第31卷 第4期
2016年8月

北京信息科技大学学报
Journal of Beijing Information Science & Technology University

Vol. 31 No. 4
Aug. 2016

文章编号:1674-6864(2016)04-0022-05 DOI: 10.16508/j.cnki.11-5866/n.2016.04.005

一种基于逻辑回归的微博内容隐私检测方法

张 玲[1],徐雅斌[1,2]

(1.北京信息科技大学 计算机学院,北京 100101;2.北京信息科技大学 网络文化与数字传播北京市重点实验室,北京 100101)

摘 要: 针对微博内容隐私问题,以新浪微博为研究平台,对微博内容的文本特征进行分析,从活动主题、参与者信息、地点信息、时间信息和用户情感五个方面,提取用于区分隐私微博和非隐私微博的特征变量,运用逻辑回归算法,提出一种基于逻辑回归的微博内容隐私检测模型。实验结果表明,该模型对隐私微博有较高的识别率,可行性较强。

关 键 词: 微博;隐私微博;隐私检测;文本特征;逻辑回归

中图分类号: TP 399 **文献标志码:** A

A micro-blog content privacy detection method based on logistic regression

ZHANG Ling[1], XU Yabing[1,2]

(1. Computer School, Beijing Information Science & Technology University, Beijing 100101, China; 2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science & Technology University, Beijing 100101, China)

Abstract: For the problem of privacy in microblog content, text characteristics of Sina mricroblog content were analyzed, and then feature variables for distinguishing the confidential and non-confidential microblogs were exacted according to activity theme and information of participant, location, time and user's emotion. By applying logic regression algorithm, a detection model of privacy in microblog content was put forward. The experimental result shows that the proposed model has high recognition rate and stronger feasibility for confidential microblogs.

Key words: micro-blog; confidential microblogs; privacy detection; text features; logistic regression

0 引言

近年来,随着智能手机、平板电脑等移动终端的普及,热爱分享身边新鲜事的人们使用微博的频率越来越高。新浪微博作为国内最大的微博网站,其注册用户已超过5亿,每天有超过1亿条微博内容产生[1]。然而,这些信息的大量公开导致隐私泄露的案例屡见不鲜,给许多用户的人身财产安全以及个人生活都造成了很大损失。因此,如何检测出微博内容中包含的隐私信息以减少隐私信息的发布,成为亟待解决的热点问题。

对此,国内外相关研究较少。文献[2-4]研究表明当前国内外微博用户主动暴露隐私的现象普遍存在。文献[5]研究了隐私信息如何在Twitter平台安全隐私控制下泄露的问题。针对泄露危害较大的酗酒状态、度假计划和疾病信息这3种隐私信息,国外文献[2]分别采用关键词匹配、命名体识别和规则匹配3种分类方法来识别。国内文献[6]采用基于贝叶斯的二级分类方法来检测。本文研究的对象是新浪微博,我们更关注后者。虽然该方法能以较高的准确率检测出这3种隐私信息,但微博内容中的隐私信息远远不止这3种,并且每一种隐私信息所包含的关键词与语法特征都大不同。所以,这种方法不具有普适性和拓展性。

通过大量观察分析发现,虽然不同微博内容中泄露的隐私信息不尽相同,泄露的程度也不一样,但

收稿日期: 2016-05-04

基金项目: 国家自然科学基金资助项目(61370139);网络文化与数字传播北京市重点实验室资助项目(ICDD201309);北京市属高等学校创新团队建设与教师职业发展计划资助项目(IDHT20130519)

作者简介: 张 玲,女,硕士研究生;通讯作者:徐雅斌,男,博士,教授。

都存在这样一个共性: 含有隐私信息的微博一般都包括 5 大特征信息,分别是活动主题、参与者信息、时间、地点和用户情感。当这些细节信息一一俱全,甚至更细节化的时候,就会发生隐私泄露。因此,本文分别采用不同的方法提取出微博内容中的 5 大特征,并在此基础上提出一种基于逻辑回归的隐私检测方法,从而提高隐私检测的准确性。

1　隐私泄露检测模型设计

1.1　特征属性的分析与获取

1.1.1　活动主题

本文选取泄露信息较多的 5 个活动主题 z,依次分别是“交通”、“娱乐”、“生活”、“学习”和“工作”,作为本文研究的目标活动主题。

本文将从背景主题、地理区域、发布时间和使用词汇 4 个方面来判定 z。

基于稀疏增量式模型,背景主题和地理区域对主题 z 的制约可以表示为:

$$P(z \mid z_b, r) = P(z \mid \theta_0, \theta_r) = \text{Multi}(z \mid \theta_0 + \theta_r) \quad (1)$$

