Result List (1)

Select All

Download

Add To Favorites

Add To Analysis DB

CN109783460A User behavior depicting and predicting method and system based on web log Viewing

CN109783460A [Chinese] **CN109783460A [English]**

Invention Title -- User behavior depicting and predicting method and system based on web log

Application No.	CN:201910089017:A
Application Date	2019.01.30
Publication No.	CN109783460A
Publication Date	2019.05.21
IPC Classification No.	G06F16/18; G06K9/62
Applicant/Assignee	UNIV BEIJING INF SCI & TECH;
Inventor	KANG HAIYAN;WANG ZIHAO;
Priority No.	CN201910089017
Priority Date	2019.01.30
CPC	

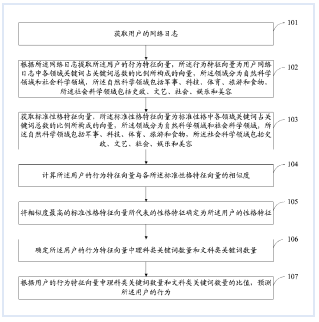
Abstract [Support Block Translation]

Chinese->English

English->Chinese Other

Abstract: The invention discloses a user behavior depicting and predicting method and system based on a web log. The method comprises the steps of obtaining a web log of a user; Extracting a behavior feature vector of the user according to the web log; Obtaining a standard character feature vector; calculating the similarity between the behavior feature vector of the user and each standard character featurevector; determining the character feature represented by the standard character feature vector with the highest similarity as the character feature of the user; determining the number of the theoretical keywords and the number of the literature keywords in the behavior feature vectors of the user; and predicting the behavior of the user according to the ratio of the number of the theoretical keywords to the number of the literature keywords in the behavior feature vectors of the user. According to the network log-based user behavior depicting and predicting method and system provided by the invention, the character and behavior of the user can be predicted, so that data support is provided for preventing the occurrence of harm.

Graphs



(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 109783460 A
(43)申请公布日 2019.05.21

(21)申请号 201910089017.1

(22)申请日 2019.01.30

(71)申请人 北京信息科技大学
地址 100000 北京市海淀区清河小营东路12号

(72)发明人 康海燕 王紫豪

(74)专利代理机构 北京高沃律师事务所 11569
代理人 杜阳阳

(51)Int.Cl.
G06F 16/18(2019.01)
G06K 9/62(2006.01)

权利要求书2页 说明书13页 附图3页

(54)发明名称

基于网络日志的用户行为刻画与预测方法及系统

(57)摘要

本发明公开了一种基于网络日志的用户行为刻画与预测方法及系统。该方法包括:获取用户的网络日志;根据网络日志提取用户的行为特征向量;获取标准性格特征向量;计算用户的行为特征向量与各标准性格特征向量的相似度;将相似度最高的标准性格特征向量所代表的性格特征确定为用户的性格特征;确定用户的行为特征向量中理科类关键词数量和文科类关键词数量;根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值,预测用户的行为。本发明提供的基于网络日志的用户行为刻画与预测方法及系统能够对用户的性格、行为进行预测,进而,为防范危害的发生提供数据支持。

CN 109783460 A

1.一种基于网络日志的用户行为刻画与预测方法,其特征在于,包括:

获取用户的网络日志;

根据所述网络日志提取所述用户的行为特征向量,所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量,所述领域分为自然科学领域和社会科学领域,所述自然科学领域包括军事、科技、体育、旅游和食物,所述社会科学领域包括史政、文艺、社会、娱乐和美容;

获取标准性格特征向量,所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量,所述领域分为自然科学领域和社会科学领域,所述自然科学领域包括军事、科技、体育、旅游和食物,所述社会科学领域包括史政、文艺、社会、娱乐和美容;

计算所述用户的行为特征向量与各所述标准性格特征向量的相似度;

将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征。

2.根据权利要求1所述的基于网络日志的用户行为刻画与预测方法,其特征在于,所述方法还包括:

确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量;

