# 基于网络日志的用户行为刻画与预测研究

康海燕[1]， 王紫豪[1,2]， 于爱民[3]， 谭雨轩[1]

(1. 北京信息科技大学 信息管理学院 北京 100192; 2. 迈阿密大学 计算机科学系 美国佛罗里达州科勒尔盖布尔斯 33146; 3. 中国科学院 信息工程研究所 北京 100093)

**摘要:** 采用基于相似度的特征聚类算法以及粗糙集模糊分析法,提出了基于网络日志的用户性格特征分析及行为预测方法.首先,构建标准性格特征向量库;然后,采用基于余弦相似度的特征聚类算法进行性格分析,该算法解决了适量样本情况下的机器学习中聚类的问题,使训练模板数据即使在数据不是足够大的情况下仍能提取特征;最后,采用基于粗糙集理论的模糊分析算法进行行为预测,该分析算法简化了分析过程,减少了建模中需考虑的因素,又能得出精确的结果.对比实验表明,该方法能较准确地分析不同用户性格特征和对其未来行为进行预判,并分析出可能对安全领域造成威胁的人群.

**关键词:** 网络日志; 余弦相似度; 粗糙集模糊分析; 用户性格特征; 行为预测技术; 安全预警

**中图分类号:** TP391 **文献标志码:** A **文章编号:** 1671-6841(2019)03-0048-07

**DOI:** 10.13705/j.issn.1671-6841.2018272

## 0 引言

随着网络与信息资源的飞速发展,网络搜索引擎已成为人们获取信息的主要途径,网络搜索日志包含了用户的行为和需求,从网络日志可以判断出一个人的性格,甚至可以预测用户接下来要做的事情,这在安全领域尤其重要,可以根据用户接下来的行为判断哪些用户可归为危险人群,如黑客经常使用社会工程学的方法,利用人的弱点进行攻击.如用户信息泄露,犯罪分子在网络上搜索用户的身份信息、手机号码等实施盗取账号资金的目的.黑客首先进行信息侦探,收集名字、电话号码、身份证号等信息,从而伪装用户以实现对服务器端的欺骗,盗取用户账户.因此,如果安全部门通过分析网络日志,便可判断出危险的人群,甚至可以知道这类人群甚至特定的人接下来要做的事,就可以提前预警,防范危害的发生.目前,用户行为分析在日志中对用户所需要的信息检索和提取行为特征的研究已经取得一定的进展.

信息检索方面,文献[1]指出了在大数据时代下,搜索引擎采用云计算分布式架构,对各个服务器进行分析,并按查找结果的相似度高低进行排序.文献[2]提出了基于上下文分词的特征表示.将分词利用 Word2vector 进行向量表示,并进行聚类.文献[3]提出了基于后缀树聚类算法的搜索引擎,将具有多重含义的关键词搜索结果以聚类的方式呈现,避免用户花更多时间筛选有用的信息,提高搜索效率.

提取行为特征方面,文献[4]中提取了 chrome 浏览器的搜索日志,利用日志存储于数据库的特点,从数据库中提取数据,但分析结果仅局限于浏览器搜索日志已搜索过的关键词和现有的 URL,并未从日志中进行进一步的数据挖掘与预测,不具有延展性.文献[5]提出了针对用户的商业意图的架构,借助电商的浏览历史记录和商品网页点击情况,统计搜索词与商品类目之间的关系,采用概率论的方法进行预测,得出了用户商业意图,但该方法仅适用于电商领域,无法应用于复杂多变的社会工程环境.文献[6]提出了用户行为具有隐蔽性强、主动性强、复杂多样的特点.该研究将行为进行分类并采用 $K$-means 算法进行行为分析,但 $K$-means算法结果精确度不高.文献[7]基于 Spark 平台,以网络日志中的 URL 作为研究对象,并对 URL 进行查询扩展,结合 Word2vec 模型,采用类层次聚类算法 ISKHC 对搜索日志进行用户聚类分析,与 $K$-means 算法相比,ISKHC 的聚类质量更高.文献[8]提出了基于多层感知器神经网络的用户垃圾邮件的分类,分类效果

**收稿日期:** 2018-10-02

**基金项目:** 国家自然科学基金项目(61370139);北京市社会科学基金项目(15JGB099,15ZHA004);高水平人才交叉培养"实培计划"(科研)基金项目(71B1810826);信息+专项基金项目(5111823610).

**作者简介:** 康海燕(1971—),男,河北石家庄人,教授,主要从事网络安全与隐私保护研究,E-mail: kanghaiyan@126.com.

相比机器学习算法更加准确，但神经网络中权值的存储消耗较大的内存空间，训练需要大量样本.文献 [9] 提出了基于卷积神经网络的特征识别，权值共享的特点减少了多层感知器神经网络的空间复杂度和训练样本数量.文献 [10] 将以神经网络为主导的深度学习算法和机器学习算法进行比较，深度学习算法的效率和准确度优于机器学习算法.

