

文章编号:1001-9081(2014)08-2209-03

doi:10.11772/j.issn.1001-9081.2014.08.2209

基于粗糙集的微博用户性别识别

黄发良^{1*}, 熊金波¹, 黄添强¹, 刘西蒙²

(1. 福建师范大学 软件学院, 福州 350007; 2. 西安电子科技大学 计算机学院, 西安 710071)

(*通信作者电子邮箱 faliang.huang@gmail.com)

摘要:针对微博消息往往会展现出性别倾向性的特点,从消息内容挖掘的角度出发提出了一种基于粗糙集的微博用户性别识别算法。设计了一种基于容差粗糙集的微博消息表示模型(TRSRM),有效地刻画微博消息的性别特征。实验结果表明,在1000个真实微博用户的微博消息的测试集中下,所提模型的准确率比特征项频数表示模型平均提高了7%,取得了更好的识别效果。

关键词:微博挖掘;性别识别;粗糙集; k 近邻分类器;网络安全

中图分类号: TP391.3; TP18 **文献标志码:**A

Gender identification of microblog users based on rough set

HUANG Faliang^{1*}, XIONG Jinbo¹, HUANG Tianqiang¹, LIU Ximeng²

(1. Faculty of Software, Fujian Normal University, Fuzhou Fujian 350007, China;

2. School of Computer Science and Technology, Xidian University, Xi'an Shaanxi 710071, China)

Abstract: Concerning gender tendency hidden in microblog messages posted by microblog users, a novel approach based on rough set theory was proposed to identify microblog user gender. In the proposed approach, a new Representation Model based on Tolerance Rough Set (TRSRM) was devised, which can effectively represent gender characteristics of microblog messages. The experimental results show that the accuracy rate of the proposed approach is 7% higher than frequency model approach by testing messages of 1000 real microblog users, and so the TRSRM achieves better recognition performance.

Key words: microblog mining; gender identification; rough set; k -Nearest Neighbor (k NN) classifier; network security

0 引言

作为新型媒体的微博,其本质是一种集成化、开发化的互联网社交服务,微博用户可借助计算机、手机等各种终端上传、下载或浏览图片、文本、视频等媒体数据来实现不同信息的即时传播与共享。自其出现以来,微博深受网民喜爱。例如,Twitter 从其 2006 年创办以来,到 2012 年其注册用户已超过 5 亿。微博的流行一方面极大地丰富了人们的交流方式,然而另一方面给作恶之人提供了一个伪装的面具,由于 Twitter 等微博平台具有开放与身份虚拟等特征,一些不法分子可以利用平台从事身份盗用、金融诈骗等各种犯罪活动^[1]。因此,研究如何发现微博用户的真实身份进而有效遏制犯罪行为有着重要的应用价值。

通过微博实名制的办法可以完全掌握微博人员的身份信息,但这会大大损害用户的微博参与度。微博用户在微博平台上所留下的微博数据为数据挖掘提供了丰富的数据来源,数据挖掘技术为揭示微博用户的真实身份独辟蹊径。表面上看,微博数据是文本数据,但与静态文本数据相比较,其具有个性化、符号化、非规范化特点,这使得传统的文本挖掘技术很难直接应用于微博数据挖掘。

由于微博平台出现的时间较短,再加上微博数据的复杂性,当前微博用户性别识别的研究成果还少见报道。Köse

等^[2]应用语义分析策略对网络聊天室的土耳其语聊天数据进行挖掘以图发现聊天人员的真实性别;Cheng 等^[3]尝试通过挖掘分析 Reuters 新闻组的 Web 文档数据来实现文档作者性别的识别;Miller 等^[4]运用流挖掘算法对 Twitter 用户的性别进行预测;Köse 等^[5]比较研究运用各种文本挖掘技术分析网聊数据并识别聊天人员的性别;唐琴等^[6]通过提取中文文本中的性别倾向性描述词、性别倾向性称谓词与人名等特征来实现文本中人物性别的识别。绝大多数现有性别识别方法都没有考虑到微博数据的口语化与个性化等特点,不能直接应用于微博用户的性别识别。相关研究^[7-9]表明,在博客语境中,男女在语言使用与讨论问题的侧重点上存在着差异。

