软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(Suppl.(2)):145−154

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于社交媒体的关联性用户属性推断*

项连城[1,2], 方 全[1,2], 桑基韬[1,2], 徐常胜[1,2], 路冬媛[3]

[1](模式识别国家重点实验室(中国科学院 自动化研究所),北京 100190)

[2](China-Singapore Institute of Digital Media, Singapore 119615)

[3](National University of Singapore, Singapore 119615)

通讯作者: 路冬媛, E-mail: dongyuanlu@gmail.com

摘 要: 挖掘用户属性对用户建模、用户检索和个性化服务等具有十分重要的意义.已有的相关研究工作都是单独挖掘各种属性,而且忽略了各属性之间的相关关系.提出一种基于超图学习的用户属性推断的方法.在超图中,顶点表示社会媒体中的用户,超边表示用户产生的内容相似性与属性之间的关系.在建好的超图模型上,把用户属性挖掘形式化成一个正则化的标签相似传播问题,可以有效推断得到用户的各种属性.利用从 Google+上收集的标记过全部属性的数据集进行了大量的实验,其结果表明了该方法在用户属性挖掘中的有效性.

关键词: 超图;用户属性挖掘;属性关系

中文引用格式: 项连城,方全,桑基韬,徐常胜,路冬媛.基于社交媒体的关联性用户属性推断.软件学报,2015,26(Suppl.(2)):145−154. http://www.jos.org.cn/1000-9825/15025.htm

英文引用格式: Xiang LC, Fang Q, Sang JT, Xu CS, Lu DY. Exploiting social media information for relational user attribute inference. Ruan Jian Xue Bao/Journal of Software, 2015,26(Suppl.(2)):145−154 (in Chinese). http://www.jos.org.cn/1000-9825/15025.htm

Exploiting Social Media Information for Relational User Attribute Inference

XIANG Lian-Cheng[1,2], FANG Quan[1,2], SANG Ji-Tao[1,2], XU Chang-Sheng[1,2], LU Dong-Yuan[3]

[1](National Laboratory of Pattern Recognition (Institute of Automation, the Chinese of Academy Sciences), Beijing 100190, China)

[2](China-Singapore Institute of Digital Media, Singapore 119615)

[3](National University of Singapore, Singapore 119615)

Abstract: Inferring user attributes is important for user profiling, retrieval, and personalization. Most existing work infers user attribute independently and ignores the relations between attributes. In this work, a new method is proposed to infer user attributes via hypergraph learning. In the hypergragh, each vertex represents a user in the social media, and the hyperedges are used to capture the similarity relations of the user generated content and the relations between attributes. The user attributes inference is formalized into a regularization label similar propagation problem in the constructed hypergraph, which can effectively infer the users' various attributes. Extensive experiments conducted on a collected dataset from Google+ with full attribute annotations demonstrate the effectiveness of the proposed approach in user attribute inference.

Key words: hypergraph; user attribute inference; relations between attribute

随着各种终端设备的发展普及和社会媒体的兴盛,用户可以随时随地上传、分享个人活动数据到在线媒体网站上,这使得用户产生的内容呈现爆炸式地增长.这些海量的由用户在社会网络中产生的行为和数据给我们带来了许多的额外信息.它们反映了各种各样的用户个人信息,包括了基本信息(如:年龄、性别、婚姻状况),兴

* 基金项目: 国家自然科学基金(61225009); 国家重点基础研究发展计划(973)(2012CB316304); 北京市自然科学基金(4131004);新加坡国家研究基金

收稿时间: 2014-06-20; 定稿时间: 2014-08-20

趣爱好(如:政治、科技、环境、运动),职业信息(如:研究院、学生、软件工程师、音乐家),情感倾向(如:积极、消极)等.我们将这些个人信息统称为用户属性.推断用户属性具有十分重要的应用价值,例如用户建模、信息检索、个性化定制和推荐等.

目前,很多的在线社交网站(例如:Facebook、新浪微博、豆瓣、QQ 等)并不能得到准确、完整的用户属性.首先,用户会填写一些简单的信息,例如性别、年龄等,但是很少会填写复杂的信息,比如说兴趣爱好等.其次,很多用户会考虑到隐私问题,不愿意向社交网站提供一些私人信息.为了解决用户属性的稀疏性的问题,一些研究工作[1−8]通过利用用户产生的网络数据来推断用户属性,例如性别[1−4]、职业[5]、政治偏向[9]、情感倾向[6]等.这些研究工作表明了用户网络行为活动对于预测用户属性的有效性.

但是现有的研究工作主要基于文本内容特征,没有考虑用户的视觉内容特征,并且都是针对单个属性进行挖掘的.即使有文献提到了多个属性,但也都是各个属性单独进行,未考虑属性之间的相应关系.实际上,一些用户属性之间存在着一定的关联性,例如一名 10 岁的女性一般不可能是已婚的,而且很可能是一个学生.图 1 显示了年龄、性别,婚姻状况和职业这 4 个用户属性之间的关系.这个统计是根据 1.05 亿 Google+用户的数据进行的.在左图中,我们可以看到小于 24 岁的用户更可能是单身,且是学生.在右图中,我们可以看到一些职业和性别有较大的相关性.因此,我们认为用户属性的关系是重要的特征,可以有效地帮助用户属性推断.

