高校区域大学生微博身份的精确识别方法①

姜 赢, 何国东, 郭雨宸, 朱玲萱

(北京师范大学珠海分校 管理学院, 珠海 519087)

摘 要: 对高校大学生微博身份进行精确识别有利于尽早的定位大学生网络谣言、高校舆情事件的起源, 为高校辅导员及相关管理部门采取线下补救措施、及时处理突发事件争取时间. 以学校提供的学生信息资料为背景, 让挖掘到的大学生微博信息尽可能地去匹配已有的背景信息, 从而识别高校区域大学生微博帐号. 分别采用 3 种阈值进行实验分析, 证明这种循环匹配的方法可以获得较好的识别效果.

关键词: 网络舆情; 微博帐号; 身份识别; 模式匹配; 学生微博

Accurate Identification Method of College Student Microblogs in Certain Area of University

JIANG Ying, HE Guo-Dong, GUO Yu-Chen, ZHU Ling-Xuan

(School of Management, Beijing Normal University, Zhuhai 519087, China)

Abstract: The accurate identification of college student microblogs is helpful to location the sources of college student rumors and university public opinion events early, which can gain time for the university tutors and related authorities to take remedial measures and deal with emergencies promptly. It matches the college student microblogs information with the student background information from the university as much as possible, so that the student microblog accounts can be identified in certain area of university. The experiments are performed on three different threshold values, and the results prove the effectiveness of the identification by this loop matching method.

Key words: online public opinion; microblog account; identification recognition; pattern matching; student microblog

在高校的微博社区中, 大学生使用微博的频率极高, 大多数的大学生都会利用微博来传达信息. 获取和分析高校大学生的微博信息有助于了解大学生学习生活状况, 以便更好地做好学生管理工作. 然而, 获取大学生微博信息的首先要从茫茫的微博“大海”中识别出大学生微博帐号. 有些高校要求学生入学时提供手机、电子邮箱信息, 其中也包括微博帐号. 从这些帐号获取微博信息虽然能解决一部分问题, 但是目前大学生个人隐私保护观念普遍较强, 并不愿意毫无保留的提供给高校专业老师、辅导员个人私密信息. 笔者在前期研究中也发现部分大学生提供虚假微博帐号给学校, 并另开通一个或多个“小号”的现象[1]. 另外, 大学生微博帐号也会随转专业、换班级、加入/退出社团等交友圈子变化而变动. 因此目前亟待一种能够快速有效识别特定区域范围(例如某个班级)大学生微博帐号的方法(微博身份识别), 才能在此基础之上获取特定群体或个体大学生微博信息并进行分析. 特别对于高校微博舆情监控与引导工作来说, 尽早的精确定位大学生网络谣言、高校舆情事件的起源至关重要. 如果快速能识别出网络舆情相关的大学生微博帐号的真实身份, 就可以立刻采取线下补救措施, 为高校辅导员及相关管理部门及时处理突发事件争取时间, 这也是本文的研究意义所在.

1 研究现状分析

徐强等通过获取微博上的用户以及用户之间的关系作为研究样本, 构建网络社区模型, 并利用 GN 算法对微博用户进行社区划分, 用于社交网络中的社区

① 基金项目:广东省省级学校德育创新项目(2015DYZD015);广东省科技计划(2014A080804001)
收稿时间:2016-04-17;收到修改稿时间:2016-05-16 [doi: 10.15888/j.cnki.csa.005527]