本文对如何利用海量微博信息识别微博用户性别进行了研究,提出了一种基于消息内容挖掘的微博用户性别识别方法。该方法首先通过分析数据,遴选出与微博用户性别相关的特征数据;然后建立一种基于容差粗糙集的微博消息表示模型;最后利用 k 近邻(k -Nearest Neighbor, k NN)分类器来识别微博用户性别。实验表明,该方法取得了很好的分类性能。

1 问题描述

尽管恶意微博用户在注册微博平台时使用虚假的性别信息,但作为微博信息传播网络中的节点,恶意微博用户在进行微博消息发布、浏览与评论的过程中不可避免地表现出其真

收稿日期:2014-04-02;修回日期:2014-05-02。

基金项目:教育部人文社会科学研究青年基金资助项目(12YJCZH074);福建省教育厅科技项目(JA13077)。

作者简介:黄发良(1975-),男,湖南永州人,副教授,博士,主要研究方向:数据挖掘;熊金波(1982-),男,湖南益阳人,讲师,博士,主要研究方向:大数据安全;黄添强(1970-),男,福建莆田人,教授,博士,主要研究方向:数据挖掘;刘西蒙(1988-),男,陕西西安人,博士研究生,主要研究方向:大数据安全。

实性别所应共同具有的传播特征。男性或女性的性别群体在参与微博消息传播的过程中留下大量的信息,例如:发布微博消息的时间、频率等使用微博的行为特征;所发布、浏览或评论的微博消息的内容特征;与其粉丝与关注人之间的关系特征等。

恶意微博用户为了捕获更多的潜在作案对象,其必然会尽量加强其微博的传播能力以结识更多的微博用户,而在用户、微博消息和用户关系三位一体的微博消息传播中,微博消息是传播的主要内容,直接影响微博消息的受关注程度和转发度,其最主要的两个传播途径是转发和粉丝,因此,本文从消息内容的角度对微博用户进行表征,具体地说,选择微博用户发布、转发、评论与私聊的消息构造一个用来表征微博用户的消息集。由文本挖掘技术可知,微博用户的性别识别问题可形式化为如下的文本二分类问题。

对于训练集为性别信息已知的微博用户集合 $U = \{(m_1, g_1), (m_2, g_2), \dots, (m_N, g_N)\}$, 其中:二元组 (m_i, g_i) 表示微博用户 u_i ,向量化消息集 $m_i = (m_{i1}, m_{i2}, \dots, m_{id})^T$ 是借助文本表示机制对微博用户 u_i 的消息集进行向量化的结果; $g_i \in \{-1, +1\}$ 是微博用户 u_i 的性别标签编码, -1 表示 female, +1 表示 male。微博用户性别识别的根本任务就是需要通过某种学习策略产生一个高性能的分类器 $g = f(m)$,使其对于性别信息不确定的微博用户能准确确定其性别分类。

2 微博消息表示模型

由第 1 章的问题描述可知,微博用户消息表示机制对微博用户性别的识别有着重要的影响。为此,本节描述两种表示模型。

2.1 特征项频数表示模型

特征项频数表示模型(Frequency Model, FM)的基本思想是:在进行微博用户消息集的向量化表示时,将一个用户的消息集视为一个文档,然后借助向量空间模型(Vector Space Model, VSM)与频数加权机制对该文档进行向量化,具体描述如下。

假定训练集中的消息集合为 $M = \{m_1, m_2, \dots, m_N\}$,微博用户 u_i 的消息集 m_i 在经过停用词过滤、词根化等预处理后生成的特征项集合 $T_i = \{t_{i1}, t_{i2}, \dots, t_{ir}\}$,整个训练集预处理后生成的特征项集合 $T = \bigcup_{i=1}^N T_i$,采用特征项频数对特征项进行赋权值,即有 $\text{vec}(m_i) = (w_{i1}, w_{i2}, \dots, w_{iN})$ 。

$$w_i = \begin{cases} freq(t_i), & \text{特征词 } t_i \text{ 出现在消息 } m_i \text{ 中} \\ 0, & \text{其他} \end{cases} \quad (1)$$