根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值,预测所述用户的行为。

3.根据权利要求1所述的基于网络日志的用户行为刻画与预测方法,其特征在于,所述计算所述用户的行为特征向量与各所述标准性格特征向量的相似度,具体包括:

计算所述用户的行为特征向量与各所述标准性格特征向量的余弦相似度;

将余弦相似度最小的标准性格特征向量确定为与所述用户行为特征向量相似度最大的标准性格特征向量。

4.根据权利要求1所述的基于网络日志的用户行为刻画与预测方法,其特征在于,所述将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征,具体包括:

将所述标准性格特征向量划分为积极性格、中级性格和消极性格三种类型;

将与用户行为特征向量相似度最大的标准性格特征向量所属类型确定为所述用户的性格类型。

5.根据权利要求2所述的基于网络日志的用户行为刻画与预测方法,其特征在于,所述根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值,预测所述用户的行为,具体包括:

当所述用户的行为特征向量中理科类关键词数量与文科类关键词数量的比值为3:1时,预测所述用户有对他人造成伤害的可能性。

6.一种基于网络日志的用户行为刻画与预测系统,其特征在于,包括:

网络日志获取模块,用于获取用户的网络日志;

用户行为特征向量提取模块,用于根据所述网络日志提取所述用户的行为特征向量,所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量,所述领域分为自然科学领域和社会科学领域,所述自然科学领域包括军事、科技、体育、旅游和食物,所述社会科学领域包括史政、文艺、社会、娱乐和美容;

标准性格特征向量获取模块，用于获取标准性格特征向量，所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

相似度计算模块，用于计算所述用户的行为特征向量与各所述标准性格特征向量的相似度；

用户性格刻画模块，用于将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征。

7.根据权利要求6所述的基于网络日志的用户行为刻画与预测系统，其特征在于，所述系统还包括：

关键词数量确定模块，用于确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量；

用户行为预测模块，用于根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为。

8.根据权利要求6所述的基于网络日志的用户行为刻画与预测系统，其特征在于，所述相似度计算模块，具体包括：

相似度计算单元，用于计算所述用户的行为特征向量与各所述标准性格特征向量的余弦相似度；

性格确定单元，用于将余弦相似度最小的标准性格特征向量确定为与所述用户行为特征向量相似度最大的标准性格特征向量。

9.根据权利要求6所述的基于网络日志的用户行为刻画与预测系统，其特征在于，所述用户性格刻画模块，具体包括：

性格类型划分单元，用于将所述标准性格特征向量划分为积极性格、中级性格和消极性格三种类型；

用户性格刻画单元，用于将与用户行为特征向量相似度最大的标准性格特征向量所属类型确定为所述用户的性格类型。

10.根据权利要求7所述的基于网络日志的用户行为刻画与预测系统，其特征在于，所述用户行为预测模块，具体包括：

用户行为预测单元，用于当所述用户的行为特征向量中理科类关键词数量与文科类关键词数量的比值为3:1时，预测所述用户有对他人造成伤害的可能性。

基于网络日志的用户行为刻画与预测方法及系统

技术领域

[0001] 本发明涉及一种基于网络日志的用户行为刻画与预测方法及系统。

背景技术

[0002] 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来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。

发明内容

[0003] 本发明的目的是提供一种基于网络日志的用户行为刻画与预测方法及系统，能够对用户的性格进行刻画预测，进而，根据用户性格预测用户的危险性，为防范危害的发生提供数据支持。

[0004] 为实现上述目的，本发明提供了如下方案：

[0005] 一种基于网络日志的用户行为刻画与预测方法，包括：

[0006] 获取用户的网络日志；

[0007] 根据所述网络日志提取所述用户的行为特征向量，所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0008] 获取标准性格特征向量，所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0009] 计算所述用户的行为特征向量与各所述标准性格特征向量的相似度；