发现[2]. 刘勘等通过随机森林算法设计微博中机器用户的识别模型有效地区分微博中的机器用户和普通用户[3]. 黄磊等将用户名和用户发表的微博文本作为表示用户的样本, 使用基于最大熵算法进行用户分类, 利用认证用户对非认证用户进行类型分类, 能够对个人用户和非个人用户进行自动分类[4]. 刘金宝等结合个人信息、帐号行为及微博内容 3 类特征的识别方法能有效识别自媒体帐号, 不同类别的特征也能够相互补充[5]. 薛云霞根据微博用户产生的相关数据对用户的个体属性进行自动识别, 包括一种基于交互式信息的半监督性别分类方法和一种基于文本和社交信息的半监督年龄回归方法[6]. 张进等提出一种改进的微博炒作账户识别方法,从账户状态、历史微博以及账户邻居 3 个方面对炒作账户的特征进行分析, 构建炒作账户特征集, 并利用数据挖掘中的朴素贝叶斯、支持向量机及 K 最近邻分类等算法对正常账户和炒作账户进行自动分类[7]. 韩忠明等构建了一个识别微博水军的概率图模型计算用户为水军的概率, 能够区分普通用户和水军的属性特征与行为特征, 将用户的属性特征作为识别水军的前提条件, 而行为特征则是判别其是否为水军的验证结果[8]. 赵岩利用僵尸粉发帖的内容特征,将文本复制检测技术应用到博文文本特征分析问题中,提出了一种基于信息指纹的微博文本查重技术, 并利用此技术实现了僵尸粉的识别[9]. 高尚等选取“加 V”和“透露职业”变量作为身份识别标准, 将 2446 个样本聚为五类(群众、学生、打拼族、达人、权威), 并分析了其在人口统计特征、信息公开程度、微博使用痕迹、微博影响力等方面的特征和差异; 又对其中 248 个重度使用者的博文进行了内容分析, 从活动、兴趣、观点完整地描述了五类群体[10]. 国外舆情监控主要是宏观趋势研究, 而微博账户身份识别的个体研究较少: Jalal Mahmud 在 Twitter 平台上推测微博用户的家庭位置, 但是无法精确到用户身份识别[11]. Kapanipathi 使用层级知识库对 Twitter 用户兴趣进行了识别和分类[12], 类似的 Zarrinkalam 也提出基于语义技术的用户兴趣识别技术[13], 也都未能精确定位到用户身份.

综上所述, 目前关于微博身份识别的研究主要是利用微博账户信息、微博内容信息, 采取数据挖掘、机器学习等方法对微博账户进行属性定性(例如, 僵尸粉、水军等)和特征分类(例如, 性别、兴趣、身份、社区等). 这些研究都本质上都是只是“模糊分类”, 无法做到“精确定位”到个人身份. 本文的研究目标不仅仅要识别大学生微博账户所属的高校区域范围(例如, 所属哪个班级、哪个社团), 还力求精确定位到单个学生的身份. 也就是说, 给定某个微博帐号之后, 要能识别出他到底是哪个学生. 然而目前基于大学生微博的高校教育相关研究也主要还是在已获得大学生微博帐号身份之后再做进一步分析(例如, 微博社交网络中的学生用户抑郁症识别方法[14]), 尚未见精确识别单个大学生微博帐号身份的相关研究报道.

2 技术原理

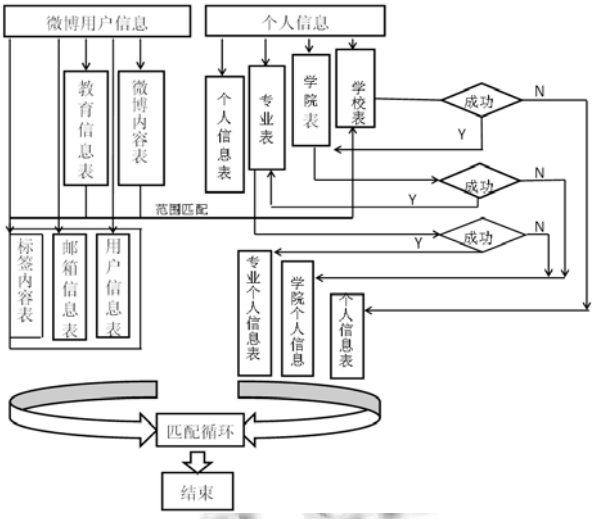

图 1 技术原理图

如图 1 所示, 高校区域大学生微博身份精确识别方法的技术原理为: 以学校提供的学生信息资料为背景, 让挖掘到的大学生微博信息尽可能地去匹配已有的背景信息, 最后计算匹配度作为识别的结果. 第一部分为范围匹配, 数据来源于教育信息表和微博内容信息表: 教育信息表含有微博用户的教育信息, 例如学校名称、学院名称以及专业名称等; 微博内容信息表是利用分词技术, 挖掘相关的关键词, 若在教育信息中, 微博用户没有填写, 那么将在微博内容中尽量提取有关于教育信息的内容, 内容关键词涉及到学校名称、学院和专业名称. 利用以上的两张微博用户表, 首先与个人信息中的学校表、学院表进行匹配, 若与个人信息表中的信息匹配不成功, 将返回专业匹配参数以及让微博用户信息与背景信息进行循环匹配, 直