其中 $freq(t_i)$ 表示特征项 t_i 在消息集 m_i 中的出现的频数。

2.2 容差粗集表示模型

特征项频数表示模型简单直观,容易实现,但其存在消息向量的稀疏性与消息向量之间的“假相似”等缺陷^[10]。为了克服这些缺陷,本文提出容差粗集表示模型(Representation Model based on Tolerance Rough Set, TRSRM)对消息进行向量化表示。容差粗集(Tolerance Rough Set)是一种代表性的粗糙集扩展模型^[11-12],其主要通过定义如下的容差关系来描述信息系统。

定义 1 容差关系(Tolerance Relation, TR)。假定信息系统 $I = (U, A)$, 其中: U 是数据对象的论域, A 是数据对象的属性集, B 是包含缺失值的属性子集, c_j 表示属性 j , $c_j(x)$ 是数据对象 x 的第 j 个属性。容差关系可形式化描述为式(2):

$$TR = \{(x, y) | x \in U \wedge y \in U \wedge \forall c_j (c_j \in B \Rightarrow (c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *))\} \quad (2)$$

消息向量可以借助容差关系的思想进行近似,对于消息集 $P = \{P_1, P_2, \dots, P_m\}$, 其中 $P_k (k = 1, 2, \dots, m)$ 为用户 u_k 是所发布、转发、评论与私聊消息中的特征项集合,其特征项集为 X ,可以定义不可明辨关系 IND ,容差关系 Ψ ,上近似 $\psi^+(X)$ 与下近似 $\psi^-(X)$ 如下:

$$IND_\lambda(t_i) = \{t_j | f_p(t_i, t_j) \geq \lambda\} \cup \{t_i\} \quad (3)$$

$$t_i \Psi t_j \Leftrightarrow t_i \in IND_\lambda(t_j) \quad (4)$$

$$\psi^+(X) = \left\{ t_i \in T \mid \frac{|IND_\lambda(t_i) \cap X|}{|IND_\lambda(t_i)|} > 0 \right\} \quad (5)$$

$$\psi^-(X) = \left\{ t_i \in T \mid \frac{|IND_\lambda(t_i) \cap X|}{|IND_\lambda(t_i)|} = 1 \right\} \quad (6)$$

其中: T 是特征词集, λ 是容差阈值, $f_p(t_i, t_j)$ 是特征词 t_i 与 t_j 共现的消息集在 M 中的比例, $\psi^+(X)$ 是特征词集 X 的上近似, $\psi^-(X)$ 是特征词集的下近似。语义上, $\psi^+(X)$ 可理解为微博用户 u_i 性别特征的内涵,而 $\psi^-(X)$ 可理解为微博用户 u_i 性别特征的外延。根据原有特征词与新增特征词存在不可明辨关系,对频数加权机制进行改进,通过性别特征的外延扩展使得一些原本不属于一个消息集的特征词被添加,这样就会大幅度降低“零相似”现象出现的概率。容差粗集表示模型的具体描述如下。

对于消息集合 M 与特征项集合 T ,其中微博用户 u_i 消息集 $m_i = \{t_1, t_2, \dots, t_n\}$,其上近似为 $V_i = \{v_{i1}, v_{i2}, \dots, v_{in}\}, w_{ij}$ 是特征词 t_j 在消息集 m'_i 中的权重,其中 m'_i 是 m_i 的上近似, w'_{ij} 是经过规范化后的特征词权重,该表示机制可以形式化为式(7)与(8):

$$w_{ij} = \begin{cases} (1 + f_{m_i}(t_j)) \times \text{lb}(N/f_M(t_j)), & \text{特征词 } t_j \text{ 出现在消息集 } m_i \text{ 中} \\ (\min_{t_k \in m_i} w_{ik}) \times \frac{\text{lb}(N/f_M(t_j))}{1 + \text{lb}(N/f_M(t_j))}, & \text{特征词 } t_j \text{ 出现在消息集 } m_i \text{ 的上近似中但} \\ & \text{不出现消息集 } m_i \text{ 中} \\ 0, & \text{特征词 } t_j \text{ 不出现在消息集 } m_i \text{ 的上近似中} \end{cases} \quad (7)$$