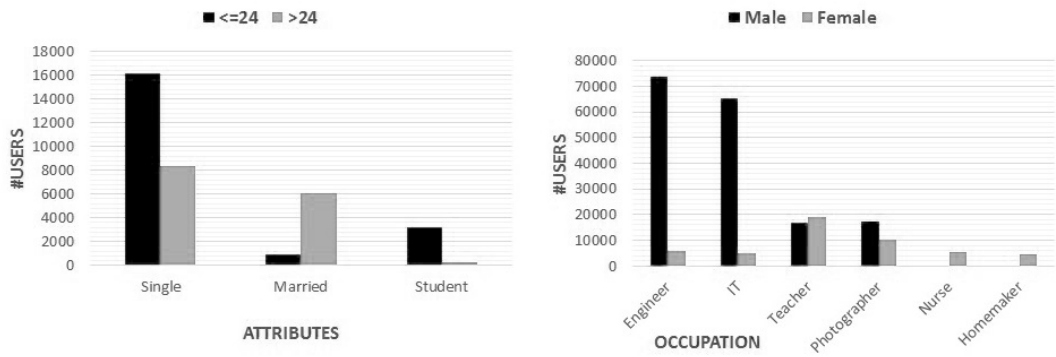


图 1 Google+数据中用户属性共现统计图

在本文中,我们提出关联性用户属性推断方法.利用超图可以有效地对用户及用户产生的丰富在线数据与各属性之间的关系进行建模,从而有效地提高用户属性推断的准确率.我们从用户产生的文本视觉内容数据提取多种丰富的特征.在超图中每个顶点对应一个用户,每条超边根据一种特征或属性的相似性来连接某些顶点.我们根据用户之间的相似度来定义超边的权重.最后根据超图中学习到的得分推断出用户属性.利用 Google+上的用户数据,以性别、年龄、婚姻关系、职业、兴趣爱好和情感倾向这 6 个用户属性为例,进行实验.实验结果表明了我们提出方法在用户属性挖掘中的优越性.

1 问题定义

在这里,我们首先定义研究的 6 种用户属性,然后形式化阐述用户属性挖掘问题.

1.1 用户属性的设定

在本研究中,我们考虑 6 种用户属性,包括 3 种档案属性——年龄、性别、婚姻状况,3 种个性化属性——职业、兴趣爱好和情感倾向.属性的值是根据关于 Google+数据的综合研究和一些之前关于用户属性挖掘的研究工作[4,5,7]进行定义的.表 1 解释了各个用户属性值的含义.6 种用户属性的定义如下:

性别:性别是二值的属性.我们用性别来区别用户是男性还是女性.

年龄:年龄是有具体数量的属性.考虑到用户真实年龄数据的普遍缺乏,在社会网络中进行准确年龄的推断是不可能的.我们对 Google+用户进行了详细的调查和观察.总的来说,Google+用户可以被分为年轻和年长的两组.因此我们将所用用户分成两个人口统计学的分组:30 岁以下的用户(年轻)和 30 岁以上的用户(年老).这种二

维分类方法虽然简单,但是在用户的建模中非常合理和有效.在文献[7]中也采用了相同的定义方法.

婚姻状况:在 Google+平台上,用户的情感状况是复杂多样的,比如说:单身,已婚,热恋等.为了更加简洁,我们将用户分为两组:未婚和已婚.

职业:根据在 Google+用户页面上的职业功能研究和文献[12]中的相关工作,职业被定义成 15 个值,如:IT 工作者,演员,摄影师等.

兴趣爱好:兴趣爱好涉及到了用户发布信息的最喜爱的话题.根据我们对收集到的 Google+数据的分析,我们定义了 12 种兴趣话题来涵盖如此之大的兴趣范畴.因为每个用户可能有多个兴趣,所以兴趣值是十二维向量,我们将兴趣推断看成是二分类问题.

情感倾向:用户发布的信息可以反映出用户具体的情感状态.例如,发布了许多有趣和快乐信息的用户很有可能是一个积极的人,同时,发布了很多消极内容则反映了用户具有消极倾向.情感倾向被用来根据用户发布信息来描述其情感极性.我们将 3 种情感倾向分别定义为:积极,消极和中性.