至循环匹配结束为止. 这样的匹配方法是为了让微博信息表与最小范围的个人信息表进行循环匹配, 对目标进行尽可能的排除操作, 目的是为缩小对象范围, 提高步骤检索效率, 从而提高匹配效率.

3　微博信息挖掘方法

3.1 微博信息存储设计

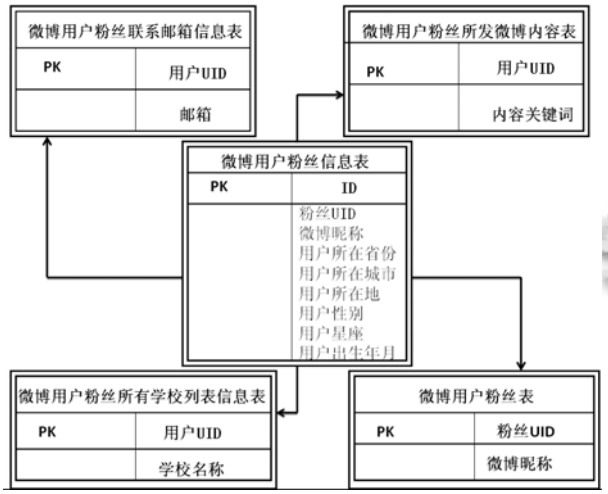

图 2　微博数据表实体关系图

如图 2 所示, 根据本次研究的主题以及匹配机制原理, 笔者将需要被挖掘的微博信息种类分为以下 5 类: (1)粉丝列表类: 授权用户的粉丝列表信息, 获取的信息有粉丝 UID、以及粉丝昵称、出生年月日、籍贯、性别等信息. (2)粉丝微博内容类: 获取粉丝近期发表的 100 条微博内容(新浪开放平台限制, 最多下载条数为 100 条), 对粉丝微博内容进行分词、提取关键词, 例如: “信息管理与信息系统”、“信管”、“人力”、“人力资源”等. (3)标签类: 粉丝微博的标签上, 大多数微博用户会设置个人特色标签, 例如: “90 后”、“星座信息”、“爱好”等; 获取标签信息后采用分词技术提取星座等关键词. (4)教育类: 获取粉丝的教育信息, 在注册微博用户过程中需要填写目前教育情况, 因此, 通过该接口可以提取用户教育情况, 一般可以提取学校名称、年级信息以及学院名称. (5)邮箱类: 获取用户填写的邮箱信息, 个人资料中存在学生邮箱信息, 与此匹配可以提高识别率.

如图 2 所示, 笔者将微博信息数据库模式采用为星型模式, 原因在于用于系统运行的是一张巨大的微博信息事实表, 因此, 为了提高灵活性以及代码易开发性, 本文将微博信息数据库模式采用微星状模式; 再者, 由于微博信息数据库中存在的数据已经进行过初步的数据预处理, 考虑到不需要在多次进行数据预处理环节, 因此采用星状模式关系数据库是最佳设计方案.

3.2 微博信息获取方式

笔者通过新浪微博 API 授权方式进行微博信息的挖掘, 在取得授权码的前提下, 在平台开放包进行信息挖掘操作. 笔者只能获得新浪微博普通权限, 受微博系统限制比较多, 因此在微博信息采集的策略上, 分为多帐号采集方式、代理 IP 采集方式. 多帐号采集方式是找到多个学校官方的微博账户, 对这些官方帐号进行授权, 获取相应的授权码; 代理 IP 采集方式, 微博系统对单一个的 IP 请求是受限制的, 那么可以采用代理 IP 突破限制, 对于普通权限, 每小时 30000 次的下载次数, 对于本次研究数据量要求是足够的.

