含在SCLIWC不同词类的词语的比例作为分类特征.

(2)统计词特征(Ngram):根据已有研究的结论,N-元词(Ngram)是最为有效的统计特征.针对男性和女性,分别选取前3000个区分性最强的一元词和二元词,并将它们合并作为NGram特征的维度,每一维度的取值为用户内容中包含该Ngram的频次.处理过程中保留了表情符等社会化词汇,因为它们也是表达性别的一种信号.度量Ngram的区分性的方法基于计算它们的类互信息.实验结果显示选择区分性强的Ngram比使用所有Ngram而不考虑其区分性的效果更好.由于特征更加紧凑,训练的效率更高.

(3)Rao et al[12].该方法综合使用了Ngram特征以及简单的用户社交统计信息,如:好友数、关注数等.因此这种方法可视为利用了局部社交网络信息.与其不同,本文对关注对象进行分群可视为对全局的用户行为进行建模.

4.4 分类器与参数设定

采用LibLinear分类器[40]进行推断.FTM以及CTM的主题数均设为200,参数学习使用吉布斯抽样方法,迭代次数设为100.

对于所有对比系统,在交叉验证过程中在训练数据上(整个数据的80%)采用4折交叉验证对参数进行调整,选择最佳参数在整个训练语料上训练模型,使用该模型在测试数据上(整个数据的20%)进行测试.

4.5 实验结果及分析

4.5.1 整体表现

表2给出了不同类型特征以及特征组合的准确率.从中可以看到,词类特征(Word Category)表现最弱,获得了65.60%的准确率.与之相比,统计词特征(Ngram)表现更为优异达到74.09%的准确率.基于用户内容主题模型(CTM)与用户关注主题模型(FTM)分别获得了75.45%和74.24%的准确率,其中CTM是最为有效的单一类型特征.实验结果说明,词类特征对于性别预测过于粗略而无法取得令人满意的效果.Ngram、CTM与FTM的表现相当.这一方面印证了前人工作的结论,Ngram特征是性别推断的重要特征,不同性别的用户倾向于使用不同的词,另一方面也说明经过降维处理的用户偏好特征能够起到正面作用:CTM比Ngram表现更好.可能的原因是:Ngram面临的数据稀疏问题得到缓解,此外主题模型实质上相当于进行了特征选择,主题区分性强的词语在用户兴趣模型建立过程中起到了更大的作用.

表2 采用不同类型特征和特征组合的表现

特征	准确率
Word Category	65.60
Ngram	74.09
CTM	75.45
FTM	74.24
Ngram + CTM	76.29
Ngram + FTM	76.06
CTM + FTM	80.16
Ngram + CTM + FTM	78.79
Rao et al[12]	75.68

将2类用户兴趣特征结合起来(CTM + FTM)取得了最好的效果,准确率达到了80.16%.Ngram分别与CTM和FTM结合时,准确率均有提升.但将三类特征全部结合起来时,表现却弱于CTM + FTM.其原因可能是CTM已经能够较好地替代Ngram特征,而使用Ngram特征可能引入更多的噪声,导致性能下降.

本文提出的方法同样超过了Rao等[12]的表现.这说明用户对不同关注对象群体的关注偏好能够更好地表达用户关注兴趣.而简单的用户关注统计数字则难以刻画.

4.5.2 在不同活跃程度用户上的表现

分析不同特征及特征组合针对活跃程度不同的用户时的表现.目的在于分析不同类型特征的健壮性,尤其是针对文本内容不够丰富的非活跃用户的表现.为此,将测试用户根据其发布微博的数量分为5组.表3给出了5组测试用户所处的不同区间及其相关统计.

表3 按照活跃程度进行划分的5个用户组相关统计

组	微博数	用户数	平均关注数
G1	[10,50)	1440	113
G2	[50,200)	2130	167
G3	[200,400)	2069	206
G4	[400,600)	1724	271
G5	[600,)	1713	422

从表3中可以看到大致有18%的用户的发表微博

数量大于600,而发表微博数量在10到200之间的用户大约占据用户总数的40%.这说明社交媒体中有相当一部分非活跃用户,其比例甚至可能远超过活跃用户.发表微博数量越多的用户关注的用户数也更多,两者具有一种近似的线性关系.然而,发表微博数目小于50的非活跃用户仍然保持一定规模的关注对象.

