基于新浪微博的社交网络垃圾用户分析与检测

孟祥飞，徐 路，王思雨
（中国民航大学计算机学院，天津 300300）

摘 要：随着信息技术和互联网的飞速发展，社交网络在人们生活中扮演着不可替代的角色。但同时，社交网络中也充斥着各种各样的广告信息，严重影响了用户的体验。一些营销团队恶意注册的大量垃圾账号也让正常用户不胜其烦。针对这些问题，首先阐述了社交网络垃圾用户产生的原因，进而分析了垃圾用户的特征，最后基于新浪微博的数据，使用 C4.5 决策树分类算法对用户进行分类。实验结果显示，该方法检测用户的准确率为 92%.

关键词：社交网络；新浪微博；垃圾用户；分类

中图分类号：TP393 文献标识码：A 文章编号：2095－6835（2014）15－0125－03

社交网络是在线社交网络（Online Social Network，“ OSN ”）的简称。社交网络服务是基于六度分隔理论，以互动交友，用户之间共同的兴趣、爱好、活动或者用户间真实的人际关系为基础，以实名或者非实名的方式在网络平台上构建的一种社会关系网络服务。Facebook 被认为是第一个真正意义上的社交网站。当今热门的 Twitter、新浪微博、腾讯微博、人人网等都属于社交网络。截至 2012-08，世界上最大的社交网站 Facebook 拥有注册用户约 10 亿人，其网络流量曾一度超过网络巨头 Google；新浪微博的最新注册用户已达到了 3 亿；人人网用户量在 2 亿左右。其中，新浪微博是最活跃、最有影响力的微博平台之一。微博的单向关注和即时推送机制使得信息在该平台上传播极为迅速，形成了“ 围观改变中国 ”的架势。

1 微博垃圾用户产生的背景

随着社交网络的快速发展，其传媒价值受到了社会各界的关注。在微博中，拥有众多粉丝的明星用户在社会舆论中有着非常重要的作用。如今，微博作为举足轻重的宣传平台，受到了广告商的青睐，他们通过发起话题、借助明星微博等方法来宣传产品。很多营销团队为了推销，注册了大量账号，专门发布广告，宣传网店、产品等信息。这些广告信息在没有监管的情况下，充斥着整个社交网络，不仅真实性无法保证，而且对用户体验产生了极大影响。另外，在新浪微博中，拥有极高粉丝数量的意见领袖的出现也给了投机者们一种营销的渠道。他们注册了大量账号，并在网上出售粉丝。当有用户向其购买时，他们就用大量的账号去关注该用户，提高该用户的关注度和影响力得分，借此吸引普通用户的注意。一些炒作团队也会使用批量注册的账号去对某一话题进行炒作，使其变成热门话题，借此达到影响舆论的目的。这些批量注册的账号不仅给服务器增加了许多负担，而且扰乱了微博的生态秩序。由其制造的层出不穷的谣言也降低了微博作为信息来源的可靠性，影响普通用户的生活。笔者通过抽取用户的关注粉丝比、链接比、互粉数、平均评论数等特征，提出一种用户行为特征的垃圾用户分类检测方法，实现了对“ 用户是否为垃圾用户 ”的检测。