3.3 微博信息预处理

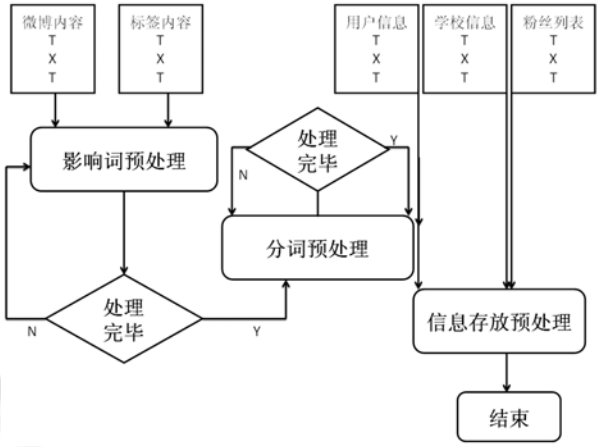

图 3　微博信息预处理流程图

被挖掘的微博信息的特点有杂乱无序和具有太多无规律的标签符号, 因此在系统利用数据之前, 需要对微博信息进行数据预处理操作. 高校微博用户以学生为主, 每所高校仅在校生数量就有几千到几万不等, 微博内容更是海量存在. 如果数据预处理完全靠人工手动处理, 将花费大量的时间和人力物力, 不具有可行性和可推广性. 本次研究将按照系统利用数据阶段把微博信息数据预处理分为系统调用前数据预处理和系统内运行时数据预处理, 所有的数据预处理操作都在系统内进行, 不利用人工手动处理.

(1) 影响词预处理方式是指将一些无关紧要的词

进行删除的, 需要利用这一方式进行预处理的信息由微博内容、标签内容. 本文将微博内容、标签内容进行影响词处理的原因在于微博内容以及标签内容所含有的英文、符号等标签过多, 尽量去除该类标签对于匹配精准度取到关键性作用.

(2) 分词预处理是将经过影响词预处理的微博内容和标签内容进行分词, 对研究中所需的关键词进行提取. 微博用户在编辑微博内容时可能会涉及到某些具有身份信息的内容, 例如, “在人力资源的专业课上, 我收获良多. ”, “人力资源”能够进行身份定位的关键词; 在标签关键词的提取方面可以提取类似于星座等关键词.

(3) 信息存放预处理指的是将微博数据按照一定的格式(位置关系等)存放在 txt 文档中, 该步骤的目的在于将微博数据存储在数据库上时易于提取, 代码的开发难度将小, 能够减少程序员的开发工作, 提高工作效率.

4 实验分析

笔者以北京师范大学珠海分校管理学院 597 名学生作为实验分析对象. 根据学院提供的学生正确个人信息, 在实验过程中, 笔者将这些个人信息分成模糊信息、较唯一信息和唯一信息; 其中模糊信息中包括了个人姓名、性别、星座、出生地、学校名称、学院名称、专业名称、年级等; 较唯一信息包括 QQ 帐号、出生年月日等; 唯一信息则含有身份证号、学号、手机号码等.

4.1 实验参数设置

笔者对模糊信息、较唯一信息、唯一信息所设置的参数是不一样的, 唯一信息的参数高, 三者之中参数最低的是模糊参数; 然而在模糊信息中的信息参数也不相同, 比如姓名参数高于性别、星座等信息, 主要是根据信息在微博人群中出现的概率大小确定的. 因此本次实验将匹配参数的设置类型分为 3 个阶段, 每个阶段的匹配参数作为准确率和召回率的阀值. 3 个阶段的阀值为参数的标准差, 参数表由下表所示.

(1) 较唯一信息匹配参数总和大于 60%

阀值 1 中较唯一信息匹配参数总和大于 60%的设定原因在于设定 60%以上匹配度为匹配合格线, 因此, 考虑信息重要程度以及出现频数讲 qq 帐号、出生年月以及微博名设为各 20%(阀值 1 号).

(2) 模糊信息考虑出现频数

阀值 2 为模糊信息考虑出现频数, 越重要出现的频数越高则设置的匹配参数越高, 但是总的模糊信息匹配参数总和不超过 60%.

