(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106886518 A
(43)申请公布日 2017.06.23

(21)申请号 201510933139.6

(22)申请日 2015.12.15

(71)申请人 国家计算机网络与信息安全管理中心
地址 100029 北京市朝阳区裕民路甲3号
申请人 北京航空航天大学

(72)发明人 董元魁 陈训逊 郎波 王博 王洋 黄亮

(74)专利代理机构 深圳市威世博知识产权代理事务所(普通合伙) 44280
代理人 陈雪梅

(51)Int.Cl.
G06F 17/30(2006.01)
G06Q 50/00(2012.01)
G06K 9/62(2006.01)

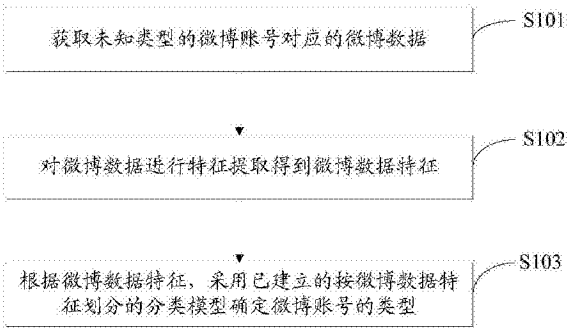
权利要求书1页 说明书15页 附图8页

(54)发明名称

一种微博账号分类的方法

(57)摘要

本发明公开了一种微博账号分类的方法，方法包括：获取未知类型的微博账号对应的微博数据，对微博数据进行特征提取得到微博数据特征，根据微博数据特征，采用已建立的按微博数据特征划分的分类模型确定微博账号的类型。通过上述方式，本发明能够准确区分普通账号跟异常账号，并且还能够识别出账号具体属于哪一类型的账号。



CN 106886518 A

1.一种微博账号分类的方法，其特征在于，所述方法包括：

获取未知类型的微博账号对应的微博数据；

对所述微博数据进行特征提取得到微博数据特征；

根据所述微博数据特征，采用已建立的按微博数据特征划分的分类模型确定所述微博账号的类型。

2.根据权利要求1所述的方法，其特征在于，所述微博账号的类型为僵尸账号、广告账号、机器账号或普通账号中的一种。

3.根据权利要求2所述的方法，其特征在于，所述在获取微博账号对应的微博数据之前，还包括：

获取已知类型的微博账号的微博数据；

对所述已知类型的微博账号的微博数据进行特征提取，得到已知类型的微博账号的微博数据特征；

对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型。

4.根据权利要求3所述的方法，其特征在于，所述对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型包括：

通过10折交叉验证的方法，对所述已知类型的微博账号的微博数据进行机器学习训练，从而建立按微博数据特征划分的分类模块。

5.根据权利要求3所述的方法，其特征在于，所述对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型包括：

分别采用随机森林、朴素贝叶斯和梯度回归分类算法中的至少一种，对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型。

6.根据权利要求5所述的方法，其特征在于，采用随机森林算法对僵尸账号的微博数据特征进行机器学习训练；采用朴素贝叶斯算法对广告账号的微博数据特征进行机器学习训练；采用梯度回归算法对机器账号的微博数据特征进行机器学习训练。

7.根据权利要求1所述的方法，其特征在于，所述获取微博账号对应的微博数据包括：

通过微博应用程序编程接口或通过网络爬虫的方式获取未知类型微博账号对应的微博数据。

8.根据权利要求1-7任一项所述的方法，其特征在于，所述微博数据特征包括用户资料特征、微博内容特征、交互行为特征和发布行为模式特征中的至少一种。

9.根据权利要求2所述的方法，其特征在于，所述根据所述微博数据特征，采用已建立的按微博数据特征划分的分类模型确定所述微博账号的类型之后，还包括：

通过分类算法对已确定的僵尸账号、广告账号、机器账号以及普通账号进行二分类。

10.根据权利要求9所述的方法，其特征在于，所述通过分类算法对已确定的僵尸账号、广告账号、机器账号以及普通账号进行二分类包括：

通过随机森林分类算法对僵尸账号和其余三种账号集合进行二分类；通过朴素贝叶斯分类算法对广告账号、普通账号以及机器账号的集合进行二分类；以及通过梯度回归分类算法对机器账号和普通账号的集合进行二分类。

一种微博账号分类的方法

技术领域

[0001] 本发明涉及一种微博账号分类的方法。

背景技术

[0002] 在当今互联网高速发展的时代，社交网络日益成为人们互联网生活的重要组成部分，其中微博这种社交网络服务更是成为了最红火的概念。微博是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过WEB、WAP以及各种客户端组建个人社区，以140字以内的文字更新信息，并实现即时分享。

[0003] 由于各类微博对用户的技术门槛都很低，同时微博应用程序编程接口的存在使得用户可以在各种移动终端上登录微博发布消息，这也加速了微博的发展。随着高速发展而来的是各种各样的问题，僵尸账号、广告账号、机器账号及其带来的虚假粉丝、内容污染和舆论误导等问题日益严重。国内外有专家学者已经开始研究异常账号的检测和垃圾内容的过滤技术。微博服务提供商也建立了诸如手机号注册、用户举报等措施来限制异常账号的泛滥。

[0004] 但是，现有的方法中，都是只能识别出异常账号跟普通账号，无法准确识别出账号具体为哪一种类型的微博账号，并且识别效率也相当低。

发明内容

[0005] 本发明主要解决的技术问题是如何提供一种能够高效准确识别微博账号类型的方法。

[0006] 有鉴于此，本发明实施例提供一种微博账号分类的方法，能够准确区分普通账号跟异常账号，并且还能够识别出账号具体属于哪一类型的账号。

[0007] 为解决上述技术问题，本发明采用的一个技术方案是：提供一种微博账号分类的方法，所述方法包括：获取未知类型的微博账号对应的微博数据；对所述微博数据进行特征提取得到微博数据特征；根据所述微博数据特征，采用已建立的按微博数据特征划分的分类模型确定所述微博账号的类型。

[0008] 其中，所述微博账号的类型为僵尸账号、广告账号、机器账号或普通账号中的一种。

[0009] 其中，所述在获取微博账号对应的微博数据之前，还包括：获取已知类型的微博账号的微博数据；对所述已知类型的微博账号的微博数据进行特征提取，得到已知类型的微博账号的微博数据特征；对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型。

[0010] 其中，所述对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型包括：通过10折交叉验证的方法，对所述已知类型的微博账号的微博数据进行机器学习训练，从而建立按微博数据特征划分的分类模块。

[0011] 其中，所述对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而

建立按微博数据特征划分的分类模型包括：分别采用随机森林、朴素贝叶斯和梯度回归分类算法中的至少一种，对所述已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型。

[0012] 其中，采用随机森林算法对僵尸账号的微博数据特征进行机器学习训练；采用朴素贝叶斯算法对广告账号的微博数据特征进行机器学习训练；采用梯度回归算法对机器账号的微博数据特征进行机器学习训练。

[0013] 其中，所述获取微博账号对应的微博数据包括：通过微博应用程序编程接口或通过网络爬虫的方式获取未知类型微博账号对应的微博数据。

[0014] 其中，所述微博数据特征包括用户资料特征、微博内容特征、交互行为特征和发布行为模式特征中的至少一种。

[0015] 其中，所述根据所述微博数据特征，采用已建立的按微博数据特征划分的分类模型确定所述微博账号的类型之后，还包括：通过分类算法对已确定的僵尸账号、广告账号、机器账号以及普通账号进行二分类。

[0016] 其中，所述通过分类算法对已确定的僵尸账号、广告账号、机器账号以及普通账号进行二分类包括：

[0017] 通过随机森林分类算法对僵尸账号和其余三种账号集合进行二分类；通过朴素贝叶斯分类算法对广告账号、普通账号以及机器账号的集合进行二分类；以及通过梯度回归分类算法对机器账号和普通账号的集合进行二分类。

[0018] 本发明的有益效果是：区别于现有技术的情况，本发明通过对微博账号对应的微博数据进行特征提取得到微博数据特征，采用已建立的按微博数据特征划分的分类模型确定微博账号的类型。由于分类模型是通过对数量庞大的已知类型微博账号对应的微博数据进行机器学习训练而得到，因此，分类模型非常全面和具有代表性，通过分类模型对微博账号的类型进行确定，从而可以对微博账号的识别和分类能够更加高效和准确。

附图说明

[0019] 图1是本发明实施例提供的一种微博账号分类的方法的流程图；

[0020] 图2是本发明实施例提供的建立按微博数据特征划分的分类模型的

[0021] 流程图；

[0022] 图3a是性别特征分析示意图；

[0023] 图3b是头像特征分析示意图；

[0024] 图3c是简介特征分析示意图；

[0025] 图3d是昵称特征分析示意图；

[0026] 图3e是微博书CDF图；

[0027] 图3f是粉丝数CDF图；

[0028] 图3g是粉丝度CDF图；

[0029] 图3h是粉丝关注比CDF图；

[0030] 图4是微博内容特征分析示意图；

[0031] 图5a是原创微博数CDF图；

[0032] 图5b是评论数CDF图；

[0033] 图6是发布行为特征分析图示意图；

[0034] 图7是特征重要度对比图示意图；

[0035] 图8是用户成分分析示意图；

[0036] 图9是本发明实施例提供的微博账号分类的装置的结构示意图。

具体实施方式

[0037] 请参阅图1，图1是本发明实施例提供的一种微博账号分类的方法的流程图，如图所示，本实施例的微博账号分类的方法包括以下步骤：

[0038] S101：获取未知类型的微博账号对应的微博数据。

[0039] 本发明实施例中，微博数据的获取可以采用微博应用程序编程接口(Application Programming Interface，API)和网络爬虫两种方法。但微博API接口对访问频率和属性获取有较大限制。因此作为本发明的优选实现方案，采用网络爬虫的方式获取微博数据。基于网络爬虫原理实现完成了微博爬虫工具，该爬虫工具能够获得微博页面上所有能呈现出的所有微博数据，并且将获得的原始微博数据进行预处理，最终存入数据库。