2 相关研究

2.1 关于垃圾用户检测的相关研究

在新浪微博兴起之前，Twitter 与 Facebook 已经拥有众多的用户。由于 Twitter 与新浪微博的结构非常相似，对新浪微博垃圾用户的检测工作可以参考 Twitter 垃圾用户的检测工作。Kurt Thomas 等学者在研究中指明，现在的垃圾用户不参与正常的社交生活，但是他们通过主动关注别人和在热门话题下发表垃圾评论来吸引正常用户点击。以往的许多研究工作是基于已有用户的数据来进行的，Zhi Yang 等人用了一种基于蜜罐的方法来检测垃圾用户，通过在社交社区中放置蜜罐，吸引垃圾用户关注，然后通过链接搜集垃圾用户的图谱（Profile），搜集文本内容、社交网络和发布模式方面的特征。在对社交网络垃圾用户的研究中，垃圾用户的定义并不是学者进行研究工作的重点，Gianluca Stringhini 等学者在其研究中将垃圾用户分为四类，针对其中的两类提取了相关特征，并用随机森林法进行分类。Alex Hai Wang 在其关于 Twitter 的研究中对各种分类算法进行了比较。他使用了决策树、神经网络、支持向量机、K-近邻、和贝叶斯分类器提取了互粉数、粉丝比和追随比，然后又根据基于内容的分析和回复数来进行分类。通过实验，得到了贝叶斯分类最精确的结果。除了新浪微博之外，中国的人人网社交平台也拥有众多的用户。Yin Zhu 在其关于人人网的研究之中，创新性地提出了利用矩阵分解的方法来进行垃圾用户的检测，定义了精确度和召回率，使用了 SVM、SF+SVM、MF+SVM、MFSR+SVM 进行用户分类工作，并对结果进行比较。

2.2 决策树算法的产生与改进

决策树算法最早是 20 世纪 50 年代由亨特在“ CLS ”（Concept Learning System）中提出，后经发展由 J.R.Quinlan 在 1979 年提出了著名的 ID3 算法。ID3 算法是建立在奥卡姆剃刀的基础上，以信息熵和信息增益为衡量标准，从而实现对数据的归纳分类，其主要是针对离散型属性数据。C4.5 决策树算法继承了 ID3 算法的优点，并对 ID3 算法进行了改进。C4.5 决策树算法在树构造过程中进行剪枝，并且用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足。C4.5 决策树算法不仅能对离散型数据、连续属性的离散化进行处理，还能够对不完整数据进行处理。

参考以上学者的研究工作，我们决定提取用户的关注粉丝比、链接比、互粉数、平均评论数等特征，使用 C4.5 决策树算法来对用户进行分类。

Analysis and Optimization of Wireless Network System Covering Issues

Han Yan

Abstract: The weak presence of wireless network coverage, had covered no main cell coverage and other issues analyzed, analysis, deal with the problem solving process and suggestions for the network optimization analysis covering the relevant indicators, covering issues processing and so provides the basis of theoretical analysis, in order to cover the issue of optimizing the different communication scenarios.

Key words: wireless networks; weak coverage; over coverage; optimization

3 C4.5 决策树构建

C4.5 决策树方法是以实例为基础的归纳学习方法，它从一个无次序、无规则的实力集合中归纳出一组采用树形结构表示的分类规则。在 C4.5 决策树使用过程中，首先根据已有的数据构建分类模型，接着用构建好的模型对样本进行分类。

在 C4.5 决策树中，根节点代表整个训练样本集，某一结点对应的子树对应着原数据集中满足某一属性测试的部分数据集。在构建过程中，从根节点开始，通过在每个节点上对某个属性的测试验证、算法递归得将数据集分为更小的数据集。将这个递归过程一直进行下去，直到某一节点对应的子树对应的数据集都属于同一类为止。图 1 所示是一个简单的决策树分类模型，其主要功能是实现“根据天气情况来决定是否去户外活动”。

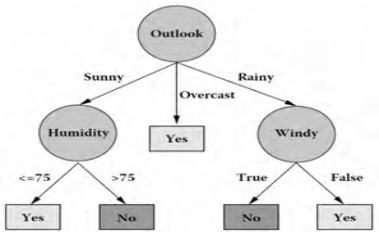

图 1 简单的决策树模型

4 特征的提取

在微博中，有许多类型的用户。除了正常的用户外，还有一些专门宣传该公司产品的用户，比如明星用户、僵尸用户等。所谓“僵尸用户”，就是并不参与正常的社交生活，而是发布垃圾链接或是专门热捧某一用户或话题的用户。根据对微博垃圾用户的理解，我们此次检查的重点就是僵尸用户。根据僵尸用户的特征，我们提取了以下特征来检测垃圾用户。