对5组测试用户,分别将数据进一步随机均分为5个部分.在训练时,从每一组测试用户中随机选取4个部分,并将来自于5组的数据合并用于训练分类模型,学习到的模型分别对每一组余下的1份数据进行测试.这样处理的原因是在实际应用的时候,仅维持一个统一的模型更加便于系统进行维护,因此模型对不同特点的用户(如:活跃用户与非活跃用户)进行推断时的健壮性尤为重要.

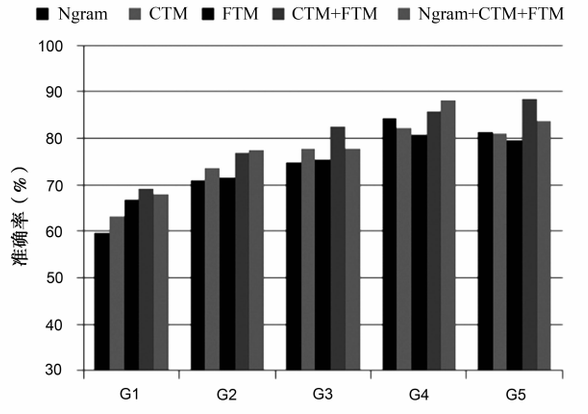

图3 针对具有不同数量微博的用户各个方法的表现

图3给出了不同的特征和特征组合在具有不同规模微博数量的用户群组上的表现.从中可以看到如下趋势:(1)用户发布内容越活跃,对其进行性别推断的准确率越高.所有的特征和特征组合都体现出这一特点.这说明丰富的内容数据更容易构建足够多的特征以避免特征稀疏问题.(2)内容相关的特征(WordCategory,Ngram,CTM)高度依赖于用户内容的规模.最明显的体现在Ngram特征,当用户发表内容足够多时(G5),其准确率超过80%,是表现最好的单独类型特征.然而当用户内容较少时,Ngram性能不如用户兴趣偏好特征.(3)对于非活跃用户,用户兴趣偏好特征CTM和FTM表现得更为健壮.例如在发布微博数小于50的用户组上,使用CTM的特征表现优于Ngram,证明对文本内容的抽象能够改善数据稀疏问题.FTM表现最好,说明对于内容较少的用户,其关注偏好兴趣能够更准确地反映其隐藏属性.(4)将用户兴趣偏好特征与其他特征相融合时,能够获得比单独使用时更好的表现.通过分析可见,用户兴趣偏好特征对于活跃用户与非活跃用户的隐藏属性推断均是有效的.对于发布内容较少的非活跃用户,使用用户兴趣偏好特征进行推断可获得更高的准确率,具有更好的健壮性.

5 结束语

本文针对中文微博用户的性别推断问题进行研究,提出了利用用户兴趣偏好建模改进推断性能的新思路.着重考察了用户的内容兴趣与关注兴趣偏好,详细比较了这些新特征与传统特征的表现并分析了针对不同活跃程度的用户不同类型特征的健壮性.实验表明,用户兴趣特征是推断用户性别的有效特征,是对传统的基于词语粒度文本分析的有力补充.特别是针对微博上数量众多的非活跃用户,用户兴趣偏好特征尤其是用户关注兴趣偏好特征能够较好地缓解数据稀疏问题,提高推断的准确率.

在未来,我们试图结合社会学与心理学中的相关理论,继续挖掘有效的用户行为特征与高级语言特征以构建更为准确的用户兴趣模型,进一步提高推断性能.

参考文献

[1] Lazer David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al. Life in the network: the coming age of computational social science[J]. Science, 2009, 323(5915): 721.

[2] Sun R. The Cambridge Handbook of Computational Psychology[M]. Cambridge University Press, 2008.