式中,z_b 为背景主题; r 为地理区域; θ_0 为背景主题分布参数; θ_r 为区域 - 主题分布参数。

发布时间对主题 z 的制约可以表示为:

$$P(t \mid z) = \text{Multi}(t \mid \theta_{zt}) \quad (2)$$

式中,t 为发布时间; θ_{zt} 为主题 z 对应的时间多项式分布参数。

同样,基于稀疏增量式模型,地理区域对主题 z 包含的词汇的制约可以表示为:

$$P(w \mid r, z) = \text{Multi}(w \mid \varphi_{rz} + \varphi_{0z}) \quad (3)$$

式中,w 为包含词汇; φ_{rz} 为主题是 z 时的地理区域 - 词汇分布参数; φ_{0z} 为主题是 z 时的背景词汇分布参数。

分布参数 θ_0、θ_r、θ_{zt}、φ_{rz}、φ_{0z} 的获取过程:

1) 从狄利克雷过程混合模型 $d(\alpha_0)$ 中抽样得到 θ_0。α_0 为 θ_0 的超参数,设定为 $50/n_z$,n_z 为目标主题 z 的个数;

2) 对于每一个地理区域 $R = 1, 2, \cdots, n_R$: 从 $d(\alpha_r)$ 中抽样得到 θ_r,并从中抽取 r,α_r 为 θ_r 的超参数,设定为 $50/n_z$;

3) 从由 θ_0 和 θ_r 构造的多项式分布中抽取目标主题 z;

4) 针对目标主题 z,从 $d(\psi)$ 中抽样得到 θ_{zt},并从中采样得到 t。ψ 为 θ_{zt} 的超参数,设定为 0.1;

5) 针对每一个目标主题 $Z = 1, 2, \cdots, n_Z$: 从 $d(\beta_0)$ 中得到 φ_{0z},从 $d(\beta_r)$ 得到 φ_{rz};

6) 对于微博 m 中的每一个词汇 $W = w_1, w_2, \cdots, w_m$: 从由 φ_{0z} 和 φ_{rz} 构造的多项式分布中抽取词汇 w_i;

7) 那么,判定微博 m 的活动主题为 z 的概率为:

$$P(z \mid m) = P(z \mid z_b, r) \times P(t \mid z) \times \prod_{w \in W} P(w \mid r, z)。$$

最终按照微博 m 属于各目标主题的概率值的高低,将 m 的活动主题归为概率最大的目标主题。

1.1.2　参与者信息

把微博中包含的人名(或者机构名)的个数 n_{pn} 和“@”的个数 n_a 分别作为衡量参与者信息这一特征的指标。

本文采用基于多特征相融合的命名实体识别方法[7]来检测微博内容中的姓名或机构名。“@”符号则可以通过微博内容结构直接解析匹配出来。

1.1.3　地点信息

将微博内容中所包含的绝对物理地点的个数 n_{ap}、相对物理地点的个数 n_{rpl} 和逻辑位置的个数 n_{ll} 作为地点信息这一特征的指标。

1) 绝对物理地点的识别。绝对物理地点指的是由行政划分的地址单元组成的物理地址,如“北京市海淀区清河小营东路 12 号”,“北京小营东路 12 号”。本文将采用一种基于隐马尔可夫模型的方法[8],作为绝对物理地点的识别方法。

2) 相对物理地点的识别。相对物理地点指的是有特定功能的物理区域,如咖啡馆、图书馆、著名景点等。由于相对物理地点指向性明确,用户会在微博内容中直接指出其关键词。因此,本文采用简单的关键词匹配的方法来识别相对物理地点。

3) 逻辑位置的识别。逻辑位置则是指网络上的虚拟位置,如微信朋友圈、qq 空间。与相对物理地点一样,逻辑位置的指向性也很明确。所以,本文也采用同样的关键词匹配方法来识别逻辑位置。

1.1.4　时间信息

将微博内容中包含的时间短语的个数 n_{tp} 和时态信息 t_s 作为时间信息这一特征的指标。

1) 时间短语识别。虽然时间短语有多个类别,但每一种类别的格式相对固定,本文将采用规则库来识别时间短语。

①规则获取。设时间短语为 $T(t_1, t_2, \cdots, t_n)$ $(n \geqslant 1)$,t_i 为基本时间元。规则获取步骤如下:

Step1 从训练样本里手动提取出所有的时间短语，去冗余后得到时间短语集合 T_{set}；

Step2 对集合 T_{set} 进行分词和词性标注，得到带有词性标注的时间短语集合 $T'_{set}=\{T_{p1},T_{p2},\cdots,T_{pn}\}$，其中 T_{pi} 表示带有词性标记的一个时间短语，定义为二元组 $\langle T_i,P_i\rangle$，T_i 为时间短语，P_i 为词性标记；

Step3 将每个 T_{pi} 按时间单元的概念分解成若干个带词性标注的时间单元，去冗余后，得到一个时间单元集合 $t_{\text{set}}=\{t_{p1},t_{p2},\cdots,t_{pn}\}$，其中 t_{pi} 表示带有词性标记的一个时间单元，定义为二元组 $\langle t_i,p_i\rangle$，t_i 为时间单元，p_i 为词性标记；

Step4 对集合中的每个 t_{pi} 进行规则转化。

②识别过程。先对未标注的微博测试语料进行分词与词性标注，结合事先构建好的时间单元规则库一一识别出时间单元，最后把紧挨在一起的时间单元合并成完整的时间短语。以上是时间短语识别的全部过程。

2) 时态判断。判定事件发生的时态主要依据是时间短语，还有一些具有时间意义的关键词，如“了”“计划”“正在”。本文采用规则匹配的方法来判断时态。

① 规则词库的建立与扩充

首先，从训练语料中选取符合表 1 情况的词语放入规则词库中。然后，采用补充学习来扩充规则库。从同义词词典里，提取出相关时间短语和具有时间语义的关键词，并把这些短语和词加进规则库。

表 1 规则库

序号	规则类型	规则（部分）	例子	时态
1	时间短语	去年，过去	去年，我也去过玉渊潭。	PAST
2		现在，此时此刻	此时此刻，我好想你!	PRESENT
3		明天，不久以后	我承诺：不久以后，我一定会回国看你的。	FUTURE
4	具有时间意义的关键词	了，着，过	我的心跟着他一起去了北京。	PAST
5		正在	我正在和我最心爱的人约会，好开心呀!	PRESENT
6		估计，计划，就要，会	这次出差结束之后，我就要恢复自由啦!	FUTURE

②判断过程

由于可能存在同时符合以上 2 种或者 2 种以上规则的现象，需要对规则进行排序。

判断过程如下：

Step1 如果微博内容含有时间短语，那么时态依据为时间短语。当时间短语个数不唯一时，但所指时态一致时，就以该时态作为微博的时态信息，反之时态为 NONE。

Step2 如果没有时间短语，那么时态依据为具有时间意义的关键词。当关键词个数不唯一时，但所指时态一致时，就以该时态作为微博的时态信息，反之时态为 NONE。

1.1.5 用户情感

本文将微博内容中包含的情感关键词个数 n_{ekw} 和表情符号个数 n_{ec} 作为用户情感这一特征的指标。由于表情符号对应的文本形式与用户在微博内容中使用的情感关键词一致，本文可采用统一的关键词匹配方法来一一识别出情感关键词和表情符号。

1.2 逻辑回归检测模型

本文采用逻辑回归检测模型用于检测微博内容中的隐私信息。条件概率 $P(Y=1\mid X)$ 是根据特征变量判断隐私泄露发生的概率，那么逻辑回归检测模型可表示为：

$$P(Y=1\mid X)=\frac{1}{1+e^{-g(x)}}=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}} \quad (4)$$

式中，X 为从微博内容中提取的特征变量，由归一化处理后的 z、n_{pn}、n_a、n_{ap}、n_{rpl}、n_{ll}、n_{tp}、t_s、n_{ekw}、n_{ec} 组成。β_0 为截距项，$\beta=(\beta_1,\beta_2,\cdots,\beta_n)$ 为自变量 X 回归系数，可通过训练集对模型进行训到。

2 实验结果分析

2.1 数据集获取

本文所用数据集是通过调用新浪微博的公共微博 API 得到的，对其进行去除助词、停用词和符号等预处理，得到 2016 年 1 月 5 日 10:00 至 2016 年 1 月 20 日 10:00 共计 1 140 174 条公共微博数据。根据接口返回的 geo 字段，剔除没有地址标识（即 geo 字段值为“null”）的微博，最终得到 262 240 条带有地理标识的微博数据作为本实验的数据集。该数据集都是抓取时间以前 300 s 以内的微博，尽管经过剔除处理，但仍具有一定的时间连续性，且不限地域