[0040] 在具体实现时，微博数据的获取除了完成基本属性值数据的获取，同时获取每个账号的最新500条微博，若微博数不足500条的，将其所有微博内容全部获取。爬取过程可以采取多台计算机分担微博数据爬取任务，避免爬取时间造成的属性差异。

[0041] S102：对微博数据进行特征提取得到微博数据特征。

[0042] 根据当前微博特点，本发明实施例抽取并扩展出4类微博数据特征：用户资料特征、微博内容特征、交互行为特征和发布行为模式特征，综合考虑多种类型账号特征能够提高账号类型识别准确率。

[0043] 其中，本发明实施例所述扩展出的4类微博数据特征的特征集合请参阅下表1(加※为本发明新提出的特征)：

[0044] 表1：微博数据特征的特征集合

[0045]

	序号	特征	类型
用户资料特征	1	昵称是否含有数字	0/1
	2	是否填写个人简介	0/1
	3	是否有头像	0/1
	4	微龄（天）※	整数
	5	微博数（原创和转发）	整数
	6	粉丝数	整数
	7	粉丝度※	0-1 小数
	8	关注数	整数
	9	关注度※	0-1 小数
	10	粉丝关注比	0-1 小数
微博内容特征	11	含图片的微博比例	0-1 小数
	12	含 URL 的微博比例	0-1 小数
	13	含话题符#的微博比例	0-1 小数
	14	含提及符@的微博比例	0-1 小数
交互行为特征	15	原创数	整数
	16	原创率※	0-1 小数
	17	转发数	整数
	18	被转发数※	整数
	19	转发率※	0-1 小数
	20	评论数※	整数
	21	被评论数※	整数
	22	回复数※	整数
	23	自转数※	整数
发布行为特征	24	微博发布时间间隔熵	小数
	25	平均每天发布微博数※	整数
	26	0-6 点的平均微博数	整数
	27	6-12 点的平均微博数	整数
	28	12-18 点的平均微博数	整数
	39	18-24 点的平均微博数	整数
	30	发布平台数	整数
	31	发布 IP 数※	整数
	32	发布 ISP 数※	整数
	33	发布省份数※	整数
	34	发布城市数※	整数
	35	最近 500 条微博内容	文本

[0046] S103:根据微博数据特征,采用已建立的按微博数据特征划分的分类模型确定微博账号的类型。

[0047] 其中,本发明实施例中的微博账号的类型为僵尸账号、广告账号、机器账号或普通账号中的一种。

[0048] 分类模型的目的是建立一个能够描述给定账号在账号类型中的出现频次或概率的分布。即利用分类模型,可以确定某一账号为哪个类型 账号的可能性更大。通过分类模型,可以对未知类型的账号进行类型识别与区分。

[0049] 具体针对一个未知类型账号进行分类时,首先输入该账号的用户身份证明(User Identification,UID),然后通过微博爬虫工具获取其相关数据,基于数值型特征集合生成基于数值型特征集合生成特征向量1和特征向量3,基于用户发布过得微博文本内容,生成

特征向量2，根据特征向量，通过分类模型采用排除法确定账号类型。

[0050] 举例而言，1)使用特征向量1判断是否是僵尸账号，若是，则停止判断，若不是，则继续下一步；2)基于用户发布过得微博文本内容，生成特征向量2；3)使用特征向量2判断是否是广告账号，若是，则停止判断，若不是，则继续下一步；4)使用特征向量3判断是否是机器账号，若是，则停止判断，若不是，则判定为普通账号。

[0051] 为了进一步确保分类的准确性，本发明实施例的方法在通过分类模型初步确定账号类型后，进一步通过分类算法对已确定类型的账号(即僵尸账号、广告账号、机器账号以及普通账号)进行二分类。

[0052] 其中，针对僵尸账号和机器账号的识别均采用抽取出的数值型特征集合构成特征向量，分别通过分类算法进行普通账号和僵尸账号、普通账号和机器账号的二分类。

[0053] 其中，作为本发明实施例的一种优选的实现方案，通过随机森林分类算法对僵尸账号和其余三种账号的集合进行二分类，通过朴素贝叶斯分类算法对广告账号、普通账号以及机器账号的集合进行二分类，以及通过梯度回归分类算法对机器账号和普通账号的集合进行二分类。

[0054] 作为一种优选，采用文本分类的通用方法进行普通账号和广告账号的二分类，以进一步确定账号为普通账号还是广告账号。

[0055] 文本分类要做如下4个预处理动作：

[0056] 1、选择微博广告和非广告文本数据集；

[0057] 2、微博文本预处理：分词、去停用词、建立词袋模型；

[0058] 3、选择文本分类使用的特征向量：词频表征特征权重；

[0059] 4、量化训练数据集和测试数据集文件。

[0060] 其中广告微博内容涉及各种电商卖家广告、代购广告、微商广告等，广告内容类型多样，但其中含有一些共同的明显的营销词汇，比如打折、优惠、包邮、购买、正品、限量等，这些具有区分性的词汇便是文本分类的关键。同时，将所有不具有广告意图的微博内容归为普通用户发布的非广告微博。

[0061] 本发明通过针对僵尸账号、广告账号和机器账号这三种异常账号，结合普通账号样本集，分别做二分类测试，对比不同分类算法分类效果，详见表2-表4。

[0062] 表2：广告账号识别分类算法的分类效果对比

[0063]

	F-score 得分		准确率		召回率	
排名	算法	得分	算法	得分	算法	得分
1	**朴素贝叶斯（高斯型）**	0.9712	**朴素贝叶斯（高斯型）**	0.9918	支持向量机（径向基核函数）	0.9923
2	随机森林	0.9299	支持向量机（线性核函数）	0.9773	逻辑回归	0.9862
3	支持向量机（线性核函数）	0.9284	K 近邻	0.9590	朴素贝叶斯（多项式型）	0.9812

[0064] 表3:僵尸账号识别分类算法的分类效果对比

[0065]

	F-score		Precision		Recalll	
排名	算法	得分	算法	得分	算法	得分
1	**随机森林**	0.9733	逻辑回归	0.9956	朴素贝叶斯（高斯型）	0.9895
2	朴素贝叶斯（伯努利型）	0.9695	朴素贝叶斯（伯努利型）	0.9808	**随机森林**	0.9712
3	K 近邻	0.9670	**随机森林**	0.9767	K 近邻	0.9659

[0066] 表4:机器账号识别分类算法的分类效果对比

[0067]

	F-score		Precision		Recalll	
排名	算法	得分	算法	得分	算法	得分
1	**梯度回归**	0.9655	随机森林	0.9845	朴素贝叶斯（高斯型）	0.9625
2	随机森林	0.9392	**梯度回归**	0.9799	**梯度回归**	0.9516
3	ExtraTrees 算法	0.8919	ExtraTrees 算法	0.9033	决策树	0.8822

[0068] 由上表2-表4的效果对比可以发现，对于广告账号和普通账号的进一步识别中用到的文本分类算法，朴素贝叶斯分类算法效果较好；对于僵尸账号和机器账号的进一步识别中用到的分类算法，随机森林和梯度回归算法更有效。

[0069] 当然，基于以上分类效果对比，在具体应用过程中，可以根据账号类型分别选取准确率(或F-score)最高的3个分类算法对账号进行二分类。

[0070] 经过本发明的方法，可以确定微博账号的分布趋势，图8是本发明实施例对预定数量的账号进行分类后所统计的用户分布示意图。

[0071] 本发明中的分类模型是基于已知类型账号的微博数据不断通过机器学习和训练而得到。本发明实施例进一步提供建立按微博数据特征划分的分类模型的方法。请参阅图2，图2是本发明实施例提供的建立按微博数据特征划分的分类模型的流程图，如图所示，建立按微博数据特征划分的分类模型包括以下步骤：

[0072] S201：获取已知类型的微博账号的微博数据。

[0073] 其中，已知类型的微博账号来源于人工标记或者电商购买的标记样本。人工标记，即手动查看每个微博账号的资料及微博动态来判定账号类型。电商购买，随着微博的盛行，电子商务网站上已经出现了很多微博服务商品，比如可以购买微博粉丝、微博账号，甚至一条微博的转发 量和点赞数都可以买到，其中卖家出售的微博粉丝，其中就是低级粉丝即僵尸账号，高级粉丝即机器账号，通过直接购买粉丝的方式，可以减少大量人力。

[0074] 在本实施例具体实现时，僵尸账号共标记2000个，其中1500个来自两个淘宝卖家的低级微博粉丝，另外500个通过人工标记。标记的依据是：1)无头像或系统默认头像；2)关注数远大于粉丝数；3)微博数较少且无转发和评论；4)用户昵称为简单的字母和数字组合或汉字和数字组合；5)用户资料填写内容少或无。综合考虑以上5个方面来判断一个账号是否为僵尸账号。通过观察微博，发现很多娱乐明星和认证公司(推销商品)的粉丝中存在大量僵尸账号，有的娱乐明星希望通过百万甚至千万级的粉丝数来提高自己的知名度，认证公司希望购买僵尸账号提高粉丝数，从而吸引普通微博用户的关注，所以僵尸账号的收集目标就集中在娱乐明星和认证公司的粉丝列表中。

[0075] 广告账号共标记1000个，全部来自人工标记。标记的依据是：1)微博内容以广告、促销和抽奖等为主；2)用户简介中有店铺链接、微信号或商品介绍；3)微博中的链接多为商品买卖链接。

[0076] 机器账号共标记2000个，其中1500个来自淘宝购买的高级微博粉丝，400个通过人工标记，100个来自相关研究中使用到的机器账号样本。人工标记的依据是：1)微博发布时间规律性强，每隔一定时间发布一条微博；2)微博内容主题是心灵鸡汤、名人名言、笑话、天气、星座运势等，微博内容也可能以广告为主，有很大嫌疑是通过调用现成的语料库来自动发布这些内容微博；3)微博内容重复度高，不同的机器账号可能使用同一些语料库；4)微博发布平台种类少，部分机器账号的微博发布平台能明显的说明使用了第三方软件，如皮皮时光机、云中小鸟、孔明社交管理等。