[3] Ingmar W, Carlos C. The demographics of web search[A]. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York: ACM, 2010. 523 - 530.

[4] Duhigg C. The power of habit: why we do what we do in life and business[J]. Random House LLC, 2012, 34(10).

[5] De Choudhury M, et al. Predicting depression via social media[A]. Proceedings of AAAI Conference on Weblogs and Social Media[C]. Palo Alto, California: AAAI Press, 2013. 128 - 137.

[6] Newman ML, et al. Gender differences in language use: An analysis of 14,000 text samples[J]. Discourse Processes, 2008, 45(3): 211 - 236.

[7] Pennebaker JW, Stone LD. Words of wisdom: language use over the life span[J]. Journal of Personality and Social Psychology, 2003, 85(2): 291 - 301.

[8] Burger JD, et al. Discriminating gender on Twitter[A]. Proceedings of Empirical Methods in Natural Language Processing[C]. Stroudsburg, PA, USA: ACL, 2011. 1301 - 1309.

[9] Gosling SD, Gaddis S, Vazire S. Personality impressions based on facebook profiles[A]. Proceedings of AAAI

Conference on Weblogs and Social Media[C]. Palo Alto, California: AAAI Press, 2007. 1 – 4.

[10] Argamon, et al. Mining the Blogosphere: Age, gender and the varieties of self-expression[J]. First Monday, 2007, 12(9).

[11] Burger JD, Henderson JC. An exploration of observable features related to blogger age[A]. Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs[C]. Palo Alto, California: AAAI Press, 2006. 15 – 20.

[12] Rao D, et al. Classifying latent user attributes in twitter[A]. Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents[C]. New York: ACM, 2010. 37 – 44.

[13] Dong N, et al. How old do you think i am?: a study of language and age in twitter[A]. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media[C]. Palo Alto, California: AAAI Press, 2013. 439 – 448.

[14] Kosinski M, Stillwell D, Graepe T. Private traits and attributes are predictable from digital records of human behavior[J]. The National Academy of Sciences, 2013, 110: 5802 – 5805.

[15] Schwartz H A, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach[J]. PloS One, 2013, 8(9).

[16] Tang C, et al. What's in a name: a study of names, gender inference, and gender behavior in facebook[J]. Database Systems for Advanced Applications, 2011, 344 – 356.

[17] Elena Z, Lise G. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles[A]. Proceedings of the 18th International Conference on World Wide Web[C]. New York: ACM, 2009. 531 – 540.

[18] Alan M, et al. You are who you know: inferring user profiles in online social networks[A]. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining[C]. New York: ACM, 2010. 251 – 260.

[19] Pennacchiotti M, Popescu A-M. Democrats, republicans and starbucks afficionados: user classification in twitter[A]. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery in Data Mining[C]. New York: ACM, 2011. 430 – 438.

[20] Golbeck, et al. Predicting personality from twitter[A]. Proceedings of the IEEE Third International Conference on Social Computing[C]. IEEE, 2011. 149 – 156.

[21] Yoram, B, et al. Personality and patterns of Facebook usage[A]. Proceedings of the 3rd Annual ACM Web Science Conference[C]. New York: ACM, 2012. 24 – 32.

[22] Daniele Q, et al. Our Twitter profiles, our selves: Predicting personality with Twitter[A]. Proceedings of the IEEE Third International Conference on Social Computing[C]. IEEE, 2011. 180 – 185.

[23] De Choudhury M, et al. Characterizing and predicting postpartum depression from shared facebook data[A]. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing[C]. New York: ACM, 2014. 626 – 638.

[24] Li Jiwei, Ritter A, Hovy E. Weakly supervised user profile extraction from Twitter[A]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics[C]. Stroudsburg, PA, USA: ACL, 2014. 165 – 174.

[25] Tausczik, YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods[J]. Journal of Language and Social Psychology, 2010, 29(1): 24 – 54.

[26] Gao, R, et al. Developing simplified Chinese psychological linguistic analysis dictionary for microblog[J]. Brain and Health Informatics, 2013, 359 – 368.