表 1 用户属性定义表

属性名称	属性值
性别	1-男性;2-女性
年龄	1-年轻(≤30);2-年长(>30)
婚姻状况	1-未婚;2-已婚
职业	1-学生;2-IT 工作者,软件工程师;3-演员,歌手,模特,主持人;4-作家,记者,编辑,评论员;5-政治家;6-运动员;7-商人,经济师,企业家,市场策略顾问,资本家;8-科学家,研究员,专家;9-摄影师,旅行者;10-医生,药剂师,美容师;11-厨师,美食家;12-工程师,专业人员,设计师;13-老师;14-艺术家,宗教人士;15-其他
兴趣爱好	1-科技,信息,网络;2-新闻,政治,军事,社会;3-经济,企业管理战略;4-娱乐,音乐,电影,时尚;5-摄影,旅游;6-美食;7-日常事务,情趣,玩笑,个人物品;8-运动,锻炼,健身;9-思考,宗教文化,文学艺术;10-健康,医疗护理,化妆;11-科学,知识;12-其他
情感倾向	1-积极(极好的,高兴的,得意的,兴奋的,快活的,欢呼的,狂喜的);2-消极(生气的,坏的,恶化的,伤痛的,尴尬的,无聊的,疯狂的,失望的,焦虑的,害怕的,厌恶的);3-中性(平常的,平静的,清醒的,工作,空白,报告,新闻,事实)

1.2 问题阐述

在本研究中,我们利用超图的方法根据用户在线的档案信息和活动信息中提取的特征进行用户属性挖掘,并考虑各个用户属性之间的关联性,提高用户属性挖掘的准确度.具体地说,我们选择其中一种用户属性(例如:职业)作为目标属性来得到一个预测该属性的模型,其他的用户属性(例如:年龄,性别,婚姻状况,兴趣爱好和情感倾向)则被称为辅助属性,用来帮助准确地推断目标属性.

给定 Google+用户集 U,每个用户 $u\in U$ 对应二维数组$[Xu, Au]$.$Xu=[x_1,x_2,\ldots,x_K]$,其中 K 为用户属性总个数,x_k 是第 k 个用户属性的用户特征.$A_u=[a_1,a_2,\ldots,a_K]$表示用户属性的真实值.将目标属性定义为 T,辅助属性定义为 H,那么整个属性值定义为 $A=[H,T]$.因此,问题正式被定义为:关联性用户属性推断方法:给定一个 Google+用户集 U,学习一个预测函数 $f(X_u,H)\rightarrow T_u$ 来推断出用户的目标属性值.

2 基于超图模型的用户属性推断

本节介绍提出的基于超图模型的关联性用户属性推断方法.图 2 是该方法的示意图.

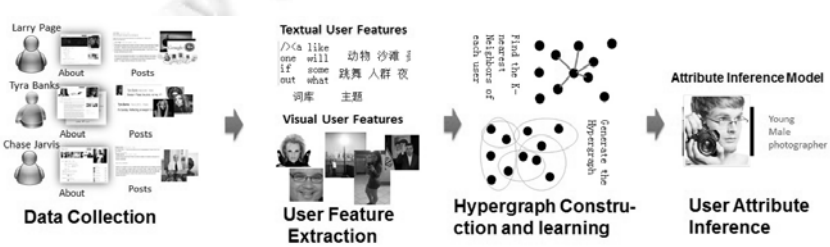


图 2 提出方法的流程图

2.1 用户特征提取

首先,我们需要将收集到的用户档案信息和发布信息进行统一的整理和转化,得到相应的用户特征.我们分别针对文本内容和视觉内容进行提取特征.

基于文本内容的特征有 3 种,分别是社会语言学特征,一元模型特征和基于主题的特征.社会语言学特征建立含有社会语言学的词和符号(例如:umm,uhhuh,>_<,><等)的词库,而一元模型则建立去掉这些符号的词库.每个词库含有一万个对某种用户属性具有一定识别力的词.根据单词是否出现作为特征的权重提取出社会语言学特征和一元模型特征.对于基于主题的特征(latent dirichlet allocation,简称 LDA)[13]是一种文档主题生成模型,被用来从档案信息和发布信息中提取用户潜在的主题.在主题提取之后,每一个用户被表示成导出的主题空间上的概率分布,而每个主题又对应由单词构成的概率分布.在本研究中,我们用了 100 个主题和 10 000 个单词来进行 LDA 处理.

基于视觉内容的特征有 3 种,分别是档案图片特征、档案图片脸部特征和发布图片特征.我们利用档案图片和发布图片两种信息来进行视觉特征提取.对于档案图片特征,每个档案图片提取出 809 维特征向量[14],其中 81 维为颜色矩,37 维为边缘直方图,120 维为小波纹理特征,59 维为局部二值模式(local binary patterns,简称 LBP)特征[15]和 512 维为通用搜索树(generalized search trees,简称 GIST)特征[16].另外,大多数档案图片包含脸部图像,它对识别面部属性非常有帮助,比如说年龄和性别.因此,我们从档案图片中检测出脸部,同样对其提取出 809 维特征向量作为档案图片脸部特征.对于发布的图片,我们明确地将每个用户上传的照片映射到一个预先定义的概念列表中.它是在观察了 88 988 张下载的发布图片后人为建立的 79 个类别的概念列表,具体见表 2.每个概念类别有近 100 张图片来训练.我们在监督式学习下训练 79 个概念分类器.每张图片提取出密集的 HOG 特征[17],接着利用 LLC(locality-constrained linear coding)[18]进行图片的编码表示.对于每一个概念,我们用 LIBLINEAR[19]训练一个 SVM 分类器.分类器的结果用 sigmoid 函数转换成概率值.因此,每张图片最终表示成 79 维向量,对应着每个概念的可能性.考虑到一个用户可能发布了很多张图片,我们通过一种聚合方法(max-pooling,最大池化)对各图片结果结果进行处理,得到每个用户聚合后的 79 维特征向量,这个向量也就是发布图片特征.

