

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2019 年 4 月 18 日 (18.04.2019)



(10) 国际公布号
WO 2019/071966 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/089707

(22) 国际申请日: 2018 年 6 月 3 日 (03.06.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710951681.3 2017 年 10 月 13 日 (13.10.2017) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区八卦岭八卦三路平安大厦六楼, Guangdong 518000 (CN)。

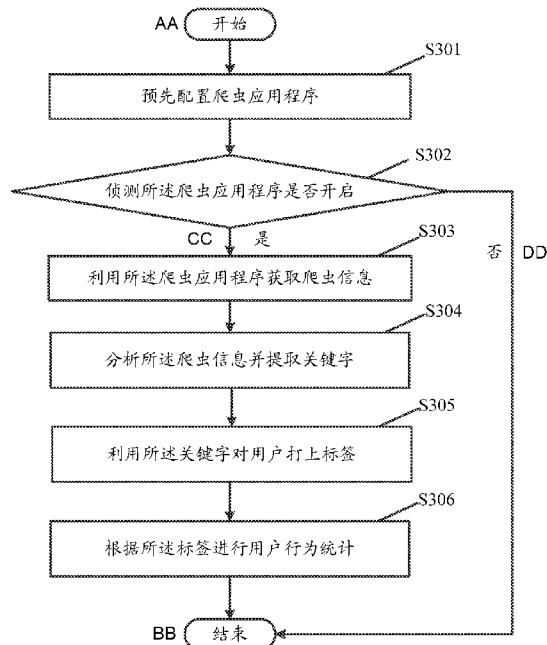
(72) 发明人: 蔡俊(CAI, Jun); 中国广东省深圳市福田区八卦岭八卦三路平安大厦六楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所(普通合伙)(SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: CRAWLER DATA-BASED USER BEHAVIOR ANALYSIS METHOD, APPLICATION SERVER AND READABLE STORAGE MEDIUM

(54) 发明名称: 基于爬虫数据的用户行为分析方法、应用服务器及可读存储介质



S301 Pre-configure a crawler application program
S302 Detect whether the crawler application program is started
S303 Use the crawler application program to acquire crawler information
S304 Analyze the crawler information and extract a keyword
S305 Use the keyword to tag a user
S306 Perform statistics on user behavior according to the tag
AA Start
BB End
CC Yes
DD No

图 3

(57) Abstract: Disclosed in the present application is a crawler data-based user behavior analysis method, the method comprising: pre-configuring a crawler application program; detecting whether the crawler application program is started; after the crawler application program is started, using the crawler application program to acquire crawler information; analyzing the crawler information and extracting a keyword; using the keyword to tag a user; and performing statistics on user behavior according to the tag. The present application further provides an application server and a computer readable storage medium. In the application server and crawler da-

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

ta-based user behavior analysis method provided by the present application, the range of user data is preliminarily delimited; in addition, tagging of the user data is processed in clustering, so as to acquire the user behavior, thereby improving the accuracy of data analysis, facilitating the business decision-making, and improving the marketing efficiency.

(57) 摘要: 本申请公开了一种基于爬虫数据的用户行为分析方法, 所述方法包括: 预先配置爬虫应用程序; 侦测所述爬虫应用程序是否开启; 在开启所述爬虫应用程序后, 利用所述爬虫应用程序获取爬虫信息; 分析所述爬虫信息并提取关键字; 利用所述关键字对用户打上标签; 及根据所述标签进行用户行为统计。本申请还提供一种应用服务器及计算机可读存储介质。本申请提供的应用服务器及基于爬虫数据的用户行为分析方法, 通过初步的划定用户数据的范围, 同时对用户数据的标签化在聚类处理, 获取用户行为, 提高了数据分析的准确性, 有助于商业决策, 提高营销效率。

基于爬虫数据的用户行为分析方法、应用服务器及可读存储介质

本申请基于巴黎公约申明享有 2017 年 10 月 13 日递交的申请号为 CN 201710951681.3、名称为“基于爬虫数据的用户行为分析方法、应用服务器及计算机可读存储介质”中国专利申请的优先权，该中国专利申请的整内容以参考的方式结合在本申请中。