[0010] 将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征。

[0011] 可选的，确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量；

[0012] 根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为。

[0013] 可选的，所述计算所述用户的行为特征向量与各所述标准性格特征向量的相似度，具体包括：

[0014] 计算所述用户的行为特征向量与各所述标准性格特征向量的余弦相似度；

[0015] 将余弦相似度最小的标准性格特征向量确定为与所述用户行为特征向量相似度最大的标准性格特征向量。

[0016] 可选的，所述将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征，具体包括：

[0017] 将所述标准性格特征向量划分为积极性格、中级性格和消极性格三种类型；

[0018] 将与用户行为特征向量相似度最大的标准性格特征向量所属类型确定为所述用户的性格类型。

[0019] 可选的，所述根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为，具体包括：

[0020] 当所述用户的行为特征向量中理科类关键词数量与文科类关键词数量的比值为3:1时，预测所述用户有对他人造成伤害的可能性。

[0021] 本发明还提供了一种基于网络日志的用户行为刻画与预测系统，包括：

[0022] 网络日志获取模块，用于获取用户的网络日志；

[0023] 用户行为特征向量提取模块，用于根据所述网络日志提取所述用户的行为特征向量，所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0024] 标准性格特征向量获取模块，用于获取标准性格特征向量，所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0025] 相似度计算模块，用于计算所述用户的行为特征向量与各所述标准性格特征向量的相似度；

[0026] 用户性格刻画模块，用于将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征。

[0027] 可选的，关键词数量确定模块，用于确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量；

[0028] 用户行为预测模块，用于根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为。

[0029] 可选的，所述相似度计算模块，具体包括：

[0030] 相似度计算单元，用于计算所述用户的行为特征向量与各所述标准性格特征向量的余弦相似度；

[0031] 性格确定单元，用于将余弦相似度最小的标准性格特征向量确定为与所述用户行为特征向量相似度最大的标准性格特征向量。

[0032] 可选的，所述用户性格刻画模块，具体包括：

[0033] 性格类型划分单元，用于将所述标准性格特征向量划分为积极性格、中级性格和

消极性格三种类型；

[0034] 用户性格刻画单元，用于将与用户行为特征向量相似度最大的标准性格特征向量所属类型确定为所述用户的性格类型。

[0035] 可选的，所述用户行为预测模块，具体包括：

[0036] 用户行为预测单元，用于当所述用户的行为特征向量中理科类关键词数量与文科类关键词数量的比值为3:1时，预测所述用户有对他人造成伤害的可能性。

[0037] 根据本发明提供的具体实施例，本发明公开了以下技术效果：本发明提供的基于网络日志的用户行为刻画与预测方法及系统，根据用户网络日志提取用户的行为特征向量，计算用户的行为特征向量与各标准性格特征向量的相似度，将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征，根据获得的用户性格特征确定用户的危险性。同时，确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量，根据理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为危险性，进行提前预警，防范危害的发生。

附图说明

[0038] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0039] 图1为本发明实施例基于网络日志的用户行为刻画与预测方法流程图；

[0040] 图2为本发明实施例基于网络日志的用户行为刻画与预测方法的又一流程图；

[0041] 图3为本发明实施例集合X的上近似、下近似、边界域示意图；

[0042] 图4为本发明实施例基于网络日志的用户行为刻画与预测系统结构示意图。

具体实施方式

[0043] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0044] 本发明的目的是提供一种基于网络日志的用户行为刻画与预测方法及系统，能够对用户的性格、行为进行预测，进而，为防范危害的发生提供数据支持。

[0045] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0046] 如图1所示，本发明提供的基于网络日志的用户行为刻画与预测方法包括以下步骤：

[0047] 步骤101：的网络日志；

[0048] 步骤102：根据所述网络日志提取所述用户的行为特征向量，所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学

领域包括史政、文艺、社会、娱乐和美容；

[0049] 步骤103：获取标准性格特征向量，所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0050] 步骤104：计算所述用户的行为特征向量与各所述标准性格特征向量的相似度；