[0077] 普通账号共标记3000个，全部来自人工标记。标记依据是：1)粉丝数和微博数较多；2)用户头像是真实照片；3)用户资料填写详细；4)微博内容有日常生活气息，如有个人生活内容分享；5)微博有被转发或评论，同时又回复。收集方法：一是从自己的现实好友出发，然后 再判断现实好友的粉丝和关注，接着递归判断粉丝的粉丝和关注、关注的粉丝和关注。而是从热门微博和热门话题下面寻找积极评论和互动的账号。

[0078] 微博数据的获取可以采用微博应用程序编程接口(Application Programming Interface，API)和网络爬虫两种方法。但微博API接口对访问频率和属性获取有较大限制。因此作为本发明的优选实现方案，采用网络爬虫的方式获取微博数据。基于网络爬虫原理实现完成了微博爬虫工具，该爬虫工具能够获得微博页面上所有能呈现出的所有微博数据，并且将获得的原始微博数据进行预处理，最终存入数据库。

[0079] 在具体实现时，微博数据的获取除了完成基本属性值数据的获取，同时获取每个账号的最新500条微博，若微博数不足500条的，将其所有微博内容全部获取。爬取过程可以采取多台计算机分担微博数据爬取任务，避免爬取时间造成的属性差异。

[0080] S202：对已知类型的微博账号的微博数据进行特征提取，得到已知类型的微博账号的微博数据特征。

[0081] 根据当前微博特点，本发明实施例抽取并扩展出4类微博数据特征：用户资料特征、微博内容特征、交互行为特征和发布行为模式特征，综合考虑多种类型账号特征能够提高账号类型识别准确率。其中，不同微博数据特征的特征集合请参阅上述表1(加※为本发明新提出的特征)，在此不再赘述。

[0082] 用户资料特征(表1中1-10号特征)来自用户比较直观的资料信息。其中微博年龄是从账号注册时间到2015年1月1日截止账号存在天数；

[0083] $$粉丝度=\frac{粉丝数}{粉丝数+关注数}，\quad 关注度=\frac{关注数}{粉丝数+关注数}；\quad 粉丝关注比=\frac{粉丝数}{关注数}。$$

[0084] 其中，图3(a)-图3(h)分别示出用户基本特征分析示意图，从图中可知，四种类型账号的男女比例分布较为随机，不具有较好区分性；头像有无、昵称和简介有无填写能够较好的区分僵尸账号和其他类型账号；机器账号由于使用自动化程序控制，所以发布微博数更多，僵尸账号几乎不发布微博；机器账号初期会发布大量某一主题微博，如笑话、星座、美景图片等特定主题类型的机器微博账号吸引了大量的粉丝，其粉丝数 远大于关注数，而僵尸账号关注数远大于粉丝数，广告账号和正常账号则粉丝数和关注数相当。

[0085] 微博内容特征(表1中11-14号特征)根据微博内容中包含的特殊内容抽取得来。

[0086] 其中，图4是微博内容特征分析示意图，从图4可知，机器账号在大量发布微博同时还会较多地@好友，希望好友能够转发该微博或进行评论等，增加机器账号的人为特性。相

反，僵尸账号几乎不@好友。所以@数可以作为区分机器账号和正常账号、僵尸账号和正常账号的特征。

[0087] 交互行为特征(表1中15-23号特征)表示微博账号和其他账号互动情况。图5(a)-图5(b)是交互行为特征分析示意图，从图5a-图5b表明，机器账号由于使用了语料库，几乎不转发微博，大部分为原创微博；80％的机器账号评论数小于150，而大约60％的正常账号评论数都超过500，即正常账号更具有评论交互意向，机器账号要实现自动评论或回复复杂度较大。

[0088] 发布行为特征(表1中24-34号特征)代表微博账号发布行为模式。通过对微博账号的观察发现，大部分机器账号以一定的时间间隔自动发布微博，有的甚至24小时连续定时发布微博，有的会稍有伪装，避开0-6点休息时间发微博。机器账号微博发布时间更有规律，普通账号则显得无规律可循。使用熵率来度量微博用户发布微博时间规律性。

[0089] 随机变量序列X＝{Xi}由一个微博用户所发微博的时间间隔随机变量组成，X_i表示第i条和第i+1条微博之间的时间间隔随机变量序列X的熵记为

[0090] $$H(X_1,\cdots\cdots,X_m)=-\sum_{i=1}^{m}P(x_i)\log P(x_i) \quad (1)$$

[0091] 其中$P(x_i)$是$P(X_i=x_i)$的概率。当已知该序列的前m-1项时，其条件信息熵记为：

[0092] $$CE(X_m|X_{m-1})=H(X_m|X_1,\cdots,X_{m-1})=H(X_1,\cdots,X_m)-H(X_1,\cdots,X_{m-1}) \quad (2)$$

[0093] 用户发微博的时间间隔构成的序列都是有限序列，而信息熵衡量的是一个无穷随机过程，无法直接用来计算有限的序列。引入修正的条件信息熵来解决序列有限性所带来的问题。修正的条件信息熵的公式如下：

[0094] $$CCE(X_m|X_1,\ldots,X_{m-1})=CE(X_m|X_1,\ldots,X_{m-1})+perc(X_m)\cdot EN(X_1) \quad (3)$$

[0095] 其中$perc(X_m)$是在长度为m的序列里面只出现过一次的序列所占的比例，$EN(X_1)$是当m＝1时的信息熵。当序列长度取[2,m]中的不同值时，分别计算出相应的修正条件信息熵的值，最终熵率取其中最小值。如果该账号是机器账号，那它的行为会有一定的规律性，因而其修正条件信息熵的值会较小。与之相反，普通账号的行为随机化程度较高，修正的条件信息熵值也会较大。

[0096] 针对行为模式特征中的信息熵，将机器账号和普通账号的发微博时间间隔序列输入后，利用式(3)得到每位账号用户的修正条件。

[0097] 图6是机器账号和普通账号各自修正条件信息熵的累积分布函数。由图6可知，机器账号的修正条件信息熵明显比普通账号的修正条件信息熵小，说明账号的发微博行为存在较强的规律性，而普通账号的发微博行为比较随机，验证了前面对用户发微博行为的分析结果。

[0098] 针对广告账号的识别，只需要检测账号发布微博内容是否为广告内容即可，所以使用发布微博文本内容这一特征，实际在文本分类中又将这一特征分解为文本特征向量；针对僵尸粉的识别，根据特征分析，选择使用是否有头像、是否填写简介、昵称是否包含数字、粉丝数、关注数、微博数这6个数值型特征即可；针对机器账号的识别根据特征分析，选择使用是否填写简介、昵称是否包含数字、粉丝数、关注数、微博数、微龄、粉丝度、关注度、粉丝关注比、微博含图片数、原创数、$原创率=\frac{原创数}{原创数+转发数}$、转发数、

转发率=$\frac{\text{转发数}}{\text{原创数+转发数}}$、评论数、被评论数、回复数、自转数、微博发布时间间隔熵、日均发布微博数、0-6点的平均微博数、6-12点的平均微博数、12-18点的平均微博数、18-24点的平均微博数、发布平台数、发布IP数、发布ISP数、发布省份数、发布城市数共29个数值型特征。特征数据分析不仅通过条形图、CDF图来展示，还通过具体分类模型计算了特征的重要。

[0099] 图7是用于普通账号和机器账号分类的34个特征的重要性排名前20对比图，通过特征重要度排名，可以进一步进行特征选择，在保证分类准确度的基础上加快账号分类速度。实际应用中可以综合考虑分类准 确度和分类速度两个指标，选择可接受的分类准确度和分类速度。

[0100] S203：对已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型。

[0101] 在具体实现时，可以采用10折交叉验证的方法，使用已标记样本数据集，训练分类模型，通过实验实际测试各个分类算法在微博账号分类中的效果。

[0102] 其中，利用第三方机器学习工具包Scikit-Learn，对不同的分类算法进行性能测试。Scikit-Learn是操作简单、高效的机器学习和数据分析工具，其中包含的机器学习模型非常丰富，包括支持向量机SVM，决策树，随机森林，梯度回归分类算法、朴素贝叶斯，GBDT，邻近算法KNN等等，可以根据数据特征选择合适的模型进行机器学习训练得到分类模型。

[0103] 以上是本发明实施例提供的一种微博账号分类的方法的详细说明，可以理解，本发明通过对微博账号对应的微博数据进行特征提取得到微博数据特征，采用已建立的按微博数据特征划分的分类模型确定微博账号的类型。由于分类模型是通过对数量庞大的已知类型微博账号对应的微博数据进行机器学习训练而得到，因此，分类模型非常全面和具有代表性，通过分类模型对微博账号的类型进行确定，从而可以对微博账号的识别和分类能够更加高效和准确。

[0104] 本发明的方法是建立在分析用户的基本资料、微博内容、交互行为、发布行为4类特征上，这4类特征可有效的描述一个微博用户的特点，实现微博账号的识别与多分类，使得账号的识别具有更高的主动性和精确性。并且能够对账号进行细分到具体的类型。

[0105] 请进一步参阅图9，图9是本发明实施例提供的一种微博账号分类的装置的结构示意图，本实施例的微博账号分类的装置用于执行上述实施例的方法。如图所示，本实施例的微博账号分类的装置100包括获取模块11、特征提取模块12以及确定模块13，其中：

[0106] 获取模块11用于获取未知类型的微博账号对应的微博数据。

[0107] 本发明实施例中，获取模块11可以采用微博应用程序编程接口(Application Programming Interface，API)和网络爬虫两种方法获取微 博数据。但微博API接口对访问频率和属性获取有较大限制。因此作为本发明的优选实现方案，采用网络爬虫的方式获取微博数据。基于网络爬虫原理实现完成了微博爬虫工具，该爬虫工具能够获得微博页面上所有能呈现出的所有微博数据，并且将获得的原始微博数据进行预处理，最终存入数据库。

[0108] 在具体实现时，微博数据的获取除了完成基本属性值数据的获取，同时获取每个账号的最新500条微博，若微博数不足500条的，将其所有微博内容全部获取。爬取过程可以采取多台计算机分担微博数据爬取任务，避免爬取时间造成的属性差异。