技术领域

本申请涉及通信技术领域，尤其涉及基于爬虫数据的用户行为分析方法、应用服务器及可读存储介质。

背景技术

大数据使得企业能够通过互联网便利地获取用户更为广泛的反馈信息，为进一步精准、快速地分析用户行为习惯、消费习惯等重要商业信息，提供了足够的数据库。伴随着对人的了解逐步深入，一个概念悄然而生：用户画像(User Profile)，完美地抽象出一个用户的信息全貌，可以看作企业应用大数据的根基。

然目前的大数据获取，数据量大而杂，无法做到数据的初步范围划定，进而影响后续用户行为分析的准确性。

发明内容

本申请提供一种基于爬虫数据的用户行为分析方法及应用服务器，通过初步的划定用户数据的范围，同时对用户数据的标签化在聚类处理，获取用户行为，提高了数据分析的准确性，有助于商业决策，提高营销效率。

本申请第一方面提供一种基于爬虫数据的用户行为分析方法，应用于应用服务器，所述方法包括：

预先配置爬虫应用程序；

侦测所述爬虫应用程序是否开启；

在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息；

分析所述爬虫信息并提取关键字；

利用所述关键字对用户打上标签；及

根据所述标签进行用户行为统计。

本申请第二方面提供一种应用服务器，所述应用服务器包括存储器、处理器，所述存储器上存储有可在所述处理器上运行的基于爬虫数据的用户行为分析程序，所述基于爬虫数据的用户行为分析程序被所述处理器执行时实现如下步骤：

预先配置爬虫应用程序；

侦测所述爬虫应用程序是否开启；
在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息；

分析所述爬虫信息并提取关键字；
利用所述关键字对用户打上标签；及
根据所述标签进行用户行为统计。

本申请第三方面提供一种计算机可读存储介质，所述计算机可读存储介质存储有基于爬虫数据的用户行为分析程序，所述基于爬虫数据的用户行为分析程序可被至少一个处理器执行，以使所述至少一个处理器执行以下步骤：

预先配置爬虫应用程序；
侦测所述爬虫应用程序是否开启；
在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息；

分析所述爬虫信息并提取关键字；
利用所述关键字对用户打上标签；及
根据所述标签进行用户行为统计。

相较于现有技术，本申请所提出的应用服务器、基于爬虫数据的用户行为分析方法及计算机可读存储介质，首先，预先配置爬虫应用程序；其次，侦测所述爬虫应用程序是否开启；然后，在开启所述爬虫应用程序后，再次，利用所述爬虫应用程序获取爬虫信息；接着，分析所述爬虫信息并提取关键字；进一步地，利用所述关键字对用户打上标签；最后，根据所述标签进行用户行为统计。这样，可以避免现有技术中大数据获取数据量大而杂，无法做到数据的初步范围划定，进而影响后续用户行为分析的准确性的弊端。通过初步的划定用户数据的范围，同时对用户数据的标签化在聚类处理，获取用户行为，提高了数据分析的准确性，有助于商业决策，提高营销效率。

附图说明

图 1 是应用服务器一可选的硬件架构的示意图；

图 2 是本申请基于爬虫数据的用户行为分析程序第一实施例的程序模块图；

图 3 为本申请基于爬虫数据的用户行为分析方法第一实施例的流程图；

图 4 为本申请基于爬虫数据的用户行为分析方法第二实施例的流程图。

附图标记:

应用服务器	1
存储器	11
处理器	12
网络接口	13
基于爬虫数据的用户行为分析程序	200
预先配置模块	201
侦测模块	202
获取模块	203
分析提取模块	204
标签模块	205
统计模块	206

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

以下结合附图对本申请的原理和特征进行描述，所举实例只用于解释本申请，并非用于限定本申请的范围。

参阅图1所示，是应用服务器1一可选的硬件架构的示意图。

所述应用服务器1可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备，该应用服务器1可以是独立的服务器，也可以是多个服务器所组成的服务器集群。

本实施例中，所述应用服务器1可包括，但不限于，可通过系统总线相互通信连接存储器11、处理器12、网络接口13。

所述应用服务器1通过网络接口13连接网络，获取资讯。所述网络可以是企业内部网（Intranet）、互联网（Internet）、全球移动通讯系统（Global System of Mobile communication, GSM）、宽带码分多址（Wideband Code Division Multiple Access, WCDMA）、4G网络、5G网络、蓝牙（Bluetooth）、Wi-Fi、通话网络等无线或有线网络。