(3) 不考虑出现频数与重要性

阀值 3 为不考虑任何出现的频率和重要性, 阀值 3 主要被作为参考参数值.

表 1 个人信息匹配参数表(%)

名称	个人信息												
	模糊信息							较唯一信息			唯一信息		
阀号值	个人姓名		性别	星座	出生地	学校名称	专业名称	QQ 帐号	出生年月日		身份证	学号	手机
1	姓	名	7	8	4	6	0	20	年	月日	0	0	0
	7	8							20	20			
2	4	4	20	15	7	20	0	15	5	10	0	0	0
3	11	11	11	11	11	12	0	11	11	11	0	0	0

表 2 循环匹配示例(10 号微博用户与 13 号学生)

微博用户	微博信息	学生信息	**个人信息**	参数返回(%)
微博名	ben 黄志军	姓名	黄志军	15
邮箱	88197886@qq.com	qq	88197886	20
qq	88197886			
性别	男	性别	男	7
星座	巨蟹座	星座	巨蟹座	8
出生地	广东珠海	出生地	福建	0
出生年	1991	出生年	1991	20
出生月日	711	出生月日	711	20
在读院校	Null	在读信息	北师大珠海分校管理学院信息管理与信息系统 1 班	0
毕业院校	Null			
合计				90

在个人姓名信息中, 主要与微博昵称进行匹配, 将姓名拆分成姓、名各自添加匹配参数, 拆分的原理在于微博用户采用真实姓名作为微博昵称; 在出生年月日信息方面, 同样采取与个人姓名的拆分方式, 拆分成出生年、出生月、出生日, 拆分的原理主要根据微博用户填写信息不完整的可能. 以上参数的设置存在匹配重要性高低之分, 在模糊信息分类中, 重要性: 个人姓名>出生地=专业名称>学院名称=星座>年级>学校名称=性别, 因此在参数设置上依据重要性由高到低设置. 在较唯一信息中, QQ 帐号信息的重要性等于出生年月日重要性. 唯一信息中, 因为新浪微博系

统不披露关于微博用户的绝密信息, 因此唯一信息不再本次研究的范围之内.

例如, 依据表 2 的匹配数据, 采取阈值号为 1 的参数进行匹配, 可得 10 号微博用户与 13 号学生的匹配度为 90%, 直到匹配完所有的学生; 最后返回所有学生的匹配度.

4.2 实验样本基本情况

本次研究的微博样本数量有 597 人, 统计的学生微博信息样本完整度情况如图 4 所示. 在个人微博资料中, 邮箱以及 qq 填写的情况是最少的, 597 人中只有 86 人和 82 人填写; 性别资料有 481 人填写, 所在地高达 458 人填写, 出生年月日大概有将近 177 人填写, 毕业院校的填写情况超 304 人. 从以上的数据可以看出, 对于微博社交平台上, 较唯一的信息填写的相对较少, qq 号码占总人数的 14.4%, 出生年月日占总人数的 30.1%; 于是这些因素都会对系统的匹配度有所影响, 根据所挖掘的信息, 必须对匹配参数做出相应的调整.