表 2 详细概念列表

概念列表																			
动物	沙滩	美女	鸟	身体	书籍	建筑	车	卡通	猫	名人	儿童	城市布	云	颜色	情侣	跳舞	人群	夜	设计	狗
饮料	电子产品	家庭	飞机	花	美食	水果	怪胎	鹅	草	草原	房子	图标	室内	昆虫	湖	山水画	叶子	男人	模型	
山	自然风光	办公室	油画	宫殿	派对	人物	表演	摄影	肖像	海报	道路	房间	雕塑	海	天空	雪	士兵	运动	斑	
点	松鼠	体育场	石头	商店	街道	日落	谈话	文本	微小植物	塔	玩具	交通	宇宙	树	手表	瀑布	女人			

2.2 超图建立

用 $G=(V,E,W)$表示超图,其中,V 表示顶点集合,E 表示超边的集合,W 表示超边所对应的权重.每一条超边 e_i 都有一个对应的权重 $w(e_i)$.超图 G 的顶点和超边之间的关系可以用一个$|V|\times|E|$的关联矩阵进行描述:

$$h(\mathrm{v},\mathrm{e})=\begin{cases}1, & \text{if } v\in e\\ 0, & \text{if } v\notin e\end{cases} \tag{1}$$

对于每一个顶点 $v_i\in V$,它的顶点度定义为

$$d(v_i)=\sum_{e\in E}w(e)h(v_i,e) \tag{2}$$

对于每一条超边 $e_i\in E$,它的超边度定义为

$$\delta(e_i)=\sum_{v\in V}h(v,e_i) \tag{3}$$

定义 D_v 和 D_e 分别为顶点度和超边度的对角矩阵,相应地,定义 W 为超边权重的对角矩阵.

在本文中,超图的顶点表示用户,超边表示用户之间的关系.我们建立两种类型的超边分别表示用户产生内容的相似性关系和用户属性相关性关系.对于建立基于用户内容相似形的超边,我们采用近邻特征用户的方法

构建.每一种用户特征都可以针对每一位用户寻找最近似的 k 个用户建立一条超边,得到与相应用户特征个数相同条边.那么,6 种用户特征就可以得到 6 种超边,依次记为 $E^{(1)},E^{(2)},\ldots,E^{(6)}$.对于构建表示用户属性关系的超边,根据已知的用户属性,属于同一个属性值的用户可以建一条超边,得到与属性值个数相同条边,可以记为 $E^{(7)}$.

根据式(1)~式(3)分别确定关联矩阵 H 和对角矩阵 D_v 和 D_e 后,依次确定不同种类的超边权重.

对于基于用户特征的超边 $E^{(1)},E^{(2)},\ldots,E^{(6)}$,其权重根据特征之间的相似度来决定:

$$w(e_i^{(t)})=\sum_{U_a,U_b\in e_i}\exp\left(-\frac{\|U_a-U_b\|^2}{\sigma^2}\right) \tag{4}$$

式中,U_a 和 U_b 表示用户 a 和 b 的相应特征值,$t\in\{1,2,3,4,5,6\}$表示是其中某一种特征的权重.为了考虑其他特征权重的公平性,我们对其进行如下统一的归一化处理:

$$w(e_i^{(t)})^*=\frac{w(e_i^{(t)})}{\max(w(e^{(t)}))} \tag{5}$$

对于基于用户属性的超边 $E^{(7)}$,我们将其权重设为 p=1.

2.3 超图学习

建立了超图之后,就可以在超图上学习,进行用户属性推断.我们参考文献[20]所提出的方法,利用正则化框架进行超图学习:

$$F^*=\arg\min_F\{Q(F)\} \tag{6}$$

$$Q(F)=\mu R_{\text{emp}}(F)+\Omega(F) \tag{7}$$

$$\mathrm{R}_{\text{emp}}(F)=\sum_{i=1}^{|V|}\|f_i-y_i\|^2 \tag{8}$$

$$\Omega(F)=\frac{1}{2}\sum_{i,j=1}^{|V|}\sum_{e\in E}\frac{1}{\delta(e)}\sum_{\{v_i,v_j\}\subseteq e}w(e)\left\|\frac{f_i}{\sqrt{d(v_i)}}-\frac{f_j}{\sqrt{d(v_j)}}\right\|^2 \tag{9}$$