需要指出的是，图1仅示出了具有组件11-13的应用服务器1，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多

或者更少的组件。

其中，所述存储器11至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD或DX存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器11可以是所述应用服务器1的内部存储单元，例如该应用服务器1的硬盘或内存。在另一些实施例中，所述存储器11也可以是所述应用服务器1的外部存储设备，例如该应用服务器1配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，所述存储器11还可以既包括所述应用服务器1的内部存储单元也包括其外部存储设备。本实施例中，所述存储器11通常用于存储安装于所述应用服务器1的操作系统和各类应用软件，例如所述基于爬虫数据的用户行为分析程序200的程序代码等。此外，所述存储器11还可以用于暂时地存储已经输出或者将要输出的各类数据。

所述处理器12在一些实施例中可以是中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器12通常用于控制所述应用服务器1的总体操作，例如执行数据交互或者通信相关的控制和处理等。本实施例中，所述处理器12用于运行所述存储器11中存储的程序代码或者处理数据，例如运行所述的基于爬虫数据的用户行为分析程序200等。

所述网络接口13可包括无线网络接口或有线网络接口，该网络接口13通常用于在所述应用服务器1与其他电子设备之间建立通信连接。

本实施例中，所述应用服务器1内安装并运行有基于爬虫数据的用户行为分析程序200，当所述基于爬虫数据的用户行为分析程序200运行时，所述应用服务器1首先，预先配置爬虫应用程序；其次，侦测所述爬虫应用程序是否开启；然后，在开启所述爬虫应用程序后，再次，利用所述爬虫应用程序获取爬虫信息；接着，分析所述爬虫信息并提取关键字；进一步地，利用所述关键字对用户打上标签；最后，根据所述标签进行用户行为统计。这样，可以避免现有技术中大数据获取数据量大而杂，无法做到数据的初步范围划定，进而影响后续用户行为分析的准确性的弊端。通过初步的划定用户数据的范围，同时对用户数据的标签化在聚类处理，获取用户行为，提高了数据分析的准确性，有助于商业决策，提高营销效率。

至此，已经详细介绍了本申请各个实施例的应用环境和相关设备

的硬件结构和功能。下面，将基于上述应用环境和相关设备，提出本申请的各个实施例。

首先，本申请提出一种基于爬虫数据的用户行为分析程序200。

参阅图2所示，是本申请基于爬虫数据的用户行为分析程序200第一实施例的程序模块图。

本实施例中，所述的基于爬虫数据的用户行为分析程序200包括一系列的存储于存储器11上的计算机程序指令，当该计算机程序指令被处理器12执行时，可以实现本申请各实施例的基于爬虫数据的用户行为的分析操作。在一些实施例中，基于该计算机程序指令各部分所实现的特定的操作，所述基于爬虫数据的用户行为分析程序200可以被划分为一个或多个模块。例如，在图2中，所述的基于爬虫数据的用户行为分析程序200可以被分割成预先配置模块201、侦测模块202、获取模块203、分析提取模块204、标签模块205及统计模块206。其中：

所述预先配置模块201，用于预先配置爬虫应用程序。具体地，所述预先配置模块201通过预设所述爬虫应用程序设置窗口，配置所述爬虫应用程序的爬取规则。

在本实施例中，所述爬取规则包括爬取的网站、内容、爬取的文件格式（文字、图片、音频、视频）、爬取的频次等等，即，所述预先配置模块201可以通过配置爬虫应用程序，获取指定范围的大数据信息，进而提高后续用户行为分析的准确性。

所述侦测模块202，用于侦测所述爬虫应用程序是否开启。

所述获取模块203，用于在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息。具体地，所述爬虫信息包括用户爬取的网站、用户爬取的文件格式、用户爬取的内容、用户的访问IP、用户爬取时间、用户爬取次数中一种或多种的组合。