[0051] 步骤105：将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征；

[0052] 作为本发明的一个实施例，在上述实施例的基础上，本发明还包括；

[0053] 步骤106：确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量；

[0054] 步骤107：根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为。

[0055] 其中，步骤104，具体包括：

[0056] 计算所述用户的行为特征向量与各所述标准性格特征向量的余弦相似度；

[0057] 将余弦相似度最小的标准性格特征向量确定为与所述用户行为特征向量相似度最大的标准性格特征向量。

[0058] 步骤105，具体包括：

[0059] 将所述标准性格特征向量划分为积极性格、中级性格和消极性格三种类型；

[0060] 将与用户行为特征向量相似度最大的标准性格特征向量所属类型确定为所述用户的性格类型。

[0061] 步骤107，具体包括：

[0062] 当所述用户的行为特征向量中理科类关键词数量与文科类关键词数量的比值为3:1时，预测所述用户有对他人造成伤害的可能性。

[0063] 作为本发明的又一实施例，如图2所示，本发明提供的基于网络日志的用户行为刻画与预测方法包括以下内容：

[0064] 1.1日志获取

[0065] 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。本发明采用搜狗实验室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。注：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。

[0066] 表1日志样本

[0067]

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
125254	[郭美	3	4	ent.qq.com/a/20060228/000178.h

[0068]

918559	美]			tm
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/2006-01-10/03537937753s.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.j1366.com/2008/photo.asp?id=261
828687165269	[富婆找鸭子]	9	3	blog.sohu.com/members/gfjyktu/1179025.html

[0069] 1.2日志的预处理

[0070] 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。

[0071] 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。

[0072] 1.2.1性格模型的构建

[0073] 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。本发明参照大五人格理论,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、中级与消极三个方面。通过结合个人特点的自身优势、不足进行客观地评价。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，会有一定几率关注体育类、旅游类的内容，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，随性，故有一定几率关注美容类。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3所示。表3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。

[0074] 1.2.2预处理算法(算法1)

[0075] 预处理算法如下(算法1)：

[0076] 输入：用户的原始日志

[0077] 输出：用户的特征向量

[0078] 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

[0079] 步骤2找出日志中的关键词，统计搜索关键词的个数

[0080] 步骤3统计关键词的上位词出现的个数

[0081] 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

[0082] 算法1举例：输出结果例如表2样本中用户125254918559和828687165269搜索的上位次比例构成的向量分别为：

[0083] 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)

[0084] 表2性格模型与性格关键词分类

[0085]

	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

[0086] 核心算法

[0087] 步骤1:相似度匹配

[0088] 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。

[0089] 对未知性格的新用户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，将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。

[0090] 基于余弦相似度特征聚类的性格分析算法如下(算法2)：

[0091] 输入：用户的特征向量(由算法1输出生成)

[0092] 输出：用户的性格特征

[0093] 步骤1构建性格模型向量集合

[0094] 步骤2构建用户测试向量

[0095] 步骤3相似度比较，找出特征向量库中余弦值最大的分量

[0096] 步骤4输出该分量对应的性格特征

[0097] 算法2举例：输出结果例如表2样本中用户125254918559的性格特征为积极性格：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)；用户828687165269的性格特征为消极性格：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)。

[0098] 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-C_n表示上位词，W_1-W_n表示上位词比例权重的平均值，每行代表标准性格特征向量库的分量，记为$\vec{r}_1$。$\vec{r}_2$代表当前用户性格特征向量，将这两个向量用于计算用户的性格趋向。

[0099] $$M=\begin{bmatrix}(C_1,W_1) & (C_1,W_2) & \cdots & (C_1,W_m)\\ (C_2,W_1) & (C_2,W_2) & \cdots & (C_2,W_m)\\ \vdots & \vdots & & \vdots\\ (C_n,W_1) & (C_n,W_2) & \cdots & (C_n,W_m)\end{bmatrix}$$

[0100] 本发明中$\vec{r}_1$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。

[0101] 步骤2关键词的上位词分类，性格分类，构建行为库与行为向量

[0102] 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、中级和消极三个等级。

[0103] 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三个等级。分析三个因素(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，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。

[0104] 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∈IND(K)，如图3所示，子集X关于知识R的上近似和下近似为：

[0105] $$\overline{R}(X)=\left\{x\middle|(\forall x\in U)\wedge([x]_R\cap X\neq\phi)\right\}$$

[0106] $$\underline{R}(X)=\left\{x\middle|(\forall x\in U)\wedge([x]_R\subseteq X)\right\}$$