式中,$F=[f_1,f_2,\ldots,f_{nc}]$是经过学习之后得到的属性推断得分矩阵,$F_{ij}$ 表示为第 i 个用户是否是属性值 j 的推断得分,n_c 是某属性的属性值个数.$R_{\text{emp}}(F)$是一个经验损失,衡量获得的属性推断得分 F 与先验得分 Y 之间的差别.先验得分 $Y=[y_1,y_2,\ldots,y_{nc}]$是一个 $n\times n_c$ 矩阵,每一列对应着一个属性值的训练集属性,例如,y_i 对应着属性值为 i 的用户先验得分.在本文中,训练集用户是该属性值时,先验得分设为 1,不是该属性值时,先验得分设为 0;测试集用户得分全设为 0.因此,y_i 是一个 $n\times 1$ 矩阵,对应着 n 个用户的属性值为 i 的先验得分,只有训练集中该属性值为 i 的用户得分为 1,其余全部为 0.$\Omega(F)$是一个平滑约束,使它减小意味着如果若干个顶点同属于多个超边,它们将拥有相似的推断得分.比如说,几个用户发布了类似的信息或者图片,他们可能具有相同的兴趣爱好或者职业等.如果两个用户的很多特征或属性相似,同属于较多条超边,则他们很可能具有相同的属性.参数μ是一个权重系数,可以调节$\Omega(F)$和 $R_{emp}(F)$之间的比重.还有需要注意的是每条超边利用超边度$\delta(e_i)$进行标准化,超边度也就是该超边所包含的顶点个数.这样不同大小的超边也可以公平对待.最理想的属性推断得分 F^*是当函数 Q(F)取得最小值的时候,见式(6).

式(9)可以简化成:

$$\begin{aligned}\Omega(F)&=\frac{1}{2}\sum_{i,j=1}^{|V|}\sum_{e\in E}\frac{1}{\delta(e)}\sum_{\{v_i,v_j\}\subseteq e}w(e)\left\|\frac{f_i}{\sqrt{d(v_i)}}-\frac{f_j}{\sqrt{d(v_j)}}\right\|^2\\&=\frac{1}{2}\sum_{i,j=1}^{|V|}\sum_{e\in E}\frac{w(e)h(v_i,e)h(v_j,e)}{\delta(e)}\left\|\frac{f_i}{\sqrt{d(v_i)}}-\frac{f_j}{\sqrt{d(v_j)}}\right\|^2\end{aligned}$$

$$= \sum_{i,j=1}^{|V|} \sum_{e\in E} \frac{w(e)h(v_i,e)h(v_j,e)}{\delta(e)} \left(\frac{f_i^2}{d(v_i)} - \frac{f_i f_j}{\sqrt{d(v_i)d(v_j)}} \right)$$

$$= \sum_{i=1}^{|V|} f_i^2 \sum_{e\in E} \frac{w(e)h(v_i,e)}{d(v_i)} \sum_{j=1}^{|V|} \frac{h(v_j,e)}{\delta(e)} - \sum_{i,j=1}^{|V|} \sum_{e\in E} \frac{f_i w(e)h(v_i,e)h(v_j,e)f_j}{\sqrt{d(v_i)d(v_j)}\delta(e)}$$

$$= \sum_{i=1}^{|V|} f_i^2 - \sum_{i,j=1}^{|V|} \sum_{e\in E} \frac{f_i w(e)h(v_i,e)h(v_j,e)f_j}{\sqrt{d(v_i)d(v_j)}\delta(e)}$$

$$= F^T F - F^T D_v^{-1/2} HWD_e^{-1} H^T D_v^{-1/2} F \tag{10}$$

定义矩阵 A 为

$$A = D_v^{-1/2} HWD_e^{-1} H^T D_v^{-1/2} \tag{11}$$

接着就可以将式(7)的函数 $Q(F)$重新写成:

$$Q(F) = F^T (I - A)F + \mu(F - Y)^T (F - Y) \tag{12}$$

为了取得式(7)函数 $Q(F)$的最小值,必须使其梯度为 0,得到下面等式:

$$\frac{\partial Q}{\partial F}|_{F=F^*} = (\mathrm{I} - A)F^* + \mu(F^* - Y) = 0 \tag{13}$$

简化后得到:

$$F^* = \frac{\mu}{1+\mu}\left(\mathrm{I} - \frac{1}{1+\mu}A\right)^{-1} Y \tag{14}$$

定义 α=1/(1+μ),则式(14)可以转化为

$$F^* = (1-\alpha)(\mathrm{I} - \alpha A)^{-1} Y \tag{15}$$

虽然整个计算过程涉及的矩阵维数较大,但是由于 $\mathrm{I}-\alpha A$ 是一个高度稀疏的矩阵,所以整个计算过程是非常高效的.在计算出了属性推断得分矩阵 F 后,它的每一行对应着超图中的每个用户,n_c 个值分别表示对该属性 n_c 个属性值的推断得分,得分最高的属性值推断为该用户的该属性值.同理可得到,所有用户所有属性的推断值.

3 实 验

在本节中,我们对本文提出的算法在用户属性挖掘中的有效性进行定性和定量的研究.