因此，针对目前信息检索与提取行为特征技术的不足，本文参考文献 [7] 中将 URL 查询进行扩展和关键词聚类的方法，对上网时用户的搜索日志，采用基于机器学习的数据挖掘算法，提出了基于网络日志的用户性格特征分析及行为预测方法，进行关键词类目以及浏览日志搜索项之外的行为预测，从而提供更广泛、精确的安全预警.

## 1　实现方案

基于网络日志的用户行为分析与预测方法流程如图 1 所示.主要包括日志获取、日志预处理、核心算法以及输出结果 4 个模块.

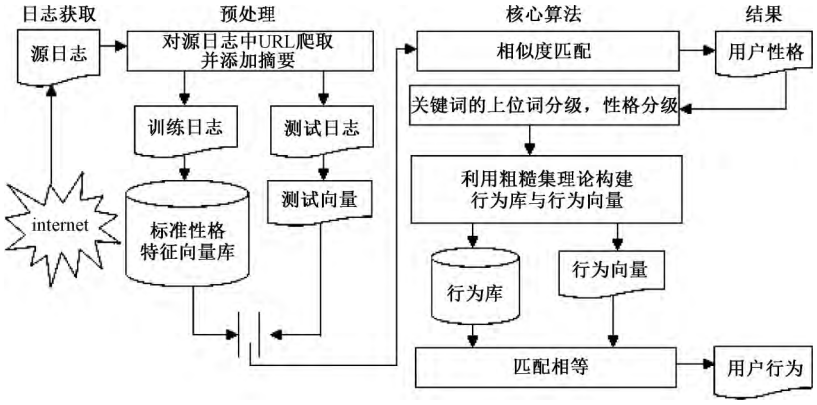


**图 1　基于网络日志的用户行为分析与预测方法流程**

**Fig.1**　Process of analysis and prediction of user behavior based on web log

### 1.1　日志获取

源日志主要来自于搜索引擎服务器或网络爬虫，将爬虫系统与各个站点连接，从而获取网络日志.目前常用的爬虫系统有百度统计、cnzz 等.本研究采用搜狗实验室 2008 年 6 月部分网页查询需求及用户点击情况的网页查询日志.数据格式为 “用户｜查询词｜该 URL 在返回结果中的排名｜用户点击的顺序号｜用户点击的 URL”.日志样本如表 1 所示.(本文选取个体用户查询信息多于 8 条的用户作为实验对象).

### 1.2　日志的预处理

**定义 1**　同义词集合. 同义词集合包含该词以及具有与该词含义相同或相近词的集合.

**定义 2**　上位词. 上位词指概念范围更广的词.如 “车”是 “汽车”的上位词 “交通工具”是 “车”的上位词.同位词集合包含于上位词中.

**1.2.1**　性格模型的构建　性格模型的构建包含两部分: (1) 性格的划分; (2) 上位词的分类与选取.本研究参照大五人格理论 [11]，性格包含开放性、责任心、外倾性、宜人性和情绪化人格.人们受社会环境、社会文化和各种思潮的影响，性格特点具有多样性，其多样性根据大五人格理论可大致分为积极、中级与消极 3 个方面.通过结合个人特点的自身优势、不足进行客观的评价.根据文献 [12]，人们接触到多元的价值观，接受主流思想，具有思维活跃、勇于展现自我、热情奔放、社会责任感强、团队合作意识、创新意识、感恩意识等特点，具体性格表现与相关分类如表 2 所示.上位词的分类可分为自然科学类词汇与社会科学类词汇.自然科学类词汇选取军事、科技、体育、旅游、食物 5 类; 社会科学类词汇选取史政、文艺、社会、娱乐、美容 5 类.划分的依据有: (1) 自然科学与社会科学本身的特点，例如科技、史政、文艺等标志明显词汇; (2) 依据人的性格特点，如社会责任感强的人，会讨论并分析国家时事，故有一定概率关注军事类内容.兴趣广泛的人通常性格开朗，会有一定概率关注体育类、旅游类内容，而体育类和旅游类需要团体的配合与周密的安排，符合自然科学类的特点.社会科学类词汇倾向于发散思维，考虑到多方面，例如社会、娱乐.社会科学类思维的人不喜拘束，随