[0107] 定义4知识简约：知识库中的知识并非同等重要，有些知识冗余。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。

[0108] 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。$V=U_{\alpha\in A}V_\alpha$，$V_\alpha$表示属性α的值域，f为U×A→V，f为信息函数，V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表达来实现。如表3所示，其中U＝{$x_1,x_2,\cdots,x_n$}为论域，A＝{$P_1,P_2,\cdots,P_n$}为全体属性集合。

[0109] 表3知识表达系统

[0110]

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

[0111] 步骤3匹配相等，输出用户行为

[0112] 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。

[0113] 基于粗糙集模糊分析的行为预测算法如下(算法3)：

[0114] 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)

[0115] 输出：用户的行为预测

[0116] 步骤1构建索引表为长度为4的二维数组作为行为分析库

[0117] 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

[0118] 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

[0119] 步骤4划分用户性格等级，存入一维数组对应的空间中

[0120] 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

[0121] 算法3举例：如表1样本中用户125254918559的行为通过采用算法3得到的预测结果为：理科类与文科类兴趣比例约为1:3，(行为具体解释：)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的行为通过采用算法3得到的预测结果为：理科类与文科类兴趣比例约为3:1，(行为具体解释：)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。

[0122] 首先将性格按照积极、中级、消极划分为三个等级；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)，

[0123] [66,100]三个区间分为三个等级。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，…，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，如表4所示。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。x_1--x_{27}表示行为的编号。

[0124] 最后将用户的属性集合与行为分析库比对，找到使得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。

[0125] 表4用户行为分析表

[0126]

行为编号	性格等级	自然科学类词汇等级	社会科学类词汇等级	行为编号	性格等级	自然科学类词汇等级	社会科学类词汇等级
x_1	1	1	1	x_{15}	2	2	3
x_2	1	1	2	x_{16}	2	3	1
x_3	1	1	3	x_{17}	2	3	2
x_4	1	2	1	x_{18}	2	3	3
x_5	1	2	2	x_{19}	3	1	1
x_6	1	2	3	x_{20}	3	1	2
x_7	1	3	1	x_{21}	3	1	3
x_8	1	3	2	x_{22}	3	2	1
x_9	1	3	3	x_{23}	3	2	2
X_{10}	2	1	1	x_{24}	3	2	3

[0127]

X_{11}	2	1	2	x_{25}	3	3	1
X_{12}	2	1	3	x_{26}	3	3	2
X_{13}	2	2	1	x_{27}	3	3	3
X_{14}	2	2	2				

[0128] 本发明还提供了一种基于网络日志的用户行为刻画与预测系统，如图4所示，该系统包括：

[0129] 网络日志获取模块401，用于获取用户的网络日志；

[0130] 用户行为特征向量提取模块402，用于根据所述网络日志提取所述用户的行为特征向量，所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0131] 标准性格特征向量获取模块403，用于获取标准性格特征向量，所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容；

[0132] 相似度计算模块404，用于计算所述用户的行为特征向量与各所述标准性格特征向量的相似度；

[0133] 用户性格刻画模块405，用于将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征；

[0134] 作为本发明的一个实施例，在上述实施例的基础上，本发明还包括：

[0135] 关键词数量确定模块406，用于确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量；

[0136] 用户行为预测模块407，用于根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为。

[0137] 其中，所述相似度计算模块404，具体包括：

[0138] 相似度计算单元，用于计算所述用户的行为特征向量与各所述标准性格特征向量的余弦相似度；

[0139] 性格确定单元，用于将余弦相似度最小的标准性格特征向量确定为与所述用户行为特征向量相似度最大的标准性格特征向量。

[0140] 所述用户性格刻画模块405，具体包括：

[0141] 性格类型划分单元，用于将所述标准性格特征向量划分为积极性格、中级性格和消极性格三种类型；

[0142] 用户性格刻画单元，用于将与用户行为特征向量相似度最大的标准性格特征向量所属类型确定为所述用户的性格类型。

[0143] 所述用户行为预测模块407，具体包括：

[0144] 用户行为预测单元，用于当所述用户的行为特征向量中理科类关键词数量与文科类关键词数量的比值为3:1时，预测所述用户有对他人造成伤害的可能性。

[0145] 本发明提供的基于网络日志的用户行为刻画与预测方法及系统，根据用户网络日志提取用户的行为特征向量，计算用户的行为特征向量与各标准性格特征向量的相似度，将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征，根据获得的用户性格特征确定用户的危险性。同时，确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量，根据理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为危险性，进行提前预警，防范危害的发生。