[0109] 特征提取模块12对微博数据进行特征提取得到微博数据特征。

[0110] 根据当前微博特点，本发明实施例抽取并扩展出4类微博数据特征：用户资料特征、微博内容特征、交互行为特征和发布行为模式特征，综合考虑多种类型账号特征能够提高账号类型识别准确率。特征提取模块12对微博数据进行特征提取，根据微博数据特征生成微博数据特征值向量。

[0111] 确定模块13根据微博数据特征，采用已建立的按微博数据特征划分的分类模型确定微博账号的类型。

[0112] 其中，本发明实施例中的微博账号的类型为僵尸账号、广告账号、机器账号或普通账号中的一种。

[0113] 分类模型的目的是建立一个能够描述给定账号在账号类型中的出现频次或概率的分布。即利用分类模型，可以确定某一账号为哪个类型账号的可能性更大。通过分类模型，可以对未知类型的账号进行类型识别与区分。

[0114] 具体针对一个未知类型账号进行分类时，首先输入该账号的用户身份证明(User Identification,UID)，然后通过微博爬虫工具获取其相关数据，基于数值型特征集合生成基于数值型特征集合生成特征向量1和特征向量3，基于用户发布过得微博文本内容，生成特征向量2，根据特征向量，通过分类模型采用排除法确定账号类型。

[0115] 举例而言，1)使用特征向量1判断是否是僵尸账号，若是，则停 止判断，若不是，则继续下一步；2)基于用户发布过得微博文本内容，生成特征向量2；3)使用特征向量2判断是否是广告账号，若是，则停止判断，若不是，则继续下一步；4)使用特征向量3判断是否是机器账号，若是，则停止判断，若不是，则判定为普通账号。

[0116] 为了进一步确保分类的准确性，本发明实施例的确定模块在通过分类模型初步确定账号类型后，进一步通过分类算法对已确定类型的账号(即僵尸账号、广告账号、机器账号以及普通账号)进行二分类。

[0117] 其中，确定模块13在对以确定类型的账号进行二分类时，针对僵尸账号和机器账号的识别均采用抽取出的数值型特征集合构成特征向量，分别通过分类算法进行普通账号和僵尸账号、普通账号和机器账号的二分类。

[0118] 其中，作为本发明实施例的一种优选的实现方案，通过随机森林分类算法对僵尸账号和其余三种账号的集合进行二分类，通过朴素贝叶斯分类算法对广告账号、普通账号以及机器账号的集合进行二分类，以及通过梯度回归分类算法对机器账号和普通账号的集合进行二分类。

[0119] 作为一种优选，采用文本分类的通用方法进行普通账号和广告账号的二分类，以进一步确定账号为普通账号还是广告账号。

[0120] 文本分类要做如下4个预处理动作：

[0121] 1、选择微博广告和非广告文本数据集；

[0122] 2、微博文本预处理：分词、去停用词、建立词袋模型；

[0123] 3、选择文本分类使用的特征向量：词频表征特征权重；

[0124] 4、量化训练数据集和测试数据集文件。

[0125] 其中广告微博内容涉及各种电商卖家广告、代购广告、微商广告等，广告内容类型多样，但其中含有一些共同的明显的营销词汇，比如打折、优惠、包邮、购买、正品、限量等，这些具有区分性的词汇便是文本分类的关键。同时，将所有不具有广告意图的微博内容归

为普通用户发布的非广告微博。

[0126] 通过实验发现，对于广告账号和普通账号的进一步识别中用到的文本分类算法，朴素贝叶斯分类算法效果较好；对于僵尸账号和机器账号 的进一步识别中用到的分类算法，集成分类算法随机森林RandomForest、AdaBoost相比较与KNN、SVM、朴素贝叶斯等单模型算法更有效。

[0127] 当然，基于以上分类效果对比，在具体应用过程中，可以根据账号类型分别选取准确率(或F-score)最高的3个分类算法对账号进行二分类。

[0128] 其中，本发明实施例提供的微博账号分类的装置还可以进一步用于建立按微博数据特征划分的分类模型。本发明中的分类模型是基于已知类型账号的微博数据不断通过机器学习和训练而得到。

[0129] 在具体实现过程中，获取模块12用于获取已知类型的微博账号的微博数据。

[0130] 其中，已知类型的微博账号来源于人工标记或者电商购买的标记样本。人工标记，即手动查看每个微博账号的资料及微博动态来判定账号类型。电商购买，随着微博的盛行，电子商务网站上已经出现了很多微博服务商品，比如可以购买微博粉丝、微博账号，甚至一条微博的转发量和点赞数都可以买到，其中卖家出售的微博粉丝，其中就是低级粉丝即僵尸账号，高级粉丝即机器账号，通过直接购买粉丝的方式，可以减少大量人力。

[0131] 在本实施例具体实现时，僵尸账号共标记2000个，其中1500个来自两个淘宝卖家的低级微博粉丝，另外500个通过人工标记。标记的依据是：1)无头像或系统默认头像；2)关注数远大于粉丝数；3)微博数较少且无转发和评论；4)用户昵称为简单的字母和数字组合或汉字和数字组合；5)用户资料填写内容少或无。综合考虑以上5个方面来判断一个账号是否为僵尸账号。通过观察微博，发现很多娱乐明星和认证公司(推销商品)的粉丝中存在大量僵尸账号，有的娱乐明星希望通过百万甚至千万级的粉丝数来提高自己的知名度，认证公司希望购买僵尸账号提高粉丝数，从而吸引普通微博用户的关注，所以僵尸账号的收集目标就集中在娱乐明星和认证公司的粉丝列表中。

[0132] 广告账号共标记1000个，全部来自人工标记。标记的依据是：1)微博内容以广告、促销和抽奖等为主；2)用户简介中有店铺链接、微 信号或商品介绍；3)微博中的链接多为商品买卖链接。

[0133] 机器账号共标记2000个，其中1500个来自淘宝购买的高级微博粉丝，400个通过人工标记，100个来自相关研究中使用到的机器账号样本。人工标记的依据是：1)微博发布时间规律性强，每隔一定时间发布一条微博；2)微博内容主题是心灵鸡汤、名人名言、笑话、天气、星座运势等，微博内容也可能以广告为主，有很大嫌疑是通过调用现成的语料库来自动发布这些内容微博；3)微博内容重复度高，不同的机器账号可能使用同一些语料库；4)微博发布平台种类少，部分机器账号的微博发布平台能明显的说明使用了第三方软件，如皮皮时光机、云中小鸟、孔明社交管理等。

[0134] 普通账号共标记3000个，全部来自人工标记。标记依据是：1)粉丝数和微博数较多；2)用户头像是真实照片；3)用户资料填写详细；4)微博内容有日常生活气息，如有个人生活内容分享；5)微博有被转发或评论，同时又回复。收集方法：一是从自己的现实好友出发，然后再判断现实好友的粉丝和关注，接着递归判断粉丝的粉丝和关注、关注的粉丝和关注。而是从热门微博和热门话题下面寻找积极评论和互动的账号。

[0135] 微博数据的获取可以采用微博应用程序编程接口(Application Programming Interface,API)和网络爬虫两种方法。但微博API接口对访问频率和属性获取有较大限制。因此作为本发明的优选实现方案,采用网络爬虫的方式获取微博数据。基于网络爬虫原理实现完成了微博爬虫工具,该爬虫工具能够获得微博页面上所有能呈现出的所有微博数据,并且将获得的原始微博数据进行预处理,最终存入数据库。

[0136] 在具体实现时,微博数据的获取除了完成基本属性值数据的获取,同时获取每个账号的最新500条微博,若微博数不足500条的,将其所有微博内容全部获取。爬取过程可以采取多台计算机分担微博数据爬取任务,避免爬取时间造成的属性差异。

[0137] 特征提取模块12用于对已知类型的微博账号的微博数据进行特征提取,得到已知类型的微博账号的微博数据特征。

[0138] 根据当前微博特点,本发明实施例抽取并扩展出4类微博数据特征:用户资料特征、微博内容特征、交互行为特征和发布行为模式特征,综合考虑多种类型账号特征能够提高账号类型识别准确率。其中,不同微博数据特征的特征集合请参阅上述表1(加※为本发明新提出的特征),在此不再赘述。

[0139] 用户资料特征(表1中1-10号特征)来自用户比较直观的资料信息。其中微博年龄是从账号注册时间到2015年1月1日截止账号存在天数;

[0140] $粉丝度=\frac{粉丝数}{粉丝数+关注数}$, $关注度=\frac{关注数}{粉丝数+关注数}$; $粉丝关注比=\frac{粉丝数}{关注数}$。

[0141] 微博内容特征(表1中11-14号特征)根据微博内容中包含的特殊内容抽取得来。机器账号在大量发布微博同时还会较多地@好友,希望好友能够转发该微博或进行评论等,增加机器账号的人为特性。相反,僵尸账号几乎不@好友。所以@数可以作为区分机器账号和正常账号、僵尸账号和正常账号的特征。

[0142] 交互行为特征(表1中15-23号特征)表示微博账号和其他账号互动情况。机器账号由于使用了语料库,几乎不转发微博,大部分为原创微博;80%的机器账号评论数小于150,而大约60%的正常账号评论数都超过500,即正常账号更具有评论交互意向,机器账号要实现自动评论或回复复杂度较大。

[0143] 发布行为特征(表1中24-34号特征)代表微博账号发布行为模式。通过对微博账号的观察发现,大部分机器账号以一定的时间间隔自动发布微博,有的甚至24小时连续定时发布微博,有的会稍有伪装,避开0-6点休息时间发微博。机器账号微博发布时间更有规律,普通账号则显得无规律可循。使用熵率来度量微博用户发布微博时间规律性。

[0144] 随机变量序列$X=\{Xi\}$由一个微博用户所发微博的时间间隔随机变量组成,X_i表示第i条和第i+1条微博之间的时间间隔随机变量序列X的熵记为

[0145] $$H(X_1,\cdots\cdots,X_m)=-\sum_{i=1}^{m}P(x_i)\log P(x_i) \quad (1)$$

[0146] 其中$P(x_i)$是$P(X_i=x_i)$的概率。当已知该序列的前m-1项时,其条件信息熵记为:

[0147] $$CE(X_m|X_{m-1})=H(X_m|X_1,\cdots,X_{m-1})=H(X_1,\cdots,X_m)-H(X_1,\cdots,X_{m-1}) \quad (2)$$