表1 日志样本

Tab.1 Log samples

| 用户 | 查询词 | 排名 | 顺序 | 用户点击的URL |
|---|---|---|---|---|
| 125254918559 | [王强] | 3 | 1 | ent.163.com/05/1231/14/26ABFP16000300C8.html |
| 125254918559 | [王强] | 4 | 2 | ent.qq.com/a/20060317/000195.htm |
| 125254918559 | [王强] | 10 | 3 | www.tyyue.com/song/124582.htm |
| 125254918559 | [郭美美] | 7 | 1 | www.520music.com/Albumlist/520music.com_2458.htm |
| 125254918559 | [郭美美] | 6 | 2 | www.znsjw.com/v/ShowSoft.asp? SoftID=5129 |
| 125254918559 | [郭美美] | 4 | 3 | www.y130.com/htm/7044.htm |
| 125254918559 | [郭美美] | 3 | 4 | ent.qq.com/a/20060228/000178.htm |
| 828687165269 | [村妓] | 4 | 1 | pr.overnightlending.com/8/93.htm |
| 828687165269 | [村妓] | 5 | 2 | blog.sohu.com/members/fgmnghdg/1094076.html |
| 828687165269 | [村妓] | 4 | 3 | pr.overnightlending.com/8/93.htm |
| 828687165269 | [林彪] | 10 | 1 | nr.book.sohu.com/20050501/n225414699.shtml |
| 828687165269 | [富婆] | 3 | 1 | news.sina.com.cn/s/2006-01-10/03 537 937 753 s.shtml |
| 828687165269 | [富婆] | 5 | 1 | news.qq.com/a/20060110/000776.htm |
| 828687165269 | [富婆找鸭子] | 2 | 1 | www.eastvenus.com/viewthread.php? tid=35523&extra=page%3D4 |
| 828687165269 | [富婆找鸭子] | 5 | 2 | www.jl366.com/2008/photo.asp? id=261 |
| 828687165269 | [富婆找鸭子] | 9 | 3 | blog.sohu.com/members/gfjyktu/1179025.html |

性,故有一定概率关注美容类,其涉猎范围广,人们关注范围广,符合社会科学类的性格特点.包含上位词的用户标准性格特征向量库如表2所示.表2中数字代表不同性格的人群搜索的具有相同上位词的关键词个数的百分比的统计平均值.

表2 性格模型与性格关键词分类

Tab.2 Classification of character models and character keywords %

| 性格 | | 自然科学类词汇 | | | | | 社会科学类词汇 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 军事 | 科技 | 体育 | 旅游 | 食物 | 史政 | 文艺 | 社会 | 娱乐 | 美容 |
| 积极性格 | 具有较强的社会责任感 | 27 | 10 | 1 | 1 | 1 | 29 | 1 | 28 | 1 | 1 |
| | 兴趣爱好广泛 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| | 自信阳光,乐于交往,热爱生活,充满热情 | 1 | 10 | 26 | 20 | 10 | 1 | 2 | 1 | 28 | 1 |
| | 家庭观念 | 15 | 20 | 1 | 1 | 1 | 20 | 20 | 20 | 1 | 1 |
| 中级性格 | 积极上进,吃苦耐劳,创新意识强 | 10 | 28 | 1 | 1 | 1 | 28 | 1 | 28 | 1 | 1 |
| | 讲究实效,注重实际,原则性强 | 20 | 28 | 1 | 1 | 1 | 10 | 1 | 28 | 9 | 1 |
| | 争强好胜 | 11 | 15 | 10 | 5 | 5 | 10 | 20 | 20 | 2 | 2 |
| | 以自我为中心 | 1 | 1 | 1 | 10 | 10 | 1 | 1 | 20 | 29 | 26 |
| 消极性格 | 自制力弱,自制力较弱,没耐心,懒散 | 1 | 1 | 10 | 10 | 5 | 1 | 1 | 1 | 40 | 30 |
| | 抗打击能力与自我愈合能力弱 | 10 | 5 | 10 | 1 | 18 | 5 | 10 | 20 | 20 | 1 |
| | 缺乏沟通技巧 | 10 | 40 | 5 | 1 | 1 | 1 | 10 | 30 | 1 | 1 |
| | 社会经验不足 | 2 | 10 | 10 | 10 | 2 | 1 | 10 | 5 | 40 | 10 |

**1.2.2 预处理算法** 预处理算法如下(算法1).

输入: 用户的原始日志.

输出: 用户的特征向量.