其中 $f_M(t_j)$ 表示特征词 j 在消息集 M 中的出现频率。

$$w'_{ij} = w_{ij} / \left(\sum_{t_j \in m'_i} w_{ij} \right) \quad (8)$$

3 实验与分析

考虑到微博用户性别识别问题在本质上是一个二分类问题,本文采用 k 近邻(kNN)作为微博用户性别分类器。 kNN 分类器简单易懂容易实现,且无需训练,其基本思想是:如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别。本文的实验环境是:CPU 是主频 1.7 GHz 的 Intel Core i5-3317U,内存 4 GB,OS 为 Windows 7。

由于当前没有通用的微博用户性别识别实验数据,本文利用微博数据爬虫从 Twitter 平台上抓取 1000 个性别已知且真实的用户(500 个男性和 500 个女性)持续 1 个月的微博消息数据,在经过停用词过滤、词根化等预处理后,该数据有特征词数为 17156。本文从以下 3 个方面测试分类器性能。

1)微博消息表示机制对识别率的影响。

本文通过随机选取 600 个微博用户(300 个男性和 300 个女性)的方式构造 8 组数据,对两种微博消息表示机制进

行10交叉的实验比较。实验结果见表1。由表1可知,TRSRM所对应的微博用户性别的识别正确率要明显高于FM,这说明TRSRM能通过引入微博用户性别特征的外延来有效提高识别正确率。

表1 微博用户性别识别率比较 %

组别	FM	TRSRM
1	75.35	81.73
2	72.17	82.76
3	75.85	80.84
4	70.66	79.61
5	76.51	83.25
6	74.33	81.08
7	74.49	81.13
8	76.94	82.22
平均	74.54	81.58

2)消息类型对识别率的影响。

本文对微博用户数据中的4种消息类型(用户发布的消息(P)、用户转发的消息(F)、用户评论的消息(R)与用户私聊的消息(W))进行组合,形成15种不同消息类型的数据集,进一步在每一类型的数据集中运用随机选择策略构造类分布平衡的10组训练样本与测试样本,训练样本与测试样本的大小与构成本分别是“300个男性+300个女性”与“150男+150女”。实验结果见表2,表中每个准确率值都是其对应10组数据的实验结果平均值。从表2可以看出,一般地,消息类型越多则性别识别率越高,然而,二者之间也绝不是简单的线性正相关关系,例如,具有最高识别率是消息类型P+R+W而不是P+F+R+W。这一方面说明绝大多数类型微博消息操作行为具有性别倾向性,另一方面也说明微博用户转发的消息从性别识别的角度看属于噪声数据。进一步分析表2可以发现,消息类型P+R诱导出的分类器具有最高男性识别率,而具有最高女性识别率的分类器是基于消息类型P+R+W,这与现有的社会学研究结果“男性比女性往往更有主见,更愿意主动公开发表观点与表达看法”^[9]是相吻合的。

表2 消息类型对识别准确率的影响

消息类型	识别准确率/%			消息类型	识别准确率/%		
	男性	女性	整体		男性	女性	整体
P	79.32	79.54	79.430	F+W	75.56	79.05	77.305
F	77.49	77.82	77.655	R+W	78.08	79.58	78.830
R	78.67	78.44	78.555	P+F+R	82.13	81.33	81.730
W	74.43	78.61	76.520	P+F+W	79.87	82.09	80.980
P+F	78.51	78.25	78.380	P+R+W	82.27	84.56	83.415
P+R	83.25	81.26	82.255	F+R+W	78.36	81.74	80.050
P+W	80.88	82.49	81.685	P+F+R+W	82.95	82.54	82.745
F+R	78.94	78.81	78.875				