3.1 实验设置

我们通过Google+开放的API来收集实验数据,关注预测用户属性,并通过人工标注用户属性的方式建立评估数据集.为了减轻标记工作,我们考虑 Google+上的名人用户,因为名人用户的档案信息可以很容易地通过其他平台(如:Facebook,维基百科 等)获得,这样不但可以减轻标记的工作量,还可以提高用户属性标记的准确率.经过筛选和过滤,最终得到 2 548 个名人用户和 846 339 条发布信息,并进行相应的人工标记.值得注意的是,我们所建立的根据用户在线产生数据进行属性推断的模型不仅可以用在名人用户上,也可以用在普通用户上.在实验中,我们将 50%的标记数据作为训练集,剩下的作为测试集.

为了进行用户属性推断结果的比较,我们测试如下几种方法:

- 单特征的超图:这种方法是利用 6 种用户特征各自单独进行构建超图,也就是说只包含基于用户特征的 $E^{(1)},E^{(2)},\ldots,E^{(6)}$中的一种超边,实现对某一种用户属性的推断.
- 多特征的超图:这种方法是同时利用 6 种用户特征进行构建超图,也就是说同时包含基于用户特征的 $E^{(1)},E^{(2)},\ldots,E^{(6)}$这 6 种超边,来实现对某一种用户属性的推断.
- 含属性之间关系的超图:这种方法就是我们所提出的关联性用户属性推断方法.它不但利用 6 种用户特征,而且还利用用户属性之间的关系来进行构建超图.也就是说同时包含 $E^{(1)},E^{(2)},\ldots,E^{(6)},E^{(7)}$这 7 种超边,来实现

对某一种用户属性的推断.值得注意的是,我们在进行某种用户属性推断的时候,并不会将所有用户的其他属性值用来构建超边.我们仍然基于训练集属性结果已知,测试集属性结果未知的假设上.所以,我们根据各个用户属性,针对其训练集,属于同一个属性值的用户可以建一条超边,得到与属性值个数相同条边,构建超边 $E^{(7)}$.

我们采用平均准确率(average precision,简称 AP)来进行衡量各种方法用户属性推断效果.准确率定义为每个属性值正确的推断个数除以实际该属性值用户个数.那么,对于每个属性的平均准确率则为各属性值准确率的平均值.其计算公式如下:

$$AP=\frac{1}{n_c}\sum_{i=1}^{n_c}\frac{right(i)}{total(i)} \tag{16}$$

式中,n_c 是该属性的属性值个数,$right(i)$和 $total(i)$分别表示属性值 i 推断正确个数和总个数.

3.2 实验结果

3.2.1 不同方法的结果对比

对于每一种用户属性,利用单属性的超图、多属性的超图、含属性之间关系的超图这 3 种不同的方法,利用之前提出的方法建立超图并学习,进行相应用户的属性推断.表 3 显示了不同方法对不同属性推断的结果.其中,超图建立中的系数 K 取为 5,超图学习中的系数α取为 0.1.从表 3 中可以看出:

不同类型的用户特征对不同的用户属性推断有不同的作用,相应的平均准确率(AP)也有所不同.平均准确率越高说明了该用户特征越有利于该用户属性的推断.例如,档案图片特征和档案图片脸部特征在进行推断用户属性“性别”的时候非常有帮助,它们所对应的平均准确率(AP)比较高.基于词是否存在的社会语言学特征和一元模型特征在各属性推断中都表现较好,尤其是属性年龄和婚姻状况中表现最好,说明我们词库的有效性及产生时所采用方法的合理性.发布图片特征利用图片基于概念的表示方法,在各个属性推断中表现也较好,特别是在属性兴趣爱好和情感倾向中,证明所选取的概念列表的有效性和采用方法的合理性.但是,我们也发现基于主题的特征在各属性推断中均未表现出良好的结果,说明 LDA 主题提取的方法很难有效地从用户发布内容中提取基于用户层面的简洁有识别能力的特征.

表 3 不同方法用户属性推断结果表

方法		性别	年龄	婚姻状况	职业	兴趣爱好	情感倾向
单特征的超图	社会语言学特征	0.519 500	0.553 915	0.583 414	0.112 311	0.514 247	0.340 801
	一元模型特征	0.548 535	0.554 602	0.577 035	0.118 026	0.513 111	0.344 840
	基于主题的特征	0.512 731	0.537 775	0.539 302	0.080 916	0.506 743	0.335 235
	档案图片特征	0.666 922	0.519 574	0.524 113	0.082 067	0.505 017	0.326 801
	档案图片脸部特征	0.632 560	0.519 574	0.527 575	0.074 188	0.494 293	0.375 301
	发布图片特征	0.571 647	0.525 412	0.506 430	0.117 545	0.529 845	0.360 823
多特征的超图		0.615 597	0.565 591	0.585 236	0.116 233	0.507 734	0.339 584
含属性之间关系的超图		0.684 590	0.683 379	0.645 064	0.110 663	0.501 922	0.339 633

多特征的超图推断结果在大多数情况下都比分别利用 6 种特征的单特征的超图推断结果好.这证明多特征的超图充分利用各种不同的用户特征,提高用户属性推断的准确性.对于个别多特征的超图结果略低于某种单特征的超图,是因为对于该属性,一些其他用户特征并不具有良好对该属性的可识别能力,加入该特征反而会带来明显的噪声,使其最后的多特征的超图推断结果有所影响,但是其结果仍比大部分单特征的超图对该属性推断结果好.