步骤1 首先对原始日志中URL进行爬取,获取网页摘要并加入用户日志中;

步骤2 找出日志中的关键词,统计搜索关键词的个数;

步骤3 统计关键词的上位词出现的个数;

步骤4 求得每类上位词占所有上位词的比例并将比值以百分比形式构成特征向量.

算法1举例: 输出结果例如表1样本中用户125254918559和828687165269,搜索的上位词比例构成的向量分别为125254918559: (0.0, 20.0, 0.0, 0.0, 0.0, 40.0, 20.0, 0.0,0.0, 20.0); 828687165269:

(25.0,56.25, 0.0, 0.0, 0.0,0.0,12.5,6.25,0.0,0.0).

### 1.3　核心算法

核心算法主要包含两个算法:(1) 基于余弦相似度特征聚类的性格分析算法;(2) 基于粗糙集模糊分析的行为预测算法.基于余弦相似度特征聚类的性格分析算法在构建上述标准性格特征向量库的基础上,对新日志中各类上位词的比例进行统计,与向量库中每一行进行余弦相似度比较,夹角越小,越接近其性格特征向量.求得余弦值最大的分量,即夹角最小的分量,即可得出其性格特征.基于粗糙集模糊分析的行为预测算法采用知识简约的方法,将性格特征向量库中的性格进一步分为积极、中级和消极 3 个等级.关键词的上位词进一步分为自然科学类和社会科学类词汇,将两类词汇的搜索比例分为 3 个等级.分析 3 个因素(性格等级,自然科学词汇等级,社会科学词汇等级)的不同组合,共有 $3^3 = 27$ 种组合,将 27 种组合参考霍兰德职业性格倾向测试结果[13],可得具有 27 种行为方向的行为库.将用户日志中的 3 个因素做相同处理,与行为库进行比对,输出符合的行为.

**1.3.1　核心算法的理论基础**　(1) 相似度计算方法.常用的相似度算法包括欧氏距离相似度和余弦相似度[14-15].本研究采用余弦相似度进行性格分析.因为余弦相似度是从方向上区分差异,而对绝对的数值不敏感,所以更多用于用户对内容评分来区分兴趣的相似度和差异,同时修正了用户间可能存在的度量标准不统一的问题.因此本文采用余弦相似度算法来计算用户性格模型之间的相似度.(2) 粗糙集理论.粗糙集是一种刻画不完整性和不确定性的数学工具[16-17],能有效地分析不精确、不一致、不完整等各种不完备的信息,还可以对数据进行分析和推理,从中发现隐含的知识,揭示潜在的规律.粗糙集理论的主要思想是利用已知的知识库,将不精确或不确定的知识用已知的知识库中的知识近似刻画.粗糙集理论核心是知识简约.

**定义 3**　近似. $K=(U,S)$ 为给定的知识库,$U$ 表示论域,$S$ 为 $U$ 上的等价关系簇.则 $\forall X \subseteq U$ 和 $U$ 上的一个等价关系 $R \in IND(K)$,如图 2 所示,子集 $X$ 关于知识 $R$ 的上近似和下近似分别为:

$$\overline{R}(X) = \{x \mid (\forall x \in U) \wedge ([x]_R \cap X \neq \varnothing)\}, \underline{R}(X) = \{x \mid (\forall x \in U) \wedge ([x]_R \subseteq X)\}.$$

**定义 4**　知识简约. 知识库中的知识并非同等重要,有知识冗余.知识简约是将一些五环或多余特征丢弃,在不影响原有分析的前提下将信息量减少.即在不影响原知识分类的情况下,将 $n$ 维信息空间 $\{x_1,x_2,\cdots,x_n\}$ 减小为 $m$ 维信息空间 $\{x_1,x_2,\cdots,x_m\}$,$(m<n)$.通过约简,产生新的决策规则.

知识表达系统是粗糙集理论中主要的知识表示方法,记为 $S=(U,A,V,f)$,其中: $U$ 表示对象的非空有限集合,即论域; $A$ 为属性的非空有限集合,即属性集; $V=\cup_{\alpha \in A} V_\alpha$,$V_\alpha$ 表示属性 $\alpha$ 的值域; $f$ 为 $U \times A \rightarrow V$,为信息函数,为每个对象的每个属性赋予一个信息值,即 $\forall \alpha \in A, x \in U, f(x,\alpha) \in V_\alpha$. 因此知识表达可以用表格来实现.如表 3 所示,其中: $U=\{x_1,x_2,\cdots,x_n\}$ 为论域; $A=\{P,P_2,\cdots,P_n\}$ 为全体属性集合.