和用户,因此有一定的代表性。

2.2　评价指标

对微博内容隐私进行建模就是建立一个分类器对微博内容的所属类别进行分类。评价分类器的性能,通常采用如表 2 所示的混淆矩阵表示。

表 2　分类器分类结果

实际类别	预测类别	
	类别 A	类别 B
类别 A	a	b
类别 B	c	d

性能评价指标通常有如下 4 个:

正确率 $T=(a+d)/(a+b+c+d)\times 100\%$;

查全率 $R=a/(a+b)\times 100\%$;

查准率 $P=a/(a+c)\times 100\%$;

漏检率 $M=b/(a+b)\times 100\%$

2.3　结果与分析

2.3.1　显著性检验和参数估计

为方便起见,本文利用 IBM SPSS Statistic 19.0 软件中的逻辑回归分析模块对模型进行训练。其中,显著性检验结果如表 3 所示。

表 3　显著性检验结果

x	B	S.E	Wals	df	Sig	Exp(B)
z	8.455	0.392	12.209	1	0.001	0.934
n_{pn}	6.781	0.436	26.327	1	0.001	0.582
n_a	2.322	0.273	18.237	1	0.000	0.037
n_{ap}	4.503	2.572	29.273	1	0.001	0.374
n_{rpl}	30.278	4.483	41.664	1	0.000	8.621
n_{ll}	11.036	1.694	50.847	1	0.000	1.235
n_{tp}	9.263	0.971	30.289	1	0.000	0.234
t_s	8.285	0.835	30.283	1	0.001	0.723
n_{ekw}	3.119	0.409	2.384	1	0.001	0.384
n_{ec}	1.003	0.086	22.252	1	0.004	0.045
β_0	2.145	0.264	19.328	1	0.000	0.234

从表 3 中我们可以看出,除了表情符号的个数 n_{ec} 这一变量的 Sig 值为 0.004,勉强通过显著性检验之外,其余变量的 Sig 值均小于临界值 0.05,说明模型所选的自变量对微博内容类别的影响是显著的,每一个变量都可以作为模型的最终变量。

B 值是各变量 x 的回归系数,也是 β 的参数估计值。从参数结果来看,所有的参数估计值都为正,说明活动主题值越大,人名(或机构名)、“@”、绝对物理地点、相对物理地点、逻辑位置、时间短语、时态信息、情感关键词和表情符号的个数越多,其置信度越高,符合之前的分析效果。相对物理点的个数 n_{rpl} 的 B 值最高,说明它对模型结果的影响最大。表情符号的个数 n_{ec} 的 B 值最小,则说明它对模型结果的影响最小。

2.3.2　分类结果分析

采用正确率 T 作为阈值 θ 选择的一个评判指标,不同 θ 下 T 的变化趋势如图 1 所示。

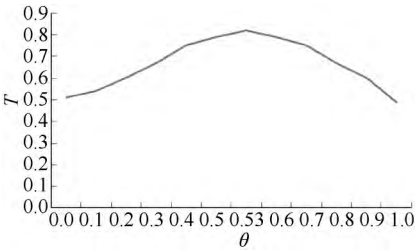

图 1　T 与 θ 关系

图 1 表明,在 $\theta=0.53$ 时,该指标的值最大,说明此时隐私检测的正确率最高,约为 0.82。因此,本文将 θ 设为 0.52,即 $P(Y=1\mid X)\geqslant 0.52$ 时就判定为隐私微博,否则为非隐私微博。分类结果如表 4、5 所示。

表 4　逻辑回归检测模型分类效果

实际类别	预测结果		
	隐私微博	非隐私微博	准确率/%
隐私微博	134 535	26 585	83.50
非隐私微博	18 202	82 918	82.00
总计准确率	—	—	82.92

表 5　评价结果　　%

正确率 T	查全率 R	查准率 P	漏检率 M
82.92	83.50	88.08	16.50

从表 4 分类结果可知,数据集里共有 161 120 条隐私微博,被判定为隐私的微博有 134 535 条,预测准确率为 83.50%。共有 101 120 条非隐私微博,被判定非隐私微博的有 82 918 条,预测准确率为 82.00%,说明模型对每一类的预测能力都比较强,且对隐私微博的判定性能要高一些。从表 5 评价结果可以看出,模型的正确率 T、查全率 R、查准率 P 都超过了 80%,而漏检率 M 则低于 20%,说明模型性能不错,可靠性较强。