我们提出的含属性之间关系的超图在用户属性性别、年龄和婚姻状况推断中,结果明显优于其他对照方法.对于用户属性情感倾向,该方法的推断结果也比多特征的超图的推断结果好.特别值得注意的是,对于用户属性年龄,该方法的推断结果比多特征的超图推断结果提高 20.83%,说明了本文所提出的含属性之间关系的超图方法在年龄这一用户属性推断中具有相当明显的优势.但是,对于用户属性职业和兴趣爱好,该方法的推断结果比多特征的超图推断结果有所下降.考虑到这两种用户属性的属性值个数较多,每个属性值对应的用户较少.在加入其他用户属性的训练集属性值时,得到的其用户属性之间关系并不具有绝对的普适性,极易受到个别特殊用户的影响,从而导致了错误性的引导,使得最后结果产生了一些偏差,用户属性推断的准确率有所下降.并

且对于职业和兴趣爱好这两种用户属性,它们和其他属性之间关系更加复杂和微妙,不同属性值的情况完全不同,也使得同样加入其他用户属性的训练集属性值时会产生一些偏差,从而降低最终的用户属性推断结果.总的来说,根据这一结果仍旧可以验证利用各属性之间关系可以帮助进行一些用户属性推断.

图 3 用户建模中部分测试样本用户的预测属性图(带下划线为错误结果)

图 3 展示了利用含属性之间关系的超图(即关联性用户属性推断方法)进行用户属性推断来用户建模的例子.我们可以看到该方法准确地推断出大部分的用户属性,例如:男性、年轻、已婚,甚至是积极、IT 工作者等.

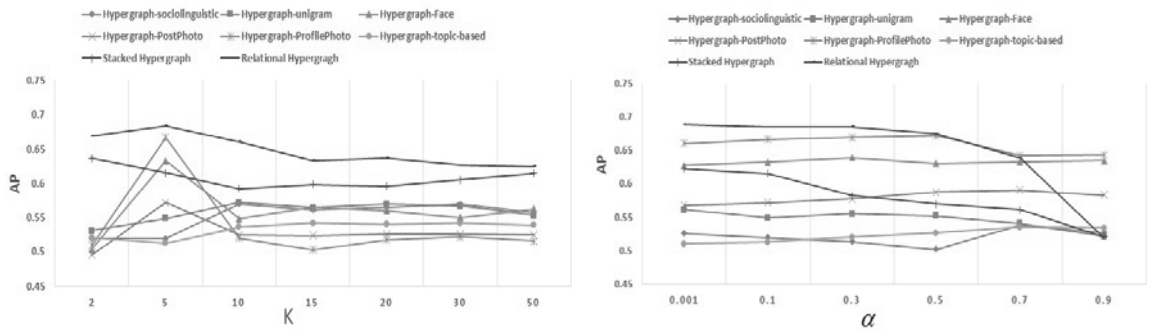
通过该实验,我们可以看出本文中提出的含属性之间关系的超图方法不但结合各种特征的优势,而且考虑到属性之间的相关关系,有效地提高用户属性推断的准确率.

3.2.2 参数设定的研究

不论对于哪种方法,选择最近似用户个数 K 在建立超边时扮演着重要的角色.对于系数 K,如果太小,则每个用户在相应的超边里关联了较少的用户;如果 K 太大,则用户在相应的超边里关联了较多的用户.用户在超边中关联过多或者过少的用户都不好,会影响最终的用户属性推断结果.在这里,我们以用户属性性别为例,对每一种用户属性推断方法都采用不同的 K 值进行用户属性推断,其实验结果如图 4(a)所示,其中系数α设为 0.1.从图中我们可以看到,K 值对不同方法的影响有所不同,但是平均准确率 AP 的整体趋势大致相同,一般都是先上升后向下降,平均准确率 AP 的最大值对应的 K 取值为 5 左右.

我们同时也考虑系数α在超图学习过程中的影响.我们同样以用户属性“性别”为例,对每一种用户属性推断方法都采用不同的α值进行用户属性推断,其实验结果如图 4(b)所示,其中系数 K 设为 5.从图中我们可以看到,α值对不同方法的影响有所不同,但是平均准确率 AP 的整体趋势大致相同,大部分都是一个不断下降的过程,当α<0.3 时,各方法的平均准确率 AP 都相对比较平稳.

根据对两个参数 K 和α取值的研究后,我们发现当 K=5 和α=0.1 时,各方法的用户属性推断结果较好,所以实验中采用该组参数设置.