[0148] 用户发微博的时间间隔构成的序列都是有限序列,而信息熵衡量的是一个无穷随机过程,无法直接用来计算有限的序列。引入修正的条件信息熵来解决序列有限性所带来的问题。修正的条件信息熵的公式如下:

[0149] $CCE(X_m|X_1,\ldots,X_{m-1})=CE(X_m|X_1,\ldots,X_{m-1})+perc(X_m)\cdot EN(X_1)$ (3)

[0150] 其中$perc(X_m)$是在长度为m的序列里面只出现过一次的序列所占的比例，$EN(X_1)$是当m＝1时的信息熵。当序列长度取[2,m]中的不同值时，分别计算出相应的修正条件信息熵的值，最终熵率取其中最小值。如果该账号是机器账号，那它的行为会有一定的规律性，因而其修正条件信息熵的值会较小。与之相反，普通账号的行为随机化程度较高，修正的条件信息熵值也会较大。

[0151] 针对行为模式特征中的信息熵，将机器账号和普通账号的发微博时间间隔序列输入后，利用式(3)得到每位账号用户的修正条件。

[0152] 针对广告账号的识别，只需要检测账号发布微博内容是否为广告内容即可，所以使用发布微博文本内容这一特征，实际在文本分类中又将这一特征分解为文本特征向量；针对僵尸粉的识别，根据特征分析，选择使用是否有头像、是否填写简介、昵称是否包含数字、粉丝数、关注数、微博数这6个数值型特征即可；针对机器账号的识别根据特征分析，选择使用是否填写简介、昵称是否包含数字、粉丝数、关注数、微博数、微龄、粉丝度、关注度、粉丝关注比、微博含图片数、原创数、$\text{原创率}=\frac{\text{原创数}}{\text{原创数+转发数}}$、转发数、$\text{转发率}=\frac{\text{转发数}}{\text{原创数+转发数}}$、评论数、被评论数、回复数、自转数、微博发布时间间隔熵、日均发布微博数、0-6点的平均微博数、6-12点的平均微博数、12-18点的平均微博数、18-24点的平均微博数、发布平台数、发布IP数、发布ISP数、发布省份数、发布城市数共29个数值型特征。特征数据分析不仅通过条形图、CDF图来展示，还通过具体分类模型计算了特征的重要。

[0153] 另外，特征提取模块12还用于对普通账号和机器账号分类的29个特征的重要性进行排名，通过排名，可以进一步进行特征选择，在保证 分类准确度的基础上加速账号分类速度。

[0154] 确定模块13对已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型。

[0155] 在具体实现时，可以采用10折交叉验证的方法，使用已标记样本数据集，训练分类模型，通过实验实际测试各个分类算法在微博账号分类中的效果。

[0156] 其中，利用第三方机器学习工具包Scikit-Learn，对不同的分类算法进行性能测试。Scikit-Learn是操作简单、高效的机器学习和数据分析工具，其中包含的机器学习模型非常丰富，包括支持向量机SVM，决策树，随机森林，梯度回归分类算法、朴素贝叶斯，GBDT，邻近算法KNN等等，可以根据数据特征选择合适的模型进行机器学习训练得到分类模型。

[0157] 以上是本发明实施例提供的一种微博账号分类的方法及装置的详细说明，可以理解，本发明通过对微博账号对应的微博数据进行特征提取得到微博数据特征，采用已建立的按微博数据特征划分的分类模型确定微博账号的类型。由于分类模型是通过对数量庞大的已知类型微博账号对应的微博数据进行机器学习训练而得到，因此，分类模型非常全面和具有代表性，通过分类模型对微博账号的类型进行确定，从而可以对微博账号的识别和分类能够更加高效和准确。

[0158] 本发明的方法是建立在分析用户的基本资料、微博内容、交互行为、发布行为4类特征上，这4类特征可有效的描述一个微博用户的特点，实现微博账号的识别与多分类，使得账号的识别具有更高的主动性和精确性。并且能够对账号进行细分到具体的类型。

[0159] 在本发明所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其 它的形式。

[0160] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0161] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0162] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)或处理器(processor)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0163] 以上所述仅为本发明的实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

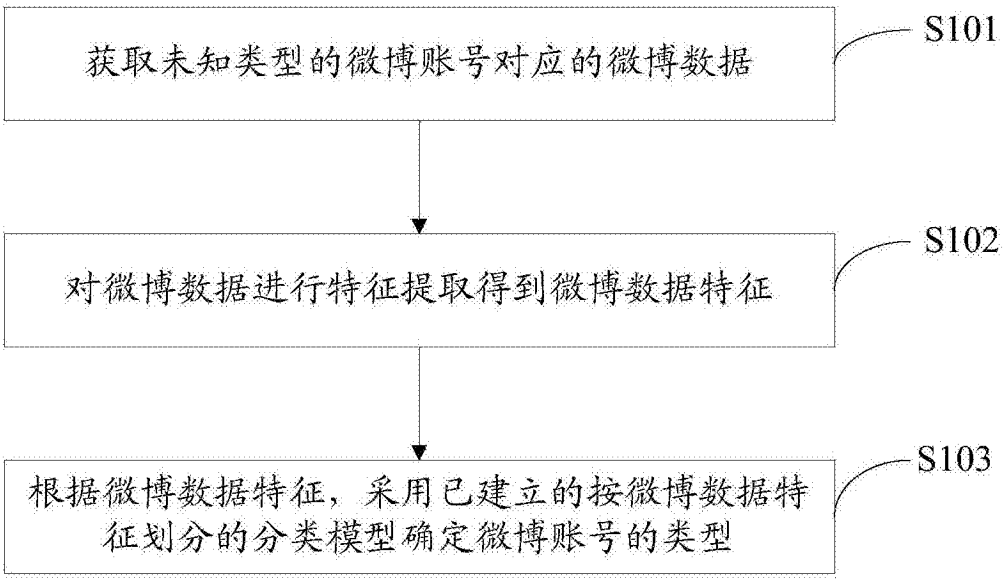

获取未知类型的微博账号对应的微博数据
S101
对微博数据进行特征提取得到微博数据特征
S102
根据微博数据特征，采用已建立的按微博数据特征划分的分类模型确定微博账号的类型
S103

图1

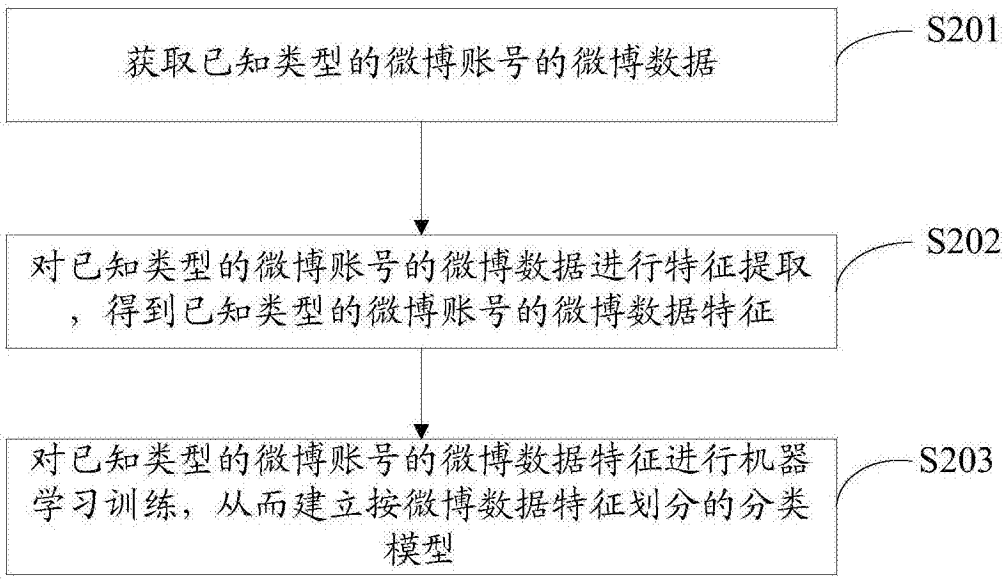

获取已知类型的微博账号的微博数据
S201
对已知类型的微博账号的微博数据进行特征提取，得到已知类型的微博账号的微博数据特征
S202
对已知类型的微博账号的微博数据特征进行机器学习训练，从而建立按微博数据特征划分的分类模型
S203

图2

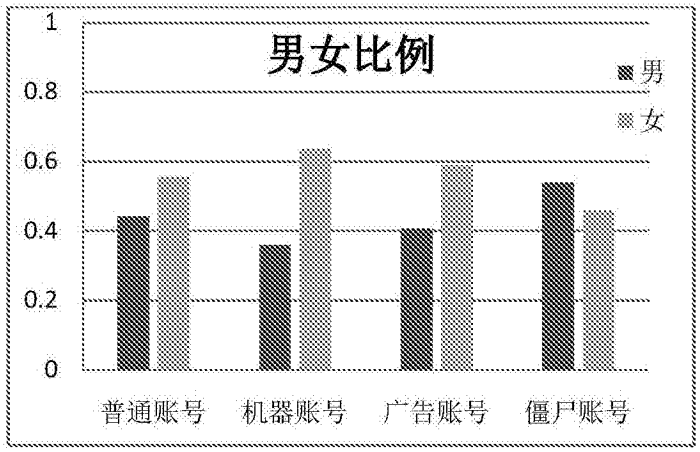

男女比例
1
0.8
0.6
0.4
0.2
0
男
女
普通账号
机器账号
广告账号
僵尸账号

图3(a)

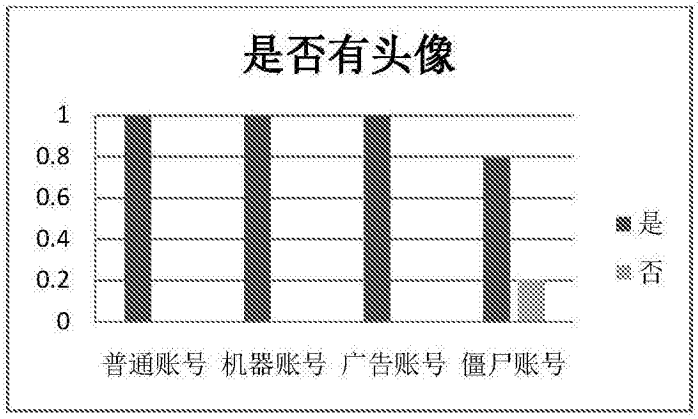

是否有头像
1
0.8
0.6
0.4
0.2
0
是
否
普通账号
机器账号
广告账号
僵尸账号

图3(b)