**表 3　知识表达系统**

**Tab.3**　Knowledge representation system

| $U$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 0 | 1 | 1 |
| $x_2$ | 1 | 0 | 1 | 1 | 1 |
| $x_3$ | 0 | 1 | 0 | 1 | 0 |
| $x_4$ | 0 | 0 | 1 | 0 | 1 |
| $x_5$ | 0 | 0 | 0 | 0 | 0 |

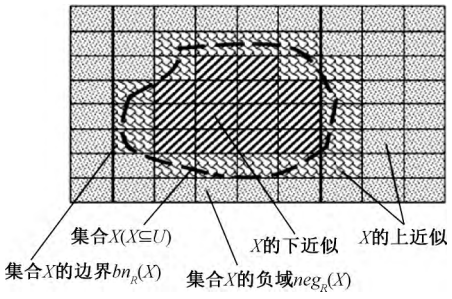


**图 2**　集合 $X$ 的上近似、下近似、边界域示意图

**Fig.2**　Schematic of the upper approximation, lower approximation, and boundary domain of set $X$

**1.3.2　基于余弦相似度特征聚类的性格分析算法**　从未知性格的新用户中提取出日志中的关键字并统计其上位词分别出现的频率,求得统计平均值.用关键字的上位词所占比例构建特征向量,并与标准性格特征向量库中的分量进行相似度计算.得出的差值越小,表明越接近该行向量,进而得出用户在标准性格特征向量库中最接近的性格,即为本用户的性格.

基于余弦相似度特征聚类的性格分析算法如下(算法 2).

输入: 用户的特征向量(由算法1输出生成).

输出: 用户的性格特征.

步骤1 构建性格模型向量集合;

步骤2 构建用户测试向量;

步骤3 相似度比较,找出特征向量库中余弦值最大的分量;

步骤4 输出该分量对应的性格特征.

算法2举例.输出结果例如表1样本中用户125254918559的性格特征: 家庭观念强、自立自强、尊老爱幼(较好的家庭观念/体恤父母/自立自强/理解父母的艰辛与良苦用心/吃苦耐劳,适应力强),归为积极性格; 用户828687165269的性格特征为: 缺乏沟通技巧(沟通不畅/表达能力差/与他人关系紧张/朋友圈子较小/容易遭遇尴尬),归为消极性格.

用户性格趋向是用户性格特征向量与标准性格特征向量库中行向量的相似度的反映.矩阵$M$为标准性格特征向量库,$C_1 \sim C_n$表示上位词,$W_1 \sim W_n$表示上位词比例权重的平均值.

$$M=\begin{bmatrix}(C_1,W_1) & (C_1,W_2) & \cdots & (C_1,W_m)\\ (C_2,W_1) & (C_2,W_2) & \cdots & (C_2,W_m)\\ \vdots & \vdots & & \vdots\\ (C_n,W_1) & (C_n,W_2) & \cdots & (C_n,W_m)\end{bmatrix}.$$

本研究中$\vec{r}_1$代表标准性格特征向量库的一个分量,$\vec{r}_2$代表当前用户性格特征向量,其中$\vec{r}_1=((C_i,W_1),(C_i,W_2),\cdots,(C_i,W_{10}))$,$\vec{r}_2=(tw_1,tw_2,\cdots,tw_{10})$. 利用两个向量的余弦夹角计算出两个向量的相似程度.

**1.3.3 基于粗糙集模糊分析的行为预测算法** 首先将性格与各个上位词搜索百分比作为知识库.然后将性格简约为积极、中级和消极.上位词简约为自然科学类词汇的百分比和社会科学类词汇的百分比.最后根据百分比所在不同区间,分别将自然科学类词汇和社会科学类词汇简约为3个等级.

基于粗糙集模糊分析的行为预测算法如下(算法3).

输入: 用户的性格特征(由算法2输出生成),用户的特征向量(由预处理算法输出生成).

输出: 用户的行为预测.

步骤1 构建长度为4的索引表,表的二维数组作为行为分析库;

步骤2 构建长度为3的一维数组分别表示性格等级、自然科学词汇等级与社会科学词汇等级;

步骤3 将用户测试向量中关键词按照自然科学类和社会科学类分别求统计平均值并划分等级,存入一维数组对应的空间中;

步骤4 划分用户性格等级,存入一维数组对应的空间中;

步骤5 将一维数组在行为分析库中匹配对应的行,输出行为.