在本实施例中，爬虫应用程序可以收集到很多爬虫信息。比如：

第一，人员注册的信息：昵称、电话号码、公司名称、注册地址、注册的IP地址

第二，爬取规则信息：爬取的网站、内容、爬取的文件格式（文字、图片、音频、视频）、爬取的频次；

第三，爬取数据提取信息：用户提取数据的时间、提取量、提取范围。

所述分析提取模块204，用于分析所述爬虫信息并提取关键字。本实施例中，所述关键字包括网站地址、网站类型、爬取内容性质、爬取文件的格式、IP、爬取时间段、爬取频率、爬取数量中的一种或多种的组合。

在本实施例中，如果爬取信息中有网站信息，那么可以提取其中

的网址或者相应网站的类型作为关键字；如果爬取信息有用户的访问IP，那么IP可以作为关键字；如果爬取信息是爬取时间，那么以该时间左右一定范围的时间段为关键字；如果爬取信息为用户爬取的内容，那么以该内容的性质为关键字，比如视频、音乐、图片等等；如果爬取信息是爬取的文件格式，那么以相应格式为关键字；如果爬取信息是爬取次数，则可以爬取频率或爬取数量为关键字。

所述标签模块205，用于利用所述关键字对用户打上标签。例如，提取网址或者相应网站的类型作为关键字时，用户标签可以根据用户访问的网址分为投资客，储户，或者投保人，当关键字为IP时，用户标签可以根据IP分为学校，医院，银行等，当关键字为视频、音乐、图片时，用户标签可以根据视频、音乐、图片的访问频率分为音乐人，编剧，粉丝等。

所述统计模块206，用于根据所述标签模块205对用户打上的标签进行用户行为统计。

具体地，所述标签模块205主要通过以下方式对用户行为进行统计：所述标签模块205根据预设指标对打上所述标签的用户进行聚类；并根据聚类后的用户的数据统计用户行为。

具体地，所述预设指标包括：近度、频度、额度。

本实施例中，统计用户行为，也可以称之为用户画像(User Profile)，用户画像的底层是机器学习，那么无论是要做客户分群还是精准营销，都先要将用户数据进行规整处理，转化为相同维度的特征向量，诸多华丽的算法才可以有用武之地，所谓的规整处理比如聚类，回归，关联，各种分类器等等。

在本实施例中，通过上述方案首先对用户进行了标签化处理，定义多种维度。然后基于同一维度去统计用户数，进而分析用户行为。在本实施方式中，所述预设指标中的近度(Recency)指的是用户第一次或最后一次的爬取时间；频度(Frequency)指的是用户的爬取频率；额度(Monetary)指的是用户的爬取数量。当然在其他实施例中，还可以选择其他指标作为聚类指标，比如爬取信息涉及评分内容，那么标签可以是评分，聚类指标可以选择爬取内容评分，或者选择用户数量作为聚类指标等等。

通过上述对标签化的数据进行聚类，可以获取用户关注度高的一些网站或者内容、也可以获取使用爬虫程序的用户分布情况等等。根据上述聚类得出的用户行为，可以辅助商业决策，提高营销效率。

通过上述程序模块201-206，本申请所提出的基于爬虫数据的用户行为分析程序200，首先，预先配置爬虫应用程序；其次，侦测所述爬虫应用程序是否开启；然后，在开启所述爬虫应用程序后，再次，

利用所述爬虫应用程序获取爬虫信息；接着，分析所述爬虫信息并提取关键字；进一步地，利用所述关键字对用户打上标签；最后，根据所述标签进行用户行为统计。这样，可以避免现有技术中大数据获取数据量大而杂，无法做到数据的初步范围划定，进而影响后续用户行为分析的准确性的弊端。通过初步的划定用户数据的范围，同时对用户数据的标签化在聚类处理，获取用户行为，提高了数据分析的准确性，有助于商业决策，提高营销效率。

此外，本申请还提出一种基于爬虫数据的用户行为分析方法。

参阅图3所示，是本申请基于爬虫数据的用户行为分析方法第一实施例的流程图。在本实施例中，根据不同的需求，图3所示的流程图中的步骤的执行顺序可以改变，某些步骤可以省略。

步骤 S301，预先配置爬虫应用程序。具体地，所述应用服务器 1 通过预设所述爬虫应用程序设置窗口，配置所述爬虫应用程序的爬取规则。

在本实施例中，所述爬取规则包括爬取的网站、内容、爬取的文件格式（文字、图片、音频、视频）、爬取的频次等等，即，所述应用服务器 1 可以通过配置爬虫应用程序，获取指定范围的大数据信息，进而提高后续用户行为分析的准确性。