[27] Huijie L, et al. User-level psychological stress detection from social media using deep neural network[A]. Proceedings of ACM International Conference on Multimedia[C]. New York: ACM, 2014. 507 – 516.

[28] 唐琴, 林鸿飞. 文本中人物性别识别研究[J]. 中文信息学报, 2010, 2: 46 – 51.
Tang Qin, Lin H. Research on gender recognition for character in text[J]. Journal of Chinese Information Processing, 2010, 24(2): 46 – 51. (in Chinese)

[29] 王晶晶, 李寿山, 黄磊. 中文微博用户性别分类方法研究[J]. 中文信息处理, 2014, 28(6): 150 – 155.
Wang Jingjing, Li Shoushan, Huang Lei. User gender classification in Chinese Microblog[J]. Journal of Chinese Information Processing, 2010, 28(6): 150 – 155. (in Chinese)

[30] Morgane C, Sonderegger M, Ruths D. Gender inference of twitter users in non-English contexts[A]. Proceedings of the Conference on Empirical Methods in Natural Language Processing[C]. Stroudsburg, PA, USA: ACL, 2013. 1136 – 1145.

[31] Zamal A, et al. Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors[A]. Proceedings of AAAI Conference on Weblogs and Social Media[C]. Palo Alto, California: AAAI Press, 2012. 387 – 390.

[32] Mislove A, et al. Understanding the demographics of twitter users[A]. Proceedings of AAAI Conference on Weblogs and Social Media[C]. Palo Alto, California:

AAAI Press,2011. 554 – 557.
[33] Liu W,Ruths D. What's in a name? using first names as features for gender inference in Twitter[A]. Proceedings of the 2013 AAAI Spring Symposium[C]. Palo Alto,California:AAAI Press,2013. 10 – 16.
[34] Ghorab MR,et al. Personalised information retrieval:survey and classification[J]. User Modeling and User-Adapted Interaction,2013,4(23):381 – 443.
[35] Bobadilla,et al. Recommender systems survey[J]. Knowledge-Based Systems,2013,46:109 – 132.
[36] Liangjie Hong,Brian D Davison. Empirical study of topic modeling in twitter[A]. Proceedings of the First Workshop on Social Media Analytics[C]. New York:ACM,2010. 80 – 88.
[37] Anderson WT,Golden LL. Lifestyle and psychographics:a critical review and recommendation[J]. Advances in Consumer Research,1984,11(1).
[38] Blei,DM,Ng AY,Jordan MI. Latent dirichlet allocation [J]. Journal of Machine Learning Research,2003,3:993 – 1022.
[39] Griffiths,TL,Steyvers M. Finding scientific topics[J]. National Academy of Sciences of the United States of America,2004,101:5228 – 5235.
[40] Jacob Cohen et al. A coef? cient of agreement for nominal scales[J]. Educational and Psychological Measurement,1960,20(1):37 – 46.
[41] Fan R. -E. ,et al. LIBLINEAR:A library for large linear classification[J]. Journal of Machine Learning Research,2008,9:1871 – 1874.

作者简介

宋　巍　男,1983 年 1 月出生,黑龙江哈尔滨人. 讲师、中国计算机学会会员、中文信息学会会员. 2006 年、2008 年和 2013 年在哈尔滨工业大学获得学士、工学硕士和工学博士学位. 现在首都师范大学信息工程学院工作,主要从事社会计算、自然语言处理和信息检索有关研究.
E-mail:wsong@ cnu. edu. cn

刘丽珍　女. 1966 年 7 月出生,山西太原人. 教授、中国人工智能学会教育工作委员会副秘书长,北京市人工智能学会理事,中国计算机学会高级会员. 1986 年、1994 年、2003 年分别在山西大学、西北大学和北京理工大学获工学学士、工学硕士和工学博士学位. 现在首都师范大学信息工程学院工作,主要从事数据挖掘、社会计算、信息检索和自然语言处理等方面的研究工作.
E-mail:liz _ liu7480@ cnu. edu. cn