4.1 微博中用户关注的人与其粉丝的比

在微博生活中，由于垃圾用户并不实际参与社交生活，知道其存在的用户不多，这样就导致其粉丝数量并不多。同时，一部分垃圾用户作为被销售的粉丝而存在，其关注的用户数量要比一般人多。因此，我们定义了以下变量：

R=following/followers. (1)

式（1）中：following——微博用户中其关注的用户数量；

followers——微博用户拥有的粉丝数。

4.2 微博用户中所有含 url 链接的比例

由于垃圾用户经常发布一些含有链接的微博吸引正常用户点击，正常用户点击进去之后，页面上就会显示一些垃圾信息，因此，我们定义了以下变量：

U=messages_containing_urls/total_messages. (2)

式（2）中：messages_containing_urls——含有链接的信息总数；

total_messages——用户发布的信息总数。

4.3 互粉数

由于垃圾用户并不参与实际社交生活，其关注的用户大部分都不是认识的好友。在正常的社交生活中，用户一般会关注自己认识的好友，用户在收到别人的关注以后，会查看自己的粉丝是否为认识的人。如果认识，就关注该好友。而垃圾用户在关注别人以后，被关注的人并不认识该垃圾用户，因此不会关注该垃圾用户。这样，垃圾用户的互粉数就会低于正常用户，因此，我们定义了以下变量：

F=friend_numbers. (3)

式（3）中：friend_numbers——用户的好友数目。

用户的好友指的是某一用户与该用户之间是互粉关系，两个用户彼此互相关注。

4.4 平均评论数

由于垃圾用户经常发布一些垃圾信息，很少有用户会评论其发布的信息，因此，其平均评论数会低于一般用户。因此，我们定义了以下变量：

Ac=total comments/total_messages. (4)

式（4）中：total_comments——某一用户所发的微博的所有评论数；

total_messages——某一用户所发微博的总数。

5 决策树算法的实现

在实现过程中，对于样本集 $S=\{X_1, X_2, \cdots, X_M\}$来说，每一个样本均可以由属性向量$\{A_1, A_2, \cdots, A_N\}$来表示，那么，根据属性 A_i（$0<i<N+1$）可以将样本集 S 分为 $C_1, C_2, \cdots, C_L$ 共 L 个子集。用于计算信息增益率的主要公式如下。

某一类别的信息熵为：

$$H(C) = -\sum_j P(C_j)\log_2(P(C_j)). \quad (5)$$

式（5）中：C_j——C 类别的一个取值，$0<j<L+1$.

按属性 A 将数据集合 C 分割后，其类别条件的信息熵为：

$$H(C|A) = -\sum_j\sum_i P(a_i)P(C_j \mid a_i)\log_2(P(C_j \mid a_i)). \quad (6)$$

式（6）中：a_i——属性 A 的一个取值，$0<i<N+1$.

信息增益为：

$$I(C, A) = H(C) - H(C \mid A). \quad (7)$$

属性 A 的信息熵为：

$$H(A) = -\sum_j P(A_j)\log_2(P(a_j)). \quad (8)$$

信息增益率为：

$$\text{gain\_ratio} = I(C, A) \mid H(A). \quad (9)$$

根据以上公式，很容易计算出在实际应用中各个属性的信息增益率。在决策树构建过程中，算法采用离散化取值空间的策略，将其转化为离散属性进行计算。

在构建过程中，主要步骤如下所示：①对于当前的每一个属性 A_i，分别计算其信息增益率 gain_ration。②选择具有最大信息增益率的属性作为其根节点的属性值。③根据根节点的属性值，对当前集合进行分类。④如果分得的某个分类节点中只有所需分类结果中的某一类，则该节点为叶子节点；否则，以当前节点为根节点，继续进行分类。⑤对所得的决策树进行剪枝，得到最终的分类决策树。

在决策树的构建过程中，通过选择信息增益率最大的属性作为测试属性，C4.5 决策树算法自上而下完成决策树的构建过程。在构建过程中，C4.5 决策树算法使用剪枝算法对初始构建的树进行剪枝，得到最后的决策树。在本实验中，最终构建的分类树如图 2 所示。