步骤 S302，侦测所述爬虫应用程序是否开启。当所述爬虫应用程序开启时，执行步骤 S303，否则，结束流程。

步骤 S303，利用所述爬虫应用程序获取爬虫信息。

具体地，所述爬虫信息包括用户爬取的网站、用户爬取的文件格式、用户爬取的内容、用户的访问 IP、用户爬取时间、用户爬取次数中一种或多种的组合。

在本实施例中，爬虫应用程序可以收集到很多爬虫信息。比如：

第一，人员注册的信息：昵称、电话号码、公司名称、注册地址、注册的 IP 地址。

第二，爬取规则信息：爬取的网站、内容、爬取的文件格式（文字、图片、音频、视频）、爬取的频次。

第三，爬取数据提取信息：用户提取数据的时间、提取量、提取范围。

步骤 S304，分析所述爬虫信息并提取关键字。

本实施例中，所述关键字包括网站地址、网站类型、爬取内容性质、爬取文件的格式、IP、爬取时间段、爬取频率、爬取数量中的一种或多种的组合。

在本实施例中，如果爬取信息中有网站信息，那么可以提取其中的网址或者相应网站的类型作为关键字；如果爬取信息有用户的访问

IP, 那么 IP 可以作为关键字; 如果爬取信息是爬取时间, 那么以该时间左右一定范围的时间段为关键字; 如果爬取信息为用户爬取的内容, 那么以该内容的性质为关键字, 比如视频、音乐、图片等等; 如果爬取信息是爬取的文件格式, 那么以相应格式为关键字; 如果爬取信息是爬取次数, 则可以爬取频率或爬取数量为关键字。

步骤 S305, 利用所述关键字对用户打上标签。例如, 提取网址或者相应网站的类型作为关键字时, 用户标签可以根据用户访问的网址分为投资客, 储户, 或者投保人, 当关键字为 IP 时, 用户标签可以根据 IP 分为学校, 医院, 银行等, 当关键字为视频、音乐、图片时, 用户标签可以根据视频、音乐、图片的访问频率分为音乐人, 编剧, 粉丝等。

步骤 S306, 根据所述标签进行用户行为统计。具体地, 所述根据所述标签进行用户行为统计的具体步骤将在本申请基于爬虫数据的用户行为分析方法第二实施例(图 4)进行详述。

通过上述步骤 S301-306, 本申请所提出的基于爬虫数据的用户行为分析方法, 所述应用服务器 1 首先, 预先配置爬虫应用程序; 其次, 侦测所述爬虫应用程序是否开启; 然后, 在开启所述爬虫应用程序后, 再次, 利用所述爬虫应用程序获取爬虫信息; 接着, 分析所述爬虫信息并提取关键字; 进一步地, 利用所述关键字对用户打上标签; 最后, 根据所述标签进行用户行为统计。这样, 可以避免现有技术中大数据获取数据量大而杂, 无法做到数据的初步范围划定, 进而影响后续用户行为分析的准确性的弊端。通过初步的划定用户数据的范围, 同时对用户数据的标签化在聚类处理, 获取用户行为, 提高了数据分析的准确性, 有助于商业决策, 提高营销效率。

参阅图4所示, 是本申请基于爬虫数据的用户行为分析方法第二实施例的流程图。在本实施例中, 根据不同的需求, 图4所示的流程图中的步骤的执行顺序可以改变, 某些步骤可以省略。

在本实施例中, 所述根据标签进行用户行为统计的步骤, 具体包括:

步骤 S401, 根据预设指标对打上标签的用户进行聚类。

本实施例中, 所述应用服务器 1 利用所述关键字对用户打上标签, 所述关键字包括网站地址、网站类型、爬取内容性质、爬取文件的格式、IP、爬取时间段、爬取频率、爬取数量中的一种或多种的组合。