算法3举例.输出结果例如表1样本中用户125254918559的行为预测结果: 理科类与文科类兴趣比例约为1∶3(社会责任感强,擅长分析政治局势/关注国家政策/喜欢政治类的新闻、报纸、名人自传); 用户828687165269的行为预测结果: 理科类与文科类兴趣比例约为3∶1(自闭/对科学内容有较高理解力,如计算机、物理、生物/可能采取技术性的极端行为,对他人造成人身/财产伤害).

首先将性格按照积极、中级、消极划分为3个等级; 然后将用户搜索日志中自然科学类中5类词汇的比例求和,得到自然科学类词汇的比例.将百分比按照$[0,33)$,$[33,66)$,$[66,100]$3个区间分为3个等级.社会科学类词汇同理.因此,$A=\{$性格等级,自然科学类词汇等级,社会科学类词汇等级$\}$,$U=\{x_1,x_2,\cdots,x_{27}\}$. 对属性集不同元素的27种组合,根据霍兰德职业性格倾向测试结果对每种情况分别进行分析,建立行为分析库,行为的集合即为$U$.其中消极性格记为1,中级性格记为2,积极性格记为3.自然科学和社会科学词频将百分比按照$[0,33)$记为1,$[33,66)$记为2,$[66,100]$记为3.$x1\sim x27$表示行为的编号.

最后将用户的属性集合与行为分析库比对,找到使用户属性集与行为分析库中的属性集相等的记录,输出符合的行为特征.

## 2　实验结果及分析

### 2.1　实验结果

操作系统为 Win7 32 位/64 位,采用 Java 语言编写,所用 IDE 为 Eclipse Java Mars.采用 2008 年搜狗实验室发布的开源匿名网络日志,选取一天内的数据,共 51.0 MB,包含 1 030 577 条日志,选取搜索项大于 8 条的个体用户作为测试数据.本实验在搜狗实验室日志中用随机抽样的方法选取 10 000 个符合条件的用户进行测试.

### 2.2　性能分析

实验内容包括 3 部分:(1) 测试个体用户性格分析的准确性;(2) 测试个体用户行为预测的准确性;(3) 测试群体中危险用户的比例.测试选取不同日志条数下的多个用户,剔除上位词频率分布仅集中在个别上位词的用户,分别进行日志条数与性格准确率关系、日志条数与行为准确率关系的测试.

**2.2.1　测试个体用户性格分析的准确性**　实验选取不同搜索条数的测试日志进行实验,测试日志不包括训练日志,性格准确率定义为分析出的性格条目描述(共 5 条)在源日志上位词中的命中率,即符合上位词内容的条目/总条数.命中率范围为(0,1).命中率越高,说明该系统对个体用户的评价越准确.本实验将不同用户按照日志条数与性格准确率作散点图,并进行曲线拟合.测试结果如图 3 所示,其中每个点代表一个用户,横坐标对应值代表该用户日志的条数,纵坐标表示准确率.

结果分析:(1) 随着个体日志条数的增加,个体性格准确率呈现先上升,后下降,再上升的趋势;(2) 准确率呈现上升趋势的日志条数区间约为(0,13)、(30,+∞),下降趋势区间为(13,30);(3) 开始日志条数较少时,随着日志条数的增加,准确率会相应增加.当日志条数相对较多时,用户搜索结果出现多样性,而与性格倾向有所偏移,形成干扰.当样本容量足够大时,符合用户性格的搜索所占比例远大于其他搜索项.因此日志条数足够多时,性格准确率显著增加.