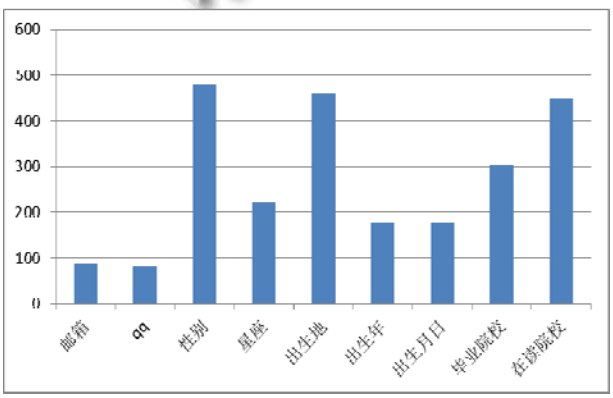

图 4　微博样本数据完整度图

较唯一信息被学生填写的次数较少, 而模糊信息被学生填写的次数较多. 如果根据定量来分配匹配参数显然是不正确的方式, 也就是说, 按照出现的频数来分配匹配参数, 性别列被填的次数最多, 但是确实最为模糊和能够被匹配上的参数, 因此只是利用定量的方法来确定参数是存在错误的, 必须引入定性变量. 获取的微博基本信息存在重要性程度不同, 例如将所有的基本信息分为唯一信息、较唯一信息和模糊信息; 唯一信息中只要能够匹配上一列, 即可对用户进行准确定位. 如果较唯一信息中, 如初生年月日、qq 号码可能会出现少量的重复和误填情况, 而模糊信息则重复出现的概率更大, 例如性别只分为男和女、出生地也可能重复. 在同等性质的信息中也可以分出不同的重要性, 例如在模糊信息中, 重复出现的概率较大的为性别和星座; 在较唯一信息中, 出生年月比 qq 号码重要.

4.3 实验结果

表 3　匹配度>60%的识别准确率和召回率表

阈值号	参数标准差	识别人数	未识别人数	匹配度>60%的总匹配数	准确率(%)	召回率(%)	最高匹配度(%)
1	7.0671	44	15	74	74.6	59.5	90
2	6.5659	59	0	906	100.0	6.51	82
3	0.3333	46	13	88	78.0	52.2	81

表 4　匹配度>80%的识别准确率和召回率表

阈值号	参数标准差	识别人数	未识别人数	匹配度>80%的总匹配数	准确率(%)	召回率(%)	最高匹配度(%)
1	7.0671	12	47	12	20.30	100	90
2	6.5659	37	22	51	62.70	72.50	82
3	0.3333	7	52	7	11.90	100	81

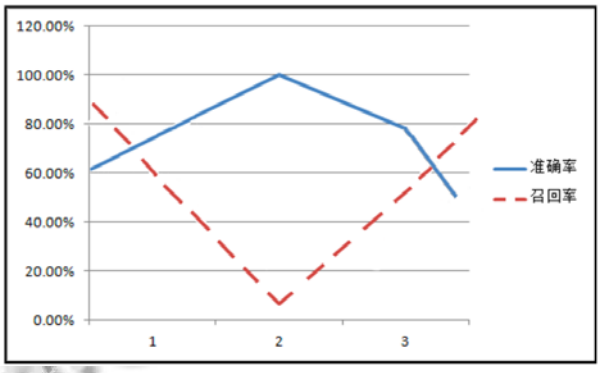

图 5　准确率与召回率比值曲线图(匹配度>60%)

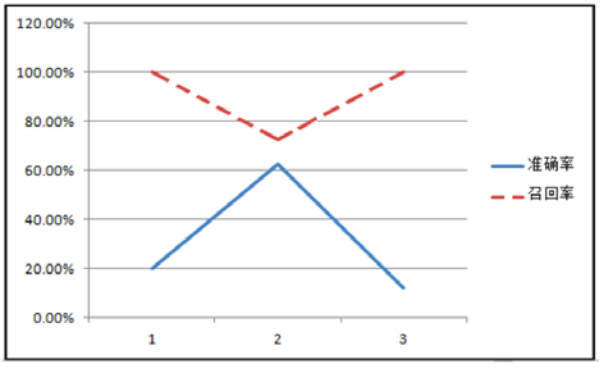

图 6　准确率与召回率比值曲线图(匹配度>80%)

从图 5 和图 6 可以看出, 准确率随着参数标准的关系为开口向下的二元一次方程关系, 而召回率与标

准差的关系为开口向上的二元一次方程关系, 在某一标准差下准确率和召回率分别取得最大值. 在第三号阀值中匹配度大于 60%和 80%的准确率和召回率最为接近; 准确率和召回率是一对相对矛盾的性能指标, 在图 5 和图 6 中可以看出, 准确度和召回率在 70%处相交, 但此时的准确度和召回率并不是本系统的最佳数值, 应该在一定的召回率的基础上提高识别准确度, 因此应该以准确度为重.