在本实施例中, 如果爬取信息中有网站信息, 那么可以提取其中的网址或者相应网站的类型作为关键字; 如果爬取信息有用户的访问

IP, 那么 IP 可以作为关键字; 如果爬取信息是爬取时间, 那么以该时间左右一定范围的时间段为关键字; 如果爬取信息为用户爬取的内容, 那么以该内容的性质为关键字, 比如视频、音乐、图片等等; 如果爬取信息是爬取的文件格式, 那么以相应格式为关键字; 如果爬取信息是爬取次数, 则可以爬取频率或爬取数量为关键字。

其中, 例如, 提取网址或者相应网站的类型作为关键字时, 用户标签可以根据用户访问的网址分为投资客, 储户, 或者投保人, 当关键字为 IP 时, 用户标签可以根据 IP 分为学校, 医院, 银行等, 当关键字为视频、音乐、图片时, 用户标签可以根据视频、音乐、图片的访问频率分为音乐人, 编剧, 粉丝等。具体地, 所述预设指标包括: 近度、频度、额度。

步骤 S402, 根据聚类后的用户的数据统计用户行为。本实施例中, 统计用户行为, 也可以称之为用户画像(User Profile), 用户画像的底层是机器学习, 那么无论是要做客户分群还是精准营销, 都先要将用户数据进行规整处理, 转化为相同维度的特征向量, 诸多华丽的算法才可以有用武之地, 所谓的规整处理比如聚类, 回归, 关联, 各种分类器等等。

在本实施例中, 通过上述方案, 首先对用户进行了标签化处理, 定义多种维度。然后基于同一维度去统计用户数, 进而分析用户行为。在本实施方式中, 所述预设指标中的近度(Recency)指的是用户第一次或最后一次的爬取时间; 频度(Frequency)指的是用户的爬取频率; 额度(Monetary)指的是用户的爬取数量。当然在其他实施例中, 还可以选择其他指标作为聚类指标, 比如爬取信息涉及评分内容, 那么标签可以是评分, 聚类指标可以选择爬取内容评分, 或者选择用户数量作为聚类指标等等。

通过上述对标签化的数据进行聚类, 可以获取用户关注度高的一些网站或者内容、也可以获取使用爬虫程序的用户分布情况等等。根据上述聚类得出的用户行为, 可以辅助商业决策, 提高营销效率。

通过上述步骤 S401-402, 本申请所提出的基于爬虫数据的用户行为分析方法, 通过对标签化的数据进行聚类, 可以获取用户关注度高的一些网站或者内容、也可以获取使用爬虫程序的用户分布情况等。而且, 根据聚类得出的用户行为, 可以辅助商业决策, 提高营销效率。。

上述本申请实施例序号仅仅为了描述, 不代表实施例的优劣。

通过以上的实施方式的描述, 本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现, 当然也可以通过硬件, 但很多情况下前者是更佳的实施方式。基于这

样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的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（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

以上所述仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是在本申请的申请构思下，利用本申请说明书及附图内容所作的等效结构变换，或直接/间接运用在其他相关的技术领域均包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种基于爬虫数据的用户行为分析方法，应用于应用服务器，其特征在于，所述方法包括：

预先配置爬虫应用程序；

侦测所述爬虫应用程序是否开启；

在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息；

分析所述爬虫信息并提取关键字；

利用所述关键字对用户打上标签；及

根据所述标签进行用户行为统计。

2、如权利要求 1 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：

根据预设指标对打上所述标签的用户进行聚类；及

根据聚类后的用户的数据统计用户行为。

3、如权利要求 2 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述预设指标包括：近度、频度、额度。

4、如权利要求 1 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述预先配置爬虫程序的步骤，具体包括：

通过预设所述爬虫应用程序设置窗口，配置所述爬虫应用程序的爬取规则。

5、如权利要求 4 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：

根据预设指标对打上所述标签的用户进行聚类；及

根据聚类后的用户的数据统计用户行为。

6、如权利要求 1 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述爬虫信息包括用户爬取的网站、用户爬取的文件格式、用户爬取的内容、用户的访问 IP、用户爬取时间、用户爬取次数中一种或多种的组合。

7、如权利要求 6 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：

根据预设指标对打上所述标签的用户进行聚类；及

根据聚类后的用户的数据统计用户行为。

8、如权利要求 1 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述关键字包括网站地址、网站类型、爬取内容性质、爬取文件的格式、IP、爬取时间段、爬取频率、爬取数量中的一种或多种的组合。