6 实验与结果分析

为了检验提取特征的有效性，我们标注了一些数据集进行训练和测试。首先随机提取了新浪微博中 4 100 用户作为垃圾用户检测的数据集。在检测方法上，我们使用了基于 C4.5 决策树分类算法、Adaboost 分类算法进行模型训练，并用 10 倍交叉验证的方法检验模型。以下表 1、表 2 为两个分类算法的检测结果。表中，T 代表分类为普通用户，F 代表分类为垃圾用户。图 3 所示为 C4.5 决策树与 Adaboost 算法的结果比较。

在实验中，我们共选取了 4 100 用户的数据，其中，正常用户为 3 508 个，垃圾用户为 592 个。从分类结果来看，C4.5 决策树算法对本文中提取的特征类结果比 Adaboost 算法分类效果好。其中，检测结果正确的用户达到 3 803 个，分类错误结果为 297 个，正确率达到 92%，同时，ROC Area 达到 86%，召回率达到 79%.

从实验结果来看，C4.5 决策树算法的分类结果比较令人满意。接着我们查看了分类错误的用户，发现这些用户作弊手段十分高明——他们会发一些正常的微博，但是在其中会夹杂一些垃圾信息。该检测方法对这些用户的分类结果较差。

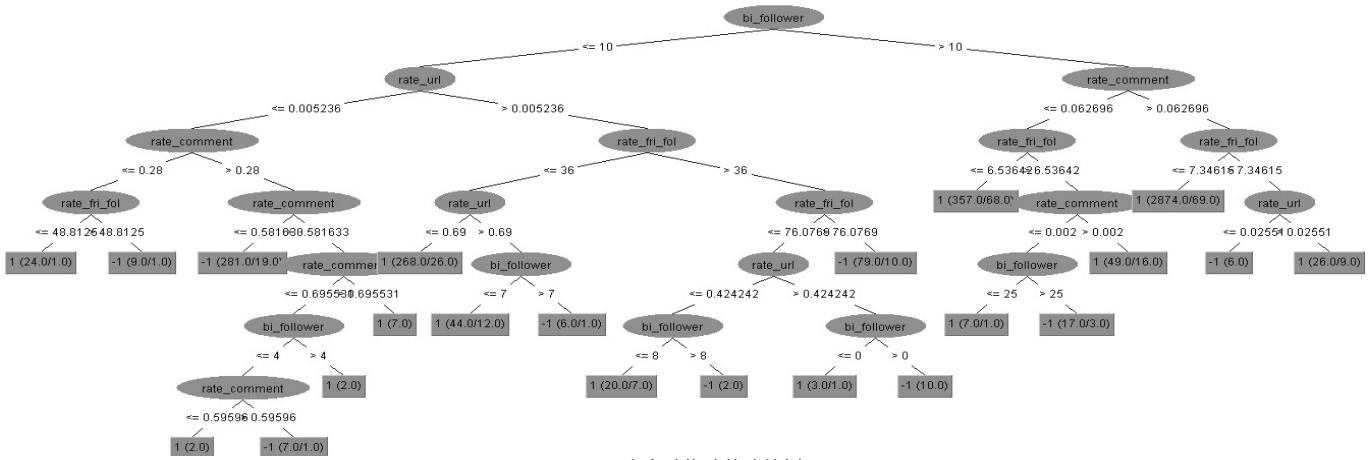

图 2 本实验构建的决策树

表 1 C4.5 决策树算法分类结果

Class	*T*	*F*	Weighted Avg
TP Rate	0.983	0.601	0.792
FP Rate	0.399	0.017	0.208
Precision	0.936	0.854	0.895
Recall	0.983	0.601	0.792
F-Measure	0.959	0.706	0.832
ROC Area	0.862	0.862	0.862