(a) (b)

图 4 不同 K 和α值对应各超图方法推断结果图

4 总结与展望

本文提出了一种基于超图学习的关联性用户属性推断方法来对社交网站中的用户属性进行推断.超图模型可以有效地对用户内容特征和属相关系进行有效地建模.在采集 Google+的数据集上的实验结果表明,通过考虑丰富的用户特征和用户属性之间的关系,我们可以在一定程度上提高用户属性推断的准确率.

由于超图方法本身的问题,考虑用户属性之间的关系对某一些属性的推断带来了一些噪声,导致并没有取得较理想的结果.在未来的研究中,我们可以进行研究和讨论自适应学习超边权重的问题.通过加入一个权重值,针对不同用户属性,采用不同的超边权重来达到降低噪声的目的,可以进一步提高用户属性推断的准确率.

致谢 在此,我们向对本文的工作给予支持和建议的同行表示感谢.

References:

[1] Garera N, Yarowsky D. Modeling latent biographic attributes in conversational genres. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009. 710−718.

[2] Sarawgi R, Gajulapalli K, Choi Y. Gender attribution: Tracing stylometric evidence beyond topic and genre. In: Proc. of the 15th Conf. on Computational Natural Language Learning. Association for Computational Linguistics, 2011. 78−86.

[3] Weber I, Castillo C. The demographics of Web search. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2010. 523−530.

[4] Zamal FA, Liu W, Ruths D. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In: Proc. of the 6th Int'l AAAI Conf. on Weblogs and Social Medis (ICWSM). 2012.

[5] Filatova E, Prager J. Occupation inference through detection and classification of biographical activities. Data & Knowledge Engineering, 2012,76:39−57.

[6] Tan CH, Lee L, Tang J, Jiang L, Zhou M, Li P. User-Level sentiment analysis incorporating social networks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2011. 1397−1405.

[7] Rao D, Yarowsky D, Shreevats A, Gupta M. Classifying latent user attributes in twitter. In: Proc. of the 2nd Int'l Workshop on Search and Mining User-Generated Contents. ACM, 2010. 37−44.

[8] Bi B, Shokouhi M, Kosinski M, Graepel T. Inferring the demographics of search users. In: Proc. of the IW3C2. 2013.

[9] Pennacchiotti M, Popescu AM. Democrats, republicans and starbucks afficionados: User classification in twitter. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2011. 430−438.

[10] Bachrach Y, Kosinski M, Graepel T, Kohli P, Stillwell D. Personality and patterns of Facebook usage. In: Proc. of the 3rd Annual ACM Web Science Conf. ACM, 2012. 24−32.

[11] Quercia D, Kosinski M, Stillwell D, Crowcroft J. Our Twitter profiles, our selves: Predicting personality with Twitter. In: Proc. of the 3rd Int'l Conf. on Social Computing (Socialcom). IEEE, 2011. 180−185.

[12] Magno G, Comarela G, Saez-Trumper D, Cha M, Almeida V. New kid on the block: Exploring the Google+ social graph. In: Proc. of the 2012 ACM Conf. on Internet Measurement Conf. ACM, 2012. 159−170.

[13] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[14] Zhu J, Hoi S CH, Lyu MR, Yan S. Near-Duplicate keyframe retrieval by nonrigid image matching. In: Proc. of the 16th ACM Int'l Conf. on Multimedia. ACM, 2008. 41−50.

[15] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition, 1996,29(1):51−59.

[16] Torralba A, Murphy KP, Freeman WT, Rubin MA. Context-Based vision system for place and object recognition. In: Proc. of the 9th IEEE Int'l Conf. on. AI Memo 2003. IEEE, 2003. 273−280.

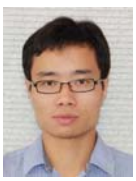
[17] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the Computer Vision and Pattern Recognition (CVPR 2005). IEEE Computer Society, 2005. 886−893.

[18] Wang JJ, Yang JC, Yu K, Lü FJ, Huang T, Gong YH. Locality-Constrained linear coding for image classification. In: Proc. of the Computer Vision and Pattern Recognition (CVPR 2010). IEEE, 2010. 3360−3367.

[19] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 2008,9:1871−1874.

[20] Zhou DY, Huang JY, Schölkopf B. Learning with hypergraphs: Clustering, classification, and embedding. Advances in Neural Information Processing Systems, 2007,19:1601.

项连城(1992－),女,浙江台州人,主要研究领域为社会媒体分析,多媒体检索,数据挖掘.



方全(1988－),男,博士生,主要研究领域为基于地理的媒体数据挖掘,社会媒体.

桑基韬(1985－),男,博士,助理研究员,主要研究领域为社会媒体分析,多媒体检索,数据挖掘.

徐常胜(1969－),男,博士,研究员,博士生导师,主要研究领域为多媒体分析/索引/检索,模式识别,计算机视觉.

路冬媛(1984－),女,博士,研究员,主要研究领域为网络数据挖掘,知识工程.