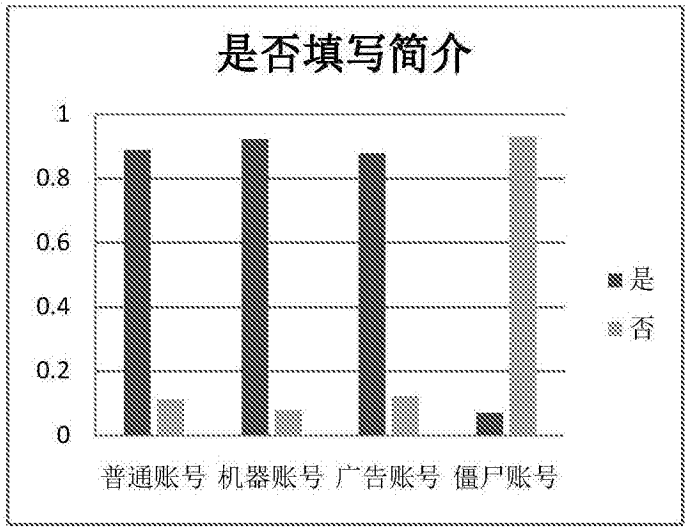

是否填写简介
1
0.8
0.6
0.4
0.2
0
是
否
普通账号
机器账号
广告账号
僵尸账号

图3(c)

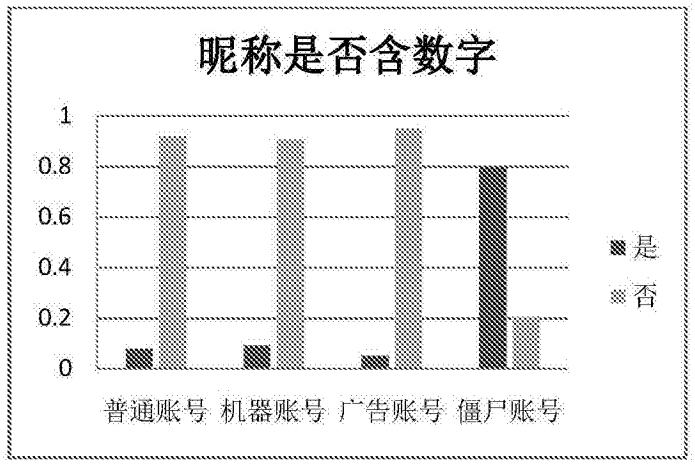

昵称是否含数字
1
0.8
0.6
0.4
0.2
0
是
否
普通账号 机器账号 广告账号 僵尸账号

图3(d)

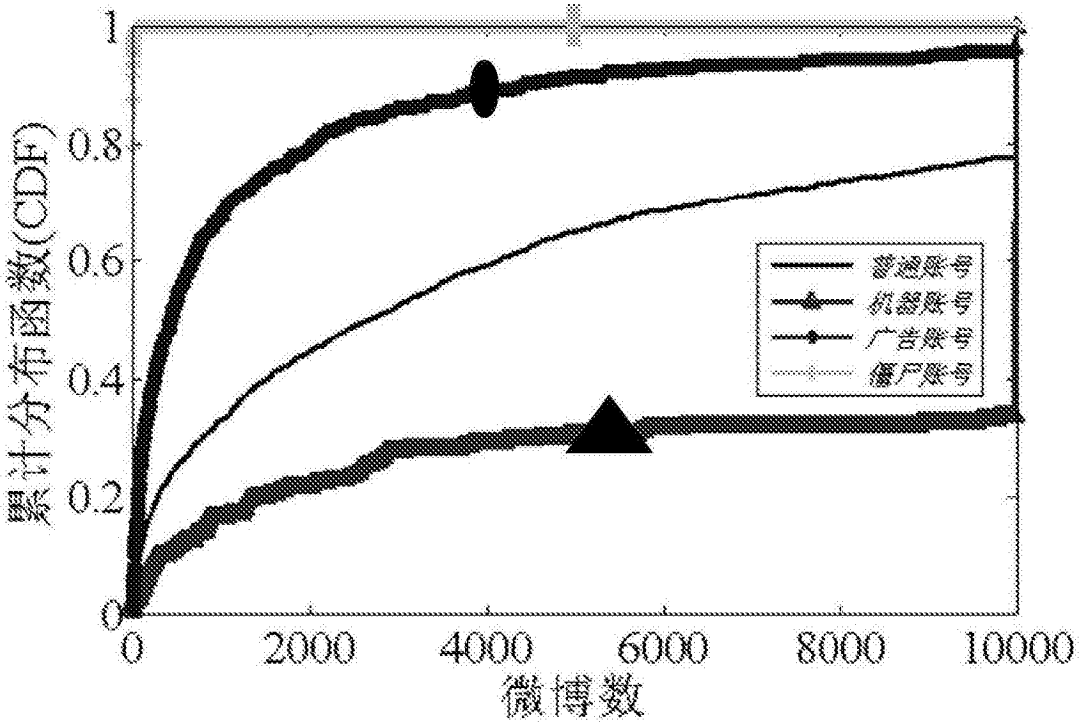

累计分布函数(CDF)
1
0.8
0.6
0.4
0.2
0
0
2000
4000
6000
8000
10000
微博数
普通账号
机器账号
广告账号
僵尸账号

图3(e)

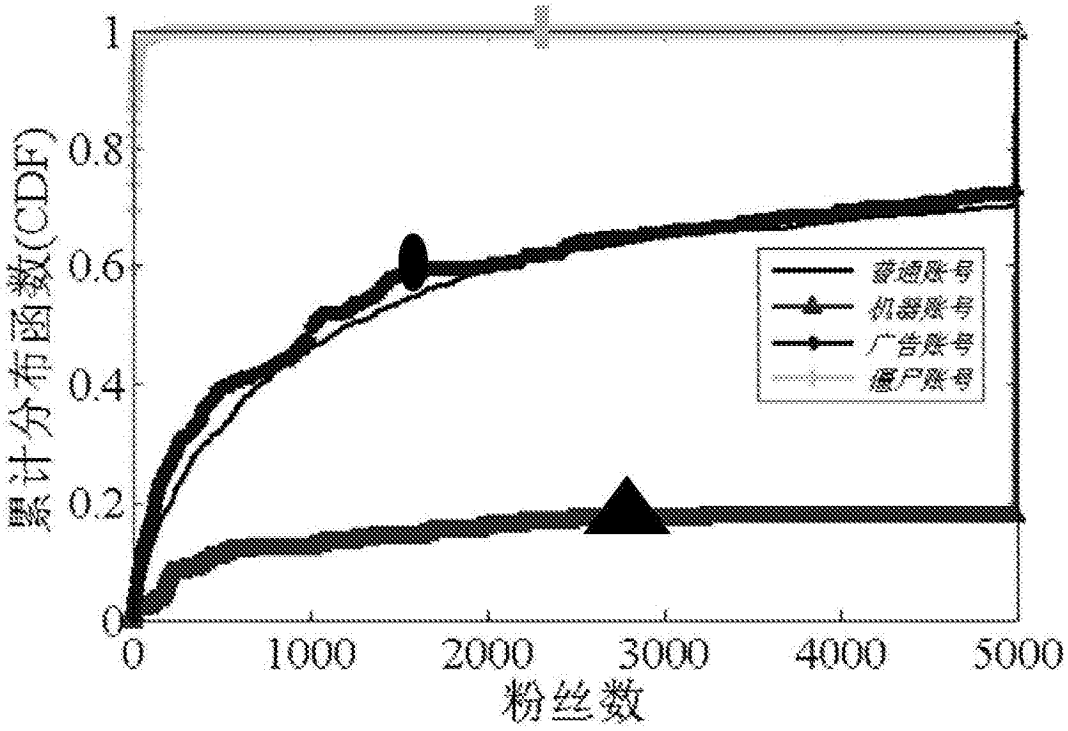

累计分布函数(CDF)
粉丝数
0
0.2
0.4
0.6
0.8
1
0
1000
2000
3000
4000
5000
普通账号
机器账号
广告账号
僵尸账号

图3(f)

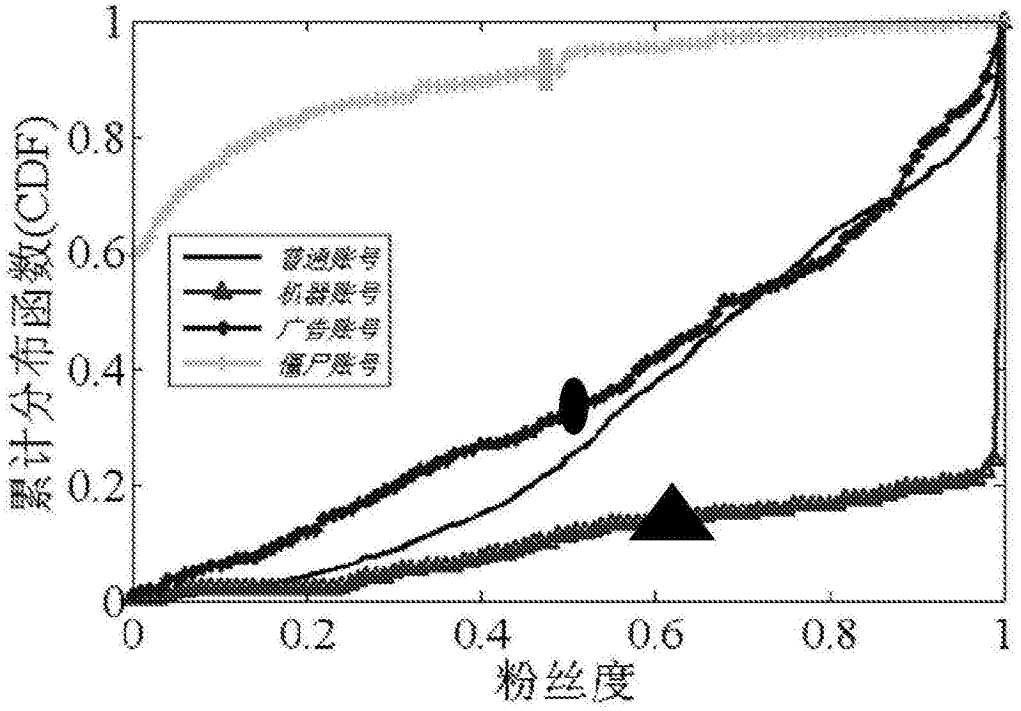

累计分布函数(CDF)
粉丝度
0
0.2
0.4
0.6
0.8
1
普通账号
机器账号
广告账号
僵尸账号

图3(g)