另外,本文用 JAVA 编码实现基于贝叶斯的检测方法[6],在本文数据集上进行测试,对比结果如图 2 所示。基于贝叶斯的检测方法主要依据的是特定类型隐私的关键词,有很大的局限性,所以隐私识别结果明显低于本文方法。

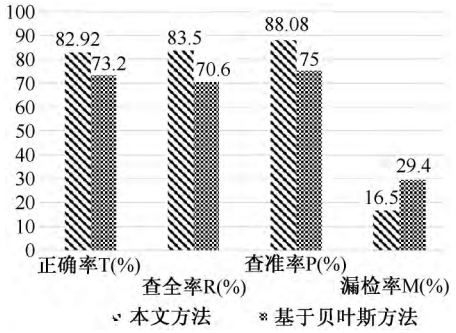

图2 微博隐私检测方法识别效果对比

3 结束语

本文从分析微博内容的5大特征出发,运用逻辑回归算法,提出一种基于逻辑回归的微博内容隐私检测模型。从对比实验结果来看,该模型对隐私微博有着较好的识别效果,能够起到有效检测微博内容隐私的作用。但是每个用户对隐私信息的关注程度不一样,如有些用户出于某一需求或者个人偏好,经常在微博内容中附加地址标识。该用户并不将地址信息作为自我隐私的一部分,导致本文提出的隐私检测模型不适应。因此,如何量化用户对微博内容中暴露的隐私的关注程度将成为我们下一步需要研究的工作。

参考文献:

[1] 曹玖新,吴江林,石伟,等. 新浪微博网信息传播分析与预测[J]. 计算机学报,2014(4):779-790.

[2] Mao H, Shuai X, Kapadia A. Loose tweets: an analysis of privacy leaks on twitter [C] // Acm Workshop on Privacy in the Electronic Society. 2011:1-12.

[3] 日月草. 微博晒行踪小心被贼盯[J]. 互联网天地,2010(3):94-95.

[4] 耿延庭. 微博普通用户主动公开隐私现象分析—以新浪微博为例[J]. 新闻世界,2014(2):163-164.

[5] Li Y, Li Y, Yan Q, et al. Privacy leakage analysis in online social networks [J]. Computers & Security, 2015, 49:239-254.

[6] 江智双. 一种基于贝叶斯的微博隐私检测方法研究[D]. 哈尔滨:哈尔滨工程大学,2013.

[7] 吴友政. 汉语问答系统关键技术研究[D]. 北京:中国科学院自动化研究所,2006.

[8] 谭同超. 有限状态机及其应用[D]. 广州:华南理工大学,2013.

(上接第21页)

on Advanced Intelligent Mechatronics, Montreal, 2010:652-657.

[6] Kuo A D. Energetics of actively powered locomotion using the simplest walking model [J]. Journal of Biomechanical Engineering, 2002, 124: 113-120.

[7] Tang Z, Yang M, Pei Z. Self-adaptive PID control strategy based on RBF neural network for robot manipulator [C] // 2010 First International Conference on Pervasive Computing Signal Processing and Applications, Harbin: IEEE, 2010: 932-935.

[8] Guo J, Wu G, Guo S. Fuzzy PID algorithm-based motion control for the spherical amphibious robot [C] // 2015 IEEE International Conference on Mechatronics and Automation, Beijing, 2015: 1583-1588.

[9] Hu Z, Guo L, Wei S, et al. Design of LQR and PID controllers for self-balancing unicycle robot [C] // international conference on information and automation, Hailar, 2014:972-977.

[10] 张化光,张欣,罗艳红,等. 自适应动态规划综述[J]. 自动化学报,2013,39(4):303-311.

[11] 林小峰,宋绍剑,宋春宁. 基于自适应动态规划的智能优化控制[M]. 北京:科学出版社,2013:118-124.

[12] 张奇志,周亚丽. 双足机器人半被动行走固定点全局稳定性分析[J]. 工程力学,2013,30(3): 431-436.

[13] Zhang Qizhi, Chew CheeMeng, Zhou Yali, et al. Iterative learning control for biped walking [C] // 2010 International Conference on Mechatronics and Automation, Xi'an, China, 2010:237-241.