表 2 Adaboost 算法分类结果

Class	*T*	*F*	Weighted Avg
TP Rate	0.968	0.608	0.788
FP Rate	0.392	0.032	0.212
Precision	0.936	0.759	0.847
Recall	0.968	0.608	0.788
F-Measure	0.951	0.675	0.813
ROC Area	0.908	0.908	0.908

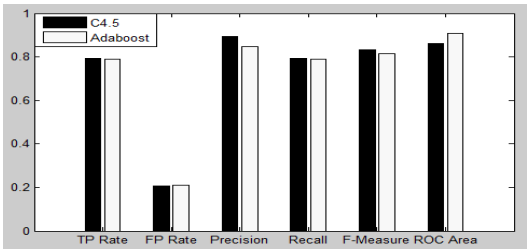

图 3 C4.5 决策树与 Adaboost 算法结果比较

7 结束语

垃圾用户检测问题是目前社交网络的难题，在一定程度上影响了社交网络的发展。我们基于多种垃圾用户及其行为特征对垃圾用户进行分类，可以快速检测出一部分垃圾用户，具有较高的准确率。由于社交网络中垃圾用户会以一些行为以逃避检测，一部分特征和规则很可能会失效，这给垃圾用户检测带来一定的困难。

参考文献

[1] 郭浩，陆余良，王宇，等.多特征微博垃圾互粉检测方法 [J]. 中国科技论文，2012，7（07）：548-551.

[2] 刘志明，刘鲁.微博网络舆情中的意见领袖识别及分析 [J]. 系统工程，2011，29（06）：8-16.

[3] Kurt Thomas，Chris Grier，Vern Paxson，et al.Suspended Accounts in Retrospect：An Analysis of Twitter Spam [G] //Proceedings of the 2011 ACM SIGCOMM conference on Internet Measurement Conference，Berlin：Germany，2011：243-258.

[4] Kyumin Lee，James Caverlee，Steve Webb.Uncovering Social Spammers：Social Honeypots+Machine Learning [G]//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval，Geneva：Switzerland，2010：435-442.

[5] Gianluca Stringhini，Christopher Kruegel，Giovanni Vigna.Detecting Spammers on Social Networks [G] // Proceedings of the 26th Annual Computer Security Applications Conference，Austin：Texas USA，2010：1-9.

[6] Alex Hai Wang.Detecting Spam Bots in Online Social Networking Sites：A Machine Learning Approach [J].Data and Applications Security and Privacy XXIV，2010：1-9.

[7] Yin Zhu，Xiao Wang，Erheng Zhong，等.Discovering Spammers in Social Networks [J]. AAAI，2012：171-177.

[8] 曹薇，张乃渊.一种基于 C4.5 决策树的 web 页面分类算法 [J].计算机系统与应用，2010，19（10）：195-198.

[9] 陈文才，黄金才.数据仓库与数据挖掘 [M].北京：人民邮电出版社，2004.

[10] 冯少荣，肖文俊.基于样本选择的决策树改进算法 [J]. 西南交通大学学报，2009，44（05）：643-647.

[11] 徐鹏，林森.基于 C4.5 决策树的流量分类方法 [J].软件学报，2009，20（10），2692-2704.

[12] Guan XQ. Research on the classifying algorithm based on decision tree [D] .Taiyuan：Shanxi University， 2006.

〔编辑：刘晓芳〕

Spam Analysis and Detection of Social Network based on Sina Weibo

Meng Xiangfei, Xu Lu, Wang Siyu

Abstract: With the rapid development of information technology and the Internet, the online social network（OSN）has played an irreplaceable role in our life. Meanwhile, however, OSN was plagued with various advertisements, which imposed severe impact to User Experience. A large amount of spam accounts registerd by malicious marketing teams also made normal users intolerable. In this paper, the reason why spam accounts appeared was illustrated firstly. Then we analyzed the feature of spam accounts. Finally, the Decision Tree algorithm was raised to classify accounts based on the data of Sina Weibo. The results show that the method is effective with a distinguishing rate of 92%.

Key words: OSN; Sina Weibo; spam account; classify