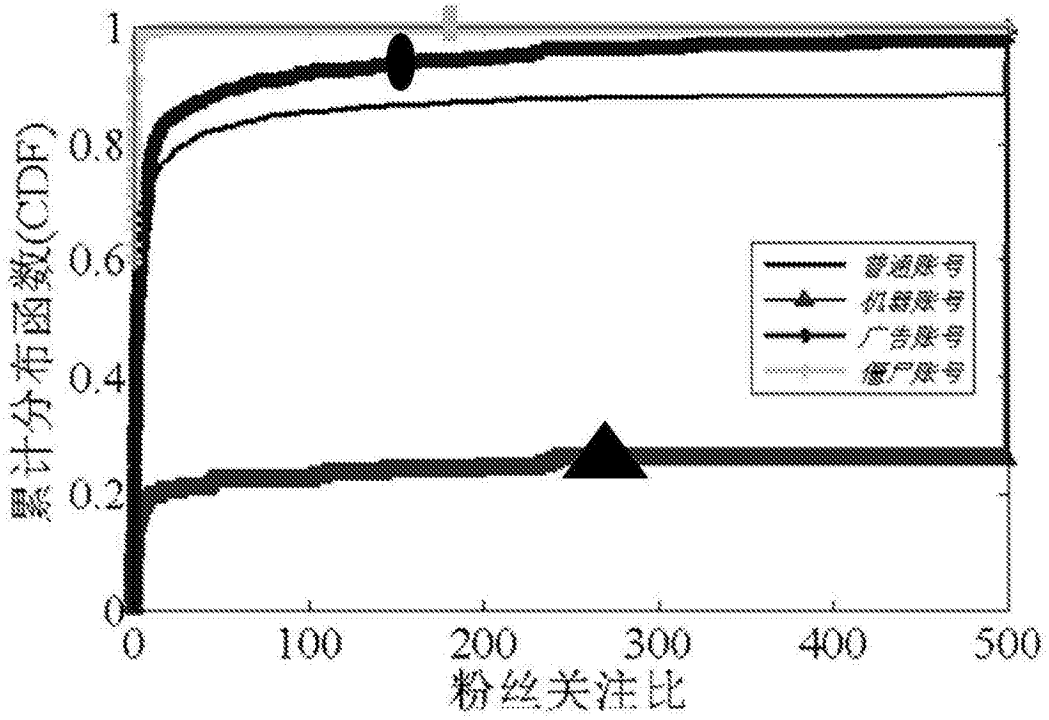

累计分布函数(CDF)
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
粉丝关注比
普通账号
机器账号
广告账号
僵尸账号

图3(h)

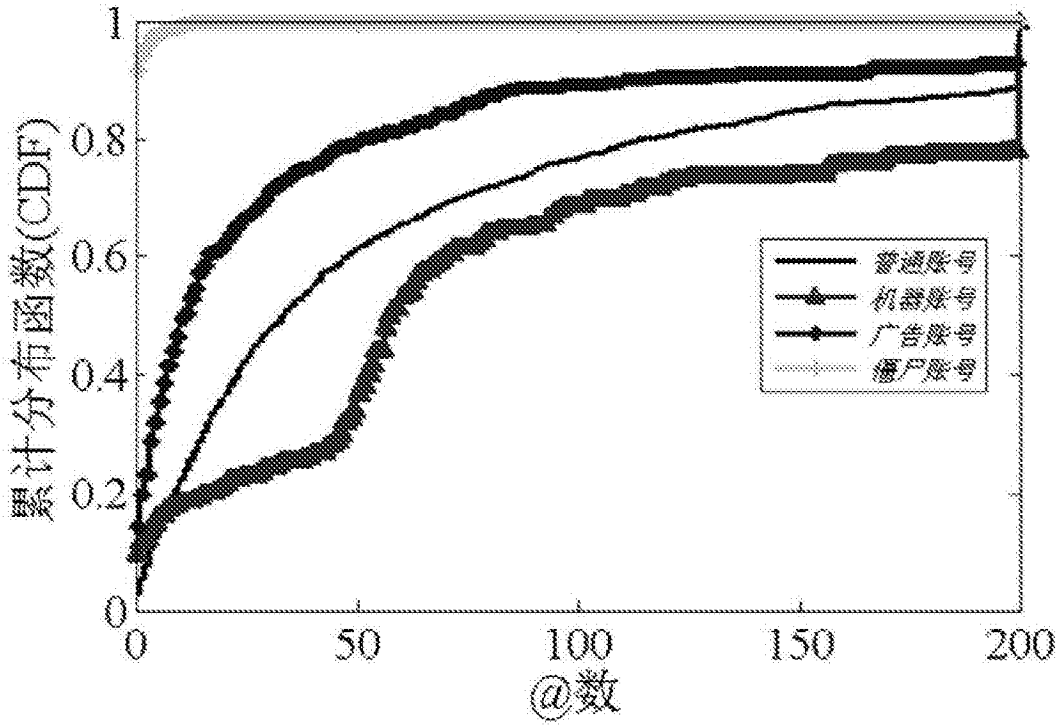

累计分布函数(CDF)
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
@数
普通账号
机器账号
广告账号
僵尸账号

图4

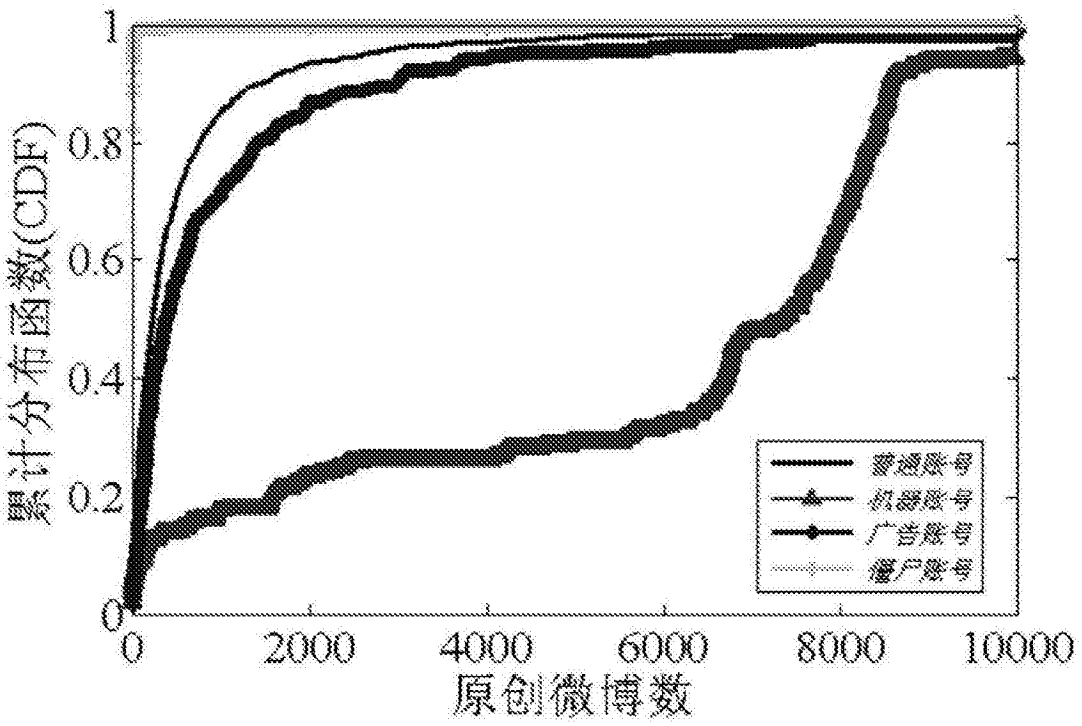

累计分布函数(CDF)
原创微博数
0
0.2
0.4
0.6
0.8
1
0
2000
4000
6000
8000
10000

图5(a)

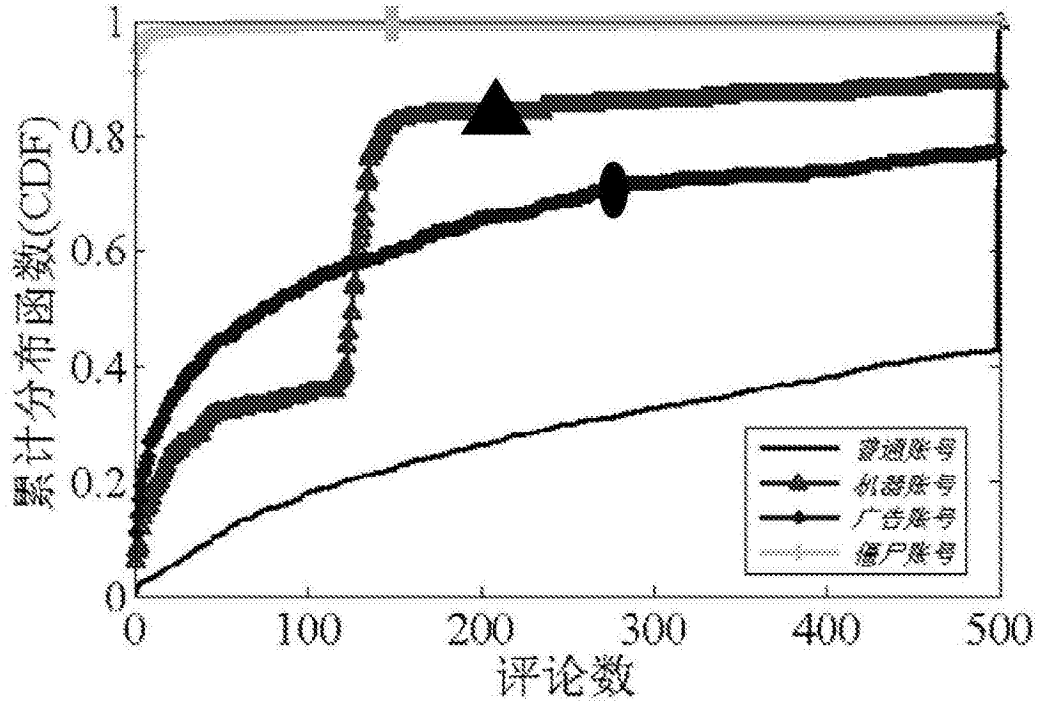

累计分布函数(CDF)
评论数
0
0.2
0.4
0.6
0.8
1
0
100
200
300
400
500

图5(b)