9、如权利要求 8 所述的基于爬虫数据的用户行为分析方法，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：
根据预设指标对打上所述标签的用户进行聚类；及
根据聚类后的用户的数据统计用户行为。

10、一种应用服务器，其特征在于，所述应用服务器包括存储器、处理器，所述存储器上存储有可在所述处理器上运行的基于爬虫数据的用户行为分析程序，所述基于爬虫数据的用户行为分析程序被所述处理器执行时实现如下步骤：

预先配置爬虫应用程序；
侦测所述爬虫应用程序是否开启；
在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息；
分析所述爬虫信息并提取关键字；
利用所述关键字对用户打上标签；及
根据所述标签进行用户行为统计。

11、如权利要求 10 所述的应用服务器，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：
根据预设指标对打上所述标签的用户进行聚类；及
根据聚类后的用户的数据统计用户行为。

12、如权利要求 11 所述的应用服务器，其特征在于，所述预设指标包括：近度、频度、额度。

13、如权利要求 10 所述的应用服务器，其特征在于，所述预先配置爬虫程序的步骤，具体包括：

通过预设所述爬虫应用程序设置窗口，配置所述爬虫应用程序的爬取规则。

14、如权利要求 13 所述的应用服务器，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：
根据预设指标对打上所述标签的用户进行聚类；及

根据聚类后的用户的数据统计用户行为。

15、如权利要求 10 所述的应用服务器，其特征在于，所述爬虫信息包括用户爬取的网站、用户爬取的文件格式、用户爬取的内容、用户的访问 IP、用户爬取时间、用户爬取次数中一种或多种的组合。

16、如权利要求 15 所述的应用服务器，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：

根据预设指标对打上所述标签的用户进行聚类；及
根据聚类后的用户的数据统计用户行为。

17、如权利要求 10 所述的应用服务器，其特征在于，所述关键字包括网站地址、网站类型、爬取内容性质、爬取文件的格式、IP、爬取时间段、爬取频率、爬取数量中的一种或多种的组合。

18、如权利要求 17 所述的应用服务器，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：

根据预设指标对打上所述标签的用户进行聚类；及
根据聚类后的用户的数据统计用户行为。

19、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有基于爬虫数据的用户行为分析程序，所述基于爬虫数据的用户行为分析程序可被至少一个处理器执行，以使所述至少一个处理器执行以下步骤：

预先配置爬虫应用程序；
侦测所述爬虫应用程序是否开启；
在开启所述爬虫应用程序后，利用所述爬虫应用程序获取爬虫信息；
分析所述爬虫信息并提取关键字；
利用所述关键字对用户打上标签；及
根据所述标签进行用户行为统计。