[0146] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的系统而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0147] 本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本发明的限制。


101
获取用户的网络日志
102
根据所述网络日志提取所述用户的行为特征向量，所述行为特征向量为用户网络日志中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容
103
获取标准性格特征向量，所述标准性格特征向量为标准性格中各领域关键词占关键词总数的比例所构成的向量，所述领域分为自然科学领域和社会科学领域，所述自然科学领域包括军事、科技、体育、旅游和食物，所述社会科学领域包括史政、文艺、社会、娱乐和美容
104
计算所述用户的行为特征向量与各所述标准性格特征向量的相似度
105
将相似度最高的标准性格特征向量所代表的性格特征确定为所述用户的性格特征
106
确定所述用户的行为特征向量中理科类关键词数量和文科类关键词数量
107
根据用户的行为特征向量中理科类关键词数量和文科类关键词数量的比值，预测所述用户的行为

图1

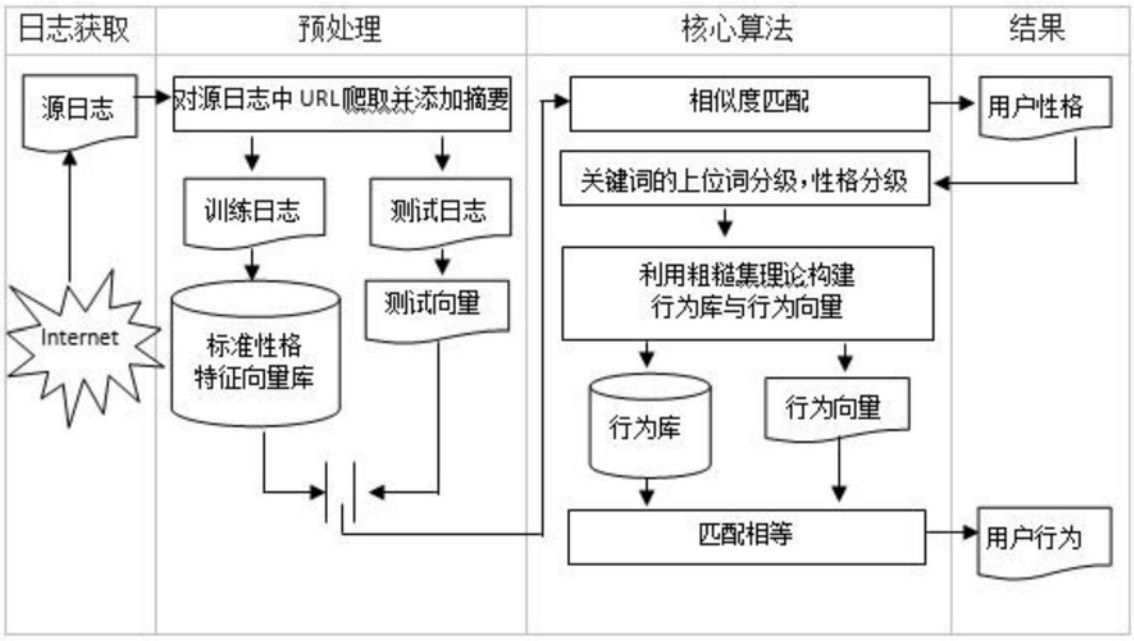

日志获取
预处理
核心算法
结果
源日志
对源日志中 URL爬取并添加摘要
相似度匹配
用户性格
关键词的上位词分级，性格分级
训练日志
测试日志
利用粗糙集理论构建
行为库与行为向量
Internet
标准性格
特征向量库
测试向量
行为库
行为向量
匹配相等
用户行为