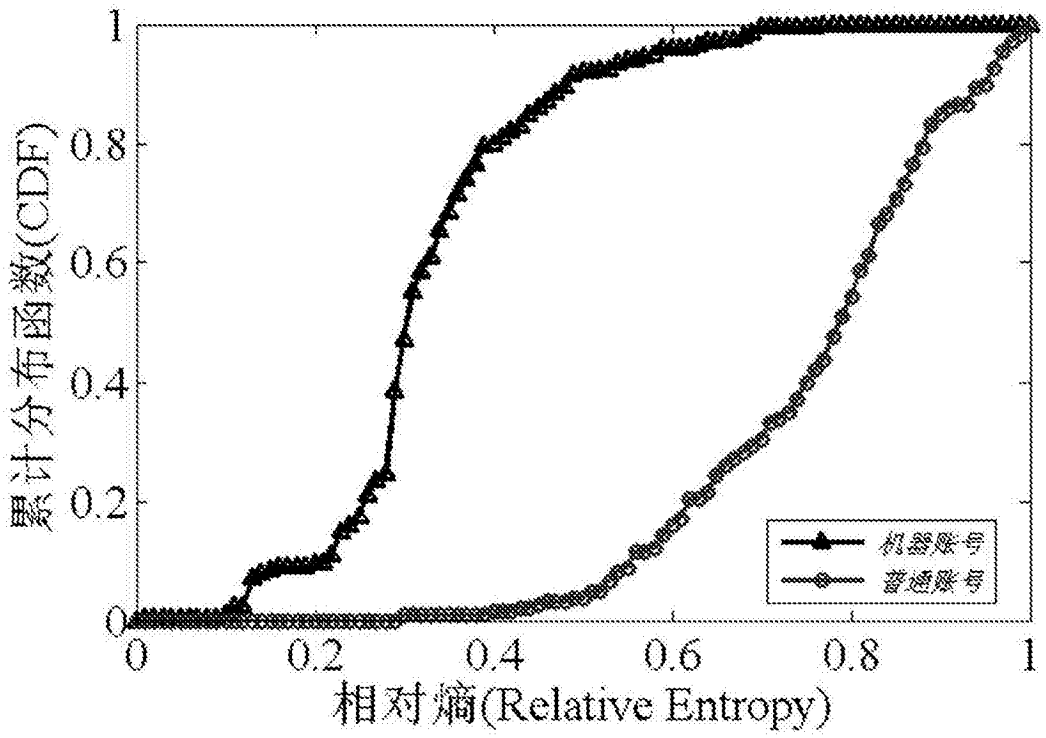

累计分布函数(CDF)
相对熵(Relative Entropy)
0
0.2
0.4
0.6
0.8
1

图6

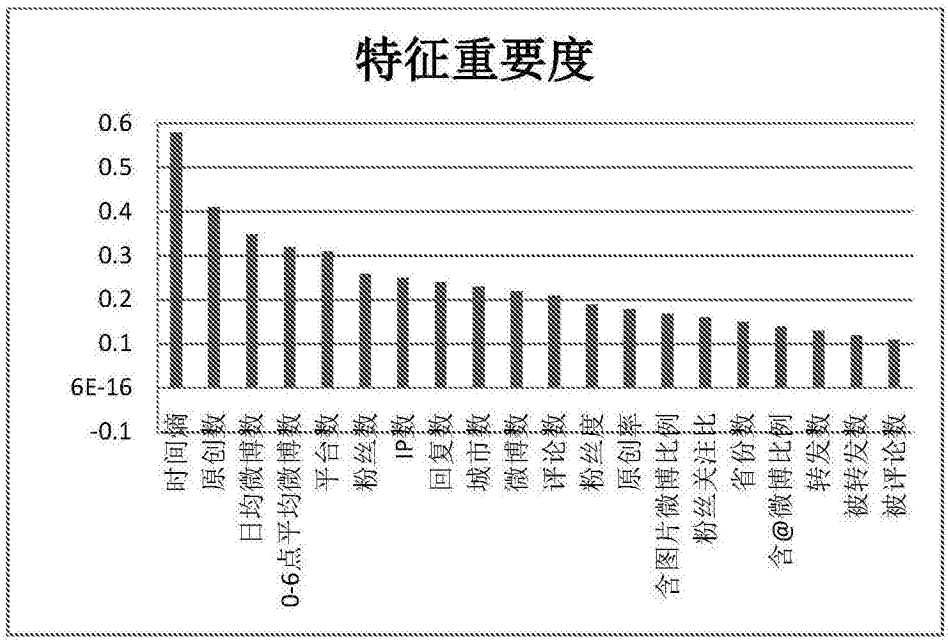

特征重要度
0.6
0.5
0.4
0.3
0.2
0.1
6E-16
-0.1
时间熵
原创数
日均微博数
0-6点平均微博数
平台数
粉丝数
IP数
回复数
城市数
微博数
评论数
粉丝度
原创率
含图片微博比例
粉丝关注比
省份数
含@微博比例
转发数
被转发数
被评论数

图7

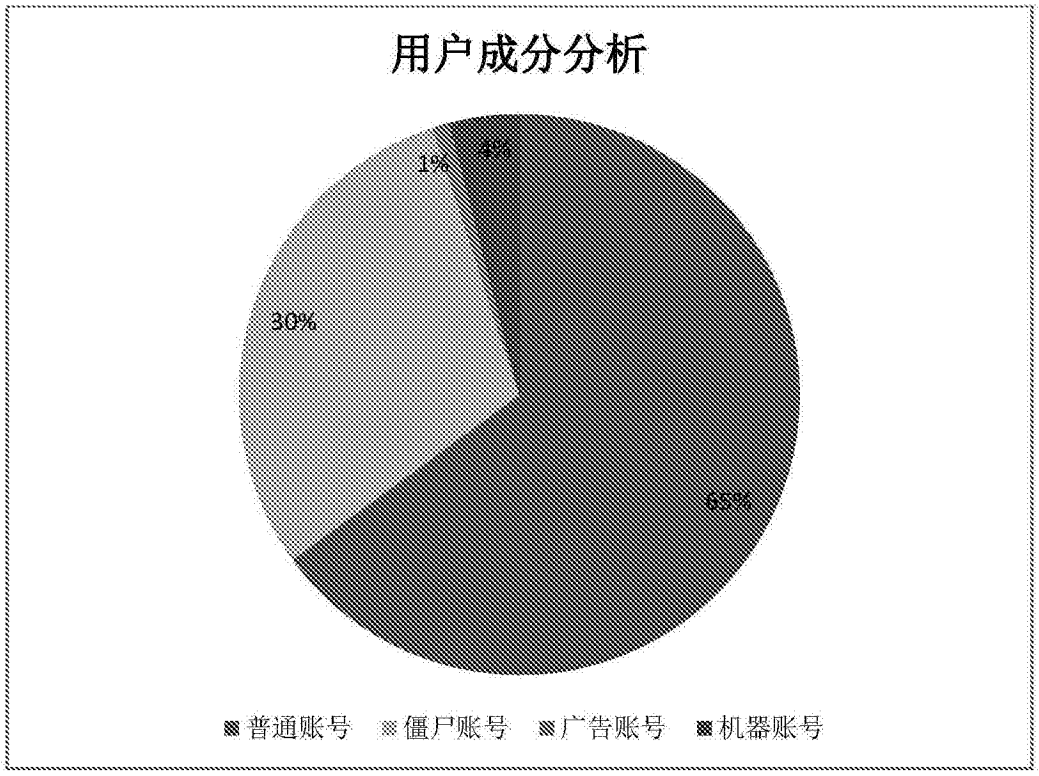

用户成分分析
1%
4%
30%
65%
普通账号
僵尸账号
广告账号
机器账号

图8

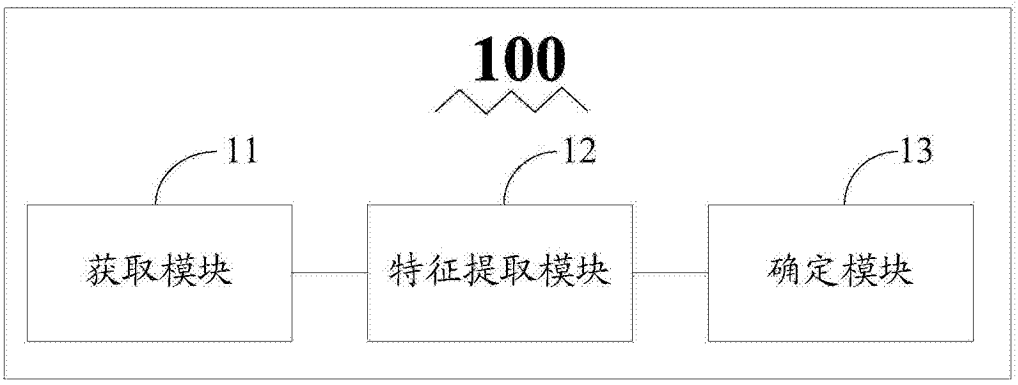

100
11
12
13
获取模块
特征提取模块
确定模块

图9