3)容差阈值对识别率的影响。

容差阈值是TRSRM模型中的一个重要参数,本文对容差阈值进行不同取值的实验,结果见图1。从图1可以看出,kNN分类正确率不是随着特征项的增加而增加的,而是呈现倒“V”字形曲线的总体趋势。由此可见,容差阈值的选取对微博用户性别识别算法性能有很大影响。

4 结语

通过微博消息内容进行微博用户性别分类是一个复杂的

问题,目前研究工作较少。本文从各种不同类型的微博消息中选取发送、转发、评论与私聊这4种类型消息,提出基于粗糙集理论的微博消息表示机制——TRSRM,利用kNN实现性别分类。实验结果表明,TRSRM能显著提高微博用户的识别正确率,同时还表明运用发送、评论与私聊消息对微博用户进行性别识别的效果最好,但不同性别的识别存在着差异,基于发送、评论与私聊消息的分类器能较好地正确识别女性微博用户,而基于发送与评论消息的分类器能更好地正确识别男性微博用户,但是在容差阈值的确定仍然是一个有待研究的难题。

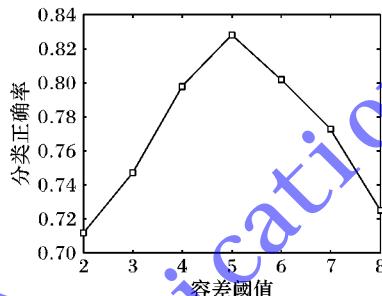


图1 容差阈值对分类正确率的影响

参考文献:

- [1] FBI San Diego. IC3 2011 Internet crime report released [EB/OL]. [2014-01-20]. <http://www.fbi.gov/sandiego/press-releases/2012/ic3-2011-Internet-crime-report-released>.
- [2] KÖSE C, ÖZYURT Ö, AMANMYRADOV G. Mining chat conversations for sex identification [C]// Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2007: 45–55.
- [3] CHENG N, CHNANDRAMOULI R, SUBBALAKSHIMI P. Author gender identification from text [J]. Digital Investigation, 2011, 8(1): 78–88.
- [4] MILLER Z, DICKINSON B, HU W. Gender prediction on twitter using stream algorithms with N -gram character features [J]. International Journal of Intelligence Science, 2012, 2(4): 143–148.
- [5] KÖSE C, ÖZYURT Ö, IKIBAS C. A comparison of textual data mining methods for sex identification in chat conversations [C]// Proceedings of the 4th Asia Information Retrieval Symposium. Berlin: Springer, 2008: 638–643.
- [6] TANG Q, LIN H. Research on gender recognition for character in text [J]. Journal of Chinese Information Processing, 2010, 24(2): 46–51.(唐琴,林鸿飞.文本中人物性别识别研究[J].中文信息学报,2010,24(2):46–51.)
- [7] LAKOFF R. Language and woman's place [M]. New York: Harper and Row, 1975.
- [8] TALBOT M M. Language and gender: an introduction [M]. New Jersey: Wiley-Blackwell, 1998.
- [9] ZHOU Y. Words that matter: gender features in the language use of weblog [D]. Hangzhou: Zhejiang University, 2007.(周炎.网络博客中的语言性别特征分析[D].杭州:浙江大学,2007.)
- [10] HUANG F, ZHANG S, HE M, et al. Clustering Web documents using hierarchical representation with multi-granularity [J]. World Wide Web, 2014, 17(1): 105–126.
- [11] PAWLAK Z. Rough sets: theoretical aspects of reasoning about data [M]. Hingham: Kluwer Academic Publishers, 1991.
- [12] AIMIRA J, LUTHER U. Generalized approximation spaces and applications[J]. Mathematische Nachrichten, 2004, 263/264(1):3–35.