图2

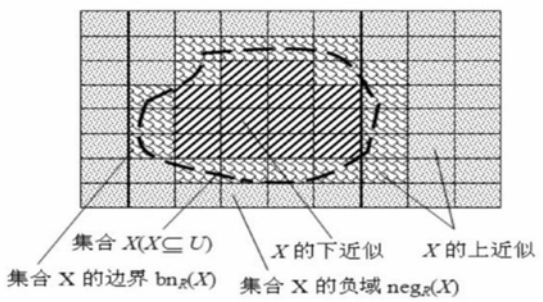

集合 X(X⊆ U)
X 的下近似
X 的上近似
集合 X 的边界 bn_R(X)
集合 X 的负域 neg_R(X)

图3

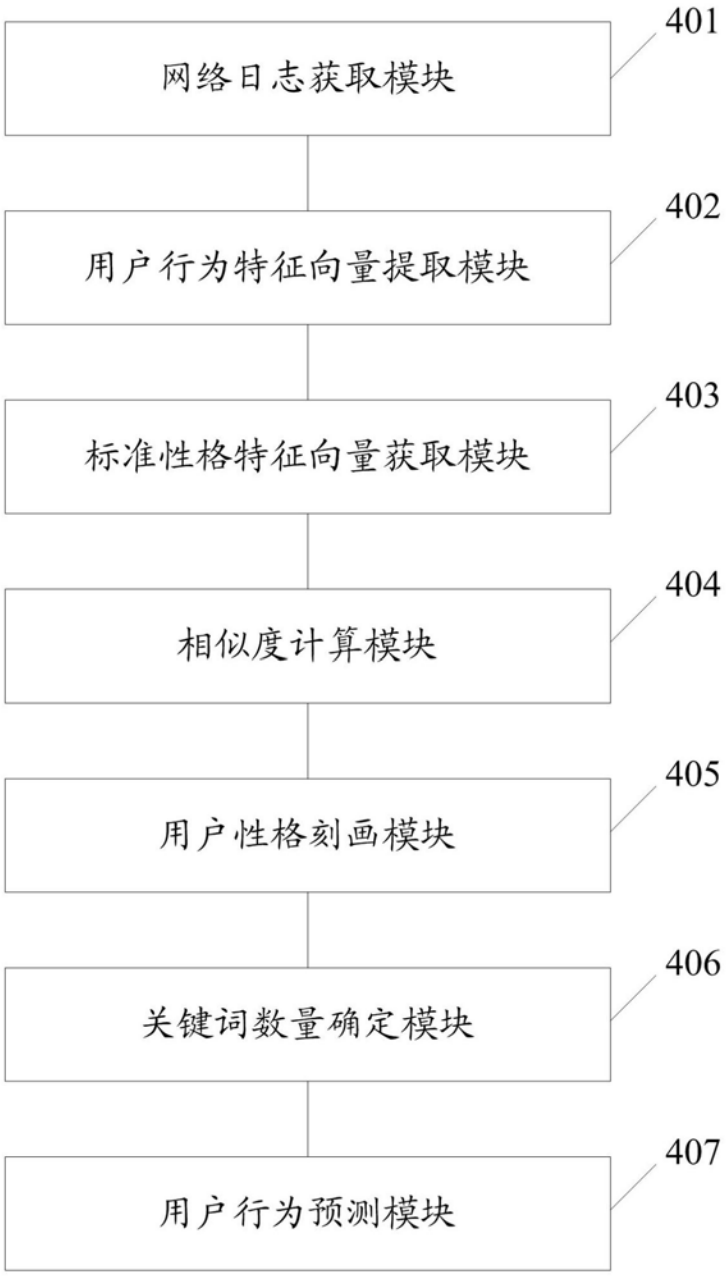

网络日志获取模块
401
用户行为特征向量提取模块
402
标准性格特征向量获取模块
403
相似度计算模块
404
用户性格刻画模块
405
关键词数量确定模块
406
用户行为预测模块
407

图4