**2.2.2　测试个体用户行为预测的准确性**　行为准确率定义与测试性格分析准确性定义类似.区别在于行为预测条数总共为 3 条.测试结果如图 4 所示.

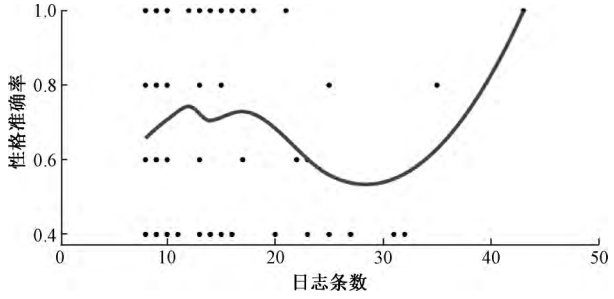


图 3　个体日志条数与性格准确率关系

**Fig.3**　Relationship between individual log number and character accuracy rate

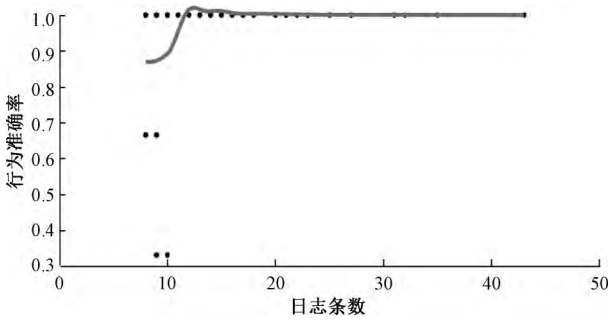


图 4　个体日志条数与行为准确率关系

**Fig.4**　Relationship between individual log number and behavior accuracy rate

结果分析:(1) 随着用户日志条数的增加,行为准确率呈先上升,后平稳的趋势;(2) 行为准确率上升的日志条数区间约为(0,13),当日志条数大于 13 时,准确率趋近于 1;(3) 当日志条数较少时,随着日志条数的增加,行为准确率增加.当日志条数足够时,由于采用粗糙集模糊分析法将性格重新进行分类,减少了性格之间的波动性,故该模型能够有效地进行行为预测.

**2.2.3　测试群体中危险用户的比例**　本实验用 10 000 个符合条件的用户进行测试,进而评估实验数据中 100 000 个用户积极、中级、消极性格用户所占比例以及危险信号与非危险信号所占比例.其中,积极性格用户约占 66%,中级性格用户约占 31%,消极性格用户约占 3%.三类用户中具有危险信号的用户约占 3%.实验证明该系统能够准确分析出危险人群.

结果分析:(1) 成功筛选出了危险人群.如表 1 所示用户 828687165269,其搜索关键词包含如“村妓”、“富婆”等词汇,用户极有可能受到不良信息的影响,并产生违法犯罪的行为;(2) 用户群体中积极性格用户占 1/3,中级性格用户占 2/3,极个别用户为消极性格;(3) 危险人群多集中在消极性格用户中.

2.2.4 实验结果对比 本实验在单机日志分析系统进行,随着样本数量的增加,评估用户性格与行为预测的准确率逐渐增加.文献[18]采用基于Hadoop的分布式日志分析系统,用*K*-means算法将用户分成5类.相比之下,本研究采用粗糙集模糊分析法的不足之处在于,行为的分析是根据先验经验建立行为模型,而非从大量日志数据分析行为类型,因此对用户行为的描述无法细化.

文献[7]的大数据平台采用3台虚拟机的Spark平台,将分类体系分为导航类、信息类和资源类.实验日志采用搜狗实验室2011年搜索日志,共100万条.去除缺失数据的日志,从中随机挑选7 000条.其中训练日志有1 000条,测试日志有6 000条.该实验进行了4组测试,分类算法和询问机制分别为:决策树+归一化熵值装袋法;朴素贝叶斯+归一化熵值装袋法;决策树+IK-Means结合归一化熵值装袋法和朴素贝叶斯结合IK-Means+归一化熵值装袋法.4组实验对3类日志分类的精确度均大于0.7,查全率均大于0.67.在该基准下,实验2.2.1在用户日志条数大于37条时,性格的精确度大于0.7;在实验2.2.2中,日志条数仅在大于13条时,行为准确率接近于1,证明该系统能够准确分析用户的行为.

## 3 结束语

安全领域所面临的威胁正逐渐从一般的纯技术攻击向结合社会工程学的智能化攻击转变.由于这种危险行为具有隐蔽性的特点,运用通常意义上的安全防范技术很难将其发现并防御.但运用网络大数据用户行为预测分析技术可以有效地做到准确分析、安全预警.综上所述,基于网络日志的用户性格特征分析及行为预测方法具有以下的创新性:第一,提出基于行为预测的主动防御方法;第二,将特征聚类与模糊分析的智能算法应用于本研究.本研究分析结果的准确性高.余弦相似度算法解决了适量样本情况下的机器学习中聚类的问题,使训练模板数据即使在数据不是足够大的情况下仍能提取特征.粗糙集模糊分析法简化了分析过程,减少了建模中需考虑的因素,又能得出精确的结果.

## 参考文献:


[1] 万冬娥.基于云计算的大数据信息检索技术[J].电子技术与软件工程,2018(3):176-176.

[2] 周顺先,蒋励,林霜巧,等.基于Word2vector的文本特征化表示方法[J].重庆邮电大学学报(自然科学版),2018,30(2):272-279.

[3] 陈建华.基于后缀树聚类算法的元搜索引擎的设计与实现[D].长春:吉林大学,2017.

[4] 杨雪,靳慧云.Chrome浏览器历史记录提取与分析[J].计算机应用与软件,2016,33(12):313-317.