20、如权利要求 19 所述的计算机可读存储介质，其特征在于，所述根据所述标签进行用户行为统计的步骤，具体包括：

根据预设指标对打上所述标签的用户进行聚类；及
根据聚类后的用户的数据统计用户行为。

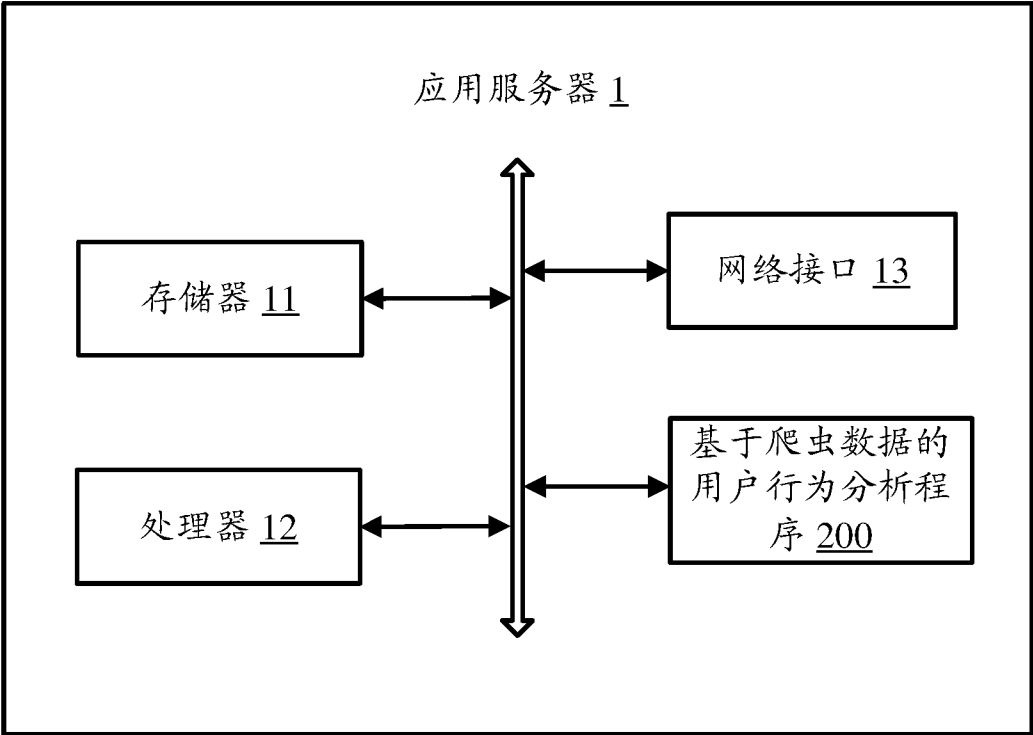


图 1

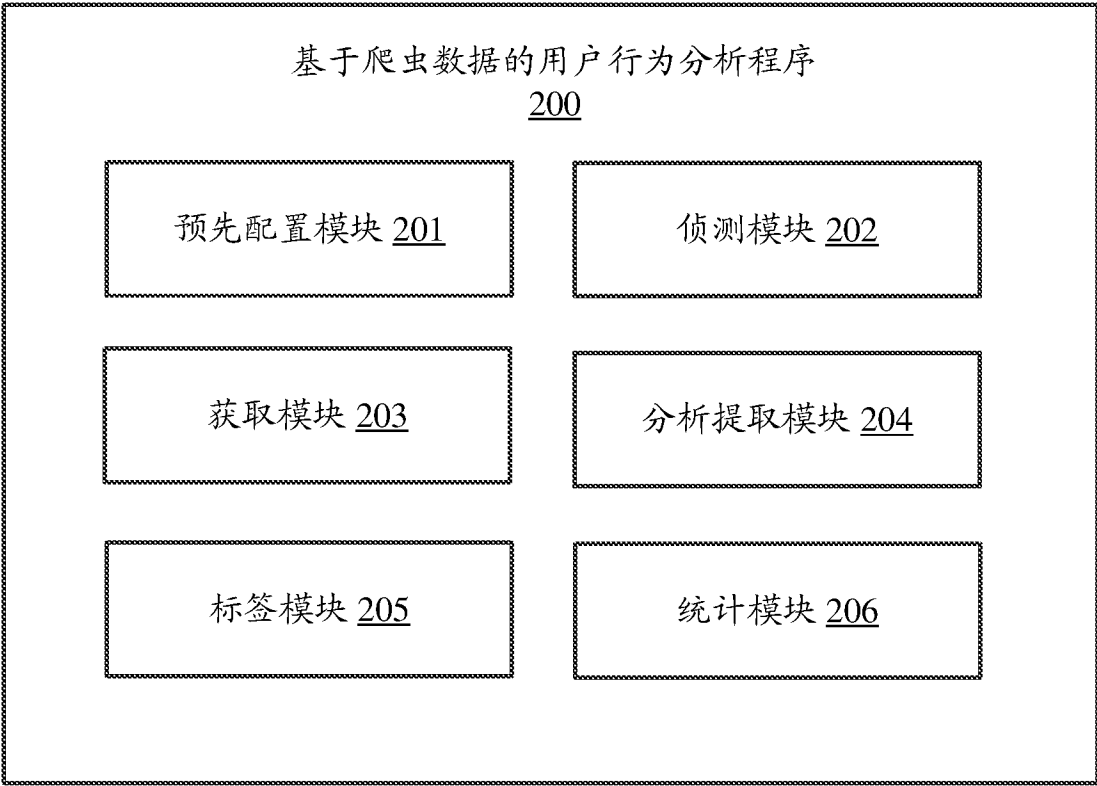


图 2