4.3 局限性分析

根据以上分析, 不难看出本方法的主要局限性在于: (1)抓取的微博内容中大部分学生的基本信息残缺, 由于基本信息的残缺导致了的匹配度较低. (2)本次研究的匹配原理采用的匹配方法过少: 本文采用的匹配方法为基本信息的绝对匹配原理, 由于匹配方法的单一导致匹配结果良莠不齐.

5 未来工作

(1) 模糊匹配: 对微博名等信息进行模糊匹配, 本文采用的微博名匹配方法为绝对匹配(匹配值为 100%), 因此存在着不足. 为了提高匹配值, 增加采用模糊匹配的方法, 例如, 微博名为“Jiang 浩”, 真实姓名为“姜浩”. 为了增加匹配值, 应该对微博名进行英文和汉字的转换、以及汉语拼音首字母的提取等操作, 利用以上操作后的结果进行匹配来增大匹配值. 例如, 模糊匹配可以利用 LCS(最长公共子序列)或 GTS(贪婪串匹配)来计算拼音字符串的匹配值[15]. 模糊匹配计算的结果是一个 0%～100%之间匹配值, 不是一个绝对的错误匹配或正确匹配. 匹配值可设置阈值范围(如 80%以上)用作匹配参数设定, 也直接用于每次循环匹配的加权参数设定.

(2) 信息内容匹配细化: 信息内容匹配细化是对信息内容进行细化匹配, 目的是让微博内容尽可能匹配用户信息增大匹配值. 信息匹配细化的原理是利用 Lucene 的搜索技术代替匹配值, 例如一名微博用户的出生地为四川绵阳, 真实出生地为四川省绵阳市, 本应该是正确的, 如果利用绝对匹配原理, 将匹配不上, 那么应该利用 Lucene 的匹配原理对该条信息进行匹配, 自定义一个返回的匹配度并返回匹配度作为匹配参数.

参考文献

1 姜赢,万里鹏,张婧,葛思坤.微博环境下高校网络舆情的监测与引导研究——以政治敏感信息的监测与引导为例.现代教育技术,2013,4:92–96.

2 徐杨,蒙祖强.基于 GN 算法的微博社区识别方法.广西大学学报(自然科学版),2013,6:1413–1417.

3 刘勘,袁蕴英,刘萍.基于随机森林分类的微博机器用户识别研究.北京大学学报(自然科学版),2015,2:289–300.

4 黄磊,李寿山,王晶晶.基于认证用户信息的微博用户类型识别方法.计算机科学与探索,2015,6:719–725.

5 刘金宝,盛达魁,张铭.微博自媒体帐号识别研究.计算机研究与发展,2015,11:2527–2534.

6 薛云霞.微博用户属性识别方法研究[硕士学位论文].苏州:苏州大学,2015.

7 张进,刘琰,罗军勇,董雨辰.基于特征分析的微博炒作账户识别方法.计算机工程,2015,4:48–54,59.

8 韩忠明,许峰敏,段大高.面向微博的概率图水军识别模型.计算机研究与发展,2013,S2:180–186.

9 赵岩.微博僵尸粉识别技术研究与实现[硕士学位论文].长沙:国防科学技术大学,2013.

10 高尚,林升栋,翁路易,梁玉麒,宋玉蓉,赵成栋.基于身份识别对中国微博活跃用户的分群研究.现代传播(中国传媒大学学报),2013,10:116–121.

11 Mahmud J, Nichols J, Drews C. Home location identification of twitter users. ACM Trans. Intell. Syst. Technol., 2014, 5(3): 1–47.

12 Kapanipathi P, Jain P, Venkataramani C, Sheth A. User interests identification on Twitter using a hierarchical knowledge base. The Semantic Web: Trends and Challenges, ESWC 2014. Lecture Notes in Computer Science, 2014, 8465: 99–113.

13 Zarrinkalam F, Fani H, Bagheri E, Kahani M, Du W. Semantics-enabled user interest detection from twitter. Proc. of 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2015. 103–110.

14 李鹏宇.微博社交网络中的学生用户抑郁症识别方法研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2014.

15 于海英.字符串相似度度量中 LCS 和 GST 算法比较.电子科技,2011,24:101–103,124.