[5] 徐晓峰.大规模用户在线行为数据分析[D].上海:上海交通大学,2013.

[6] 周雪.基于网络日志的用户行为分析与研究[D].北京:北京邮电大学,2017.

[7] 周三多.基于大数据平台的搜索日志分析技术研究[D].重庆:重庆邮电大学,2017.

[8] BOGAWAR P S, BHOYAR K K. Soft computing approaches to classification of emails for sentiment analysis[C]//Proceedings of the International Conference on Informatics and Analytics-ICIA-16. Pondicherry, 2016: 1-7.

[9] SZEGEDY C, LIU W, JIA Y A, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, 2015: 1-9.

[10] PATERAKIS N G, MOCANU E, GIBESCU M, et al. Deep learning versus traditional machine learning methods for aggregated energy demand prediction[C]//IEEE PES Innovative Smart Grid Technologies Conference. Torino, 2017: 1-6.

[11] BAGBY R M, WIDIGER T A. Five factor model personality disorder scales: an introduction to a special section on assessment of maladaptive variants of the five factor model[J].Psychological assesment,2018,30(1):1-9.

[12] 冯海龙.当代大学生性格特点分析[J].统计与管理,2015(1):78-79.

[13] 李乐.基于职业性格特点开展大学生专业教育及职业指导的探索[J].长春教育学院学报,2015,31(24): 121-123.

[14] 李清华,康海燕,苑晓姣,等.个性化搜索中用户兴趣模型匿名化研究[J].西安交通大学学报,2013,47(4):131-136.

[15] 康海燕,马跃雷,苑晓姣,等.面向网络搜索日志的发布方法研究[J].信息安全研究,2016,2(3):251-257.

[16] 于洪,王国胤,姚一豫.决策粗糙集理论研究现状与展望[J].计算机学报,2015,38(8):1628-1639.

[17] 韩朝,苗夺谦,任福继,等.基于粗糙集知识发现的开放领域中文问答检索[J].计算机研究与发展,2018,55(5):958-967.

[18] 邓小盾.一种基于大数据的网络日志分析模型构建研究[J].电子设计工程,2017,25(23):97-100.


(下转第60页)

# Multi-scale Region Features Algorithm for Fine-grained Classification

XIONG Changzhen, JIANG Jie
(*Beijing Key Lab of Urban Intelligent Traffic Control Technology, North China University of Technology, Beijing* 100144, *China*)

**Abstract:** Intending to reduce the influence of complex background on fine-grained classification, as well as to study the global information and local information of the target objects extracted from the convolutional neural network for fine-grained tasks, a fine-grained classification method based on multi-scale region feature was proposed. The method FASTER-RCNN framework was to train three convolution models to locate multi-scale object regions. Then the bounding box constraint and Helen constraint were applied to improve the location accuracy of the detected object. Finally, the extracted multi-scaled region features were combined to train a SVM classifier for fine-grained classification. The proposed method was tested in Caltech-UCSD bird datasets and CompCars vehicle datasets. The results showed that the accuracy of classification in Caltech-UCSD bird datasets was 82.8%. It increased by 7.5 % than the method without multi-scale region features. Compared with part-based RCNN, it increased by 8.9 %. The results showed that the accuracy of classification in CompCars was 93.5%. It increased by 8.3 % than the method without multi-scale region features. Compared with GoogleNet, it increased by 2.3 %.

**Key words:** fine-grained recognition; neural network fine-tuning; box constraint; Helen constraint algorithm

(责任编辑: 方惠敏)

(上接第54页)

# Analysis and Prediction of User Behavior Based on Web Log

KANG Haiyan[1], WANG Zihao[1,2], YU Aimin[3], TAN Yuxuan[1]
(1. *School of Information Management, Beijing Information Science and Technology University, Beijing* 100192, *China*;
2. *Department of Computer Science, University of Miami, Coral Gables, FL* 33146, *USA*;
3. *Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093, *China*)

**Abstract:** The feature clustering algorithm based on similarity and the fuzzy analysis method based on rough set were used. A method of analysis and prediction of user behavior based on web log was proposed. Firstly, a standard character eigenvector library was constructed. Then, a character clustering algorithm based on cosine similarity was used for character analysis. Finally, a fuzzy analysis algorithm based on rough set theory was used to perform behavior prediction. Results showed that the method accurately analyze the personality characteristics of users and predict their future behaviors, and identify the groups that might pose a threat to the security field.

**Key words:** web log; cosine similarity; rough set fuzzy analysis; user personality trait; behavior prediction technology; security warning

(责任编辑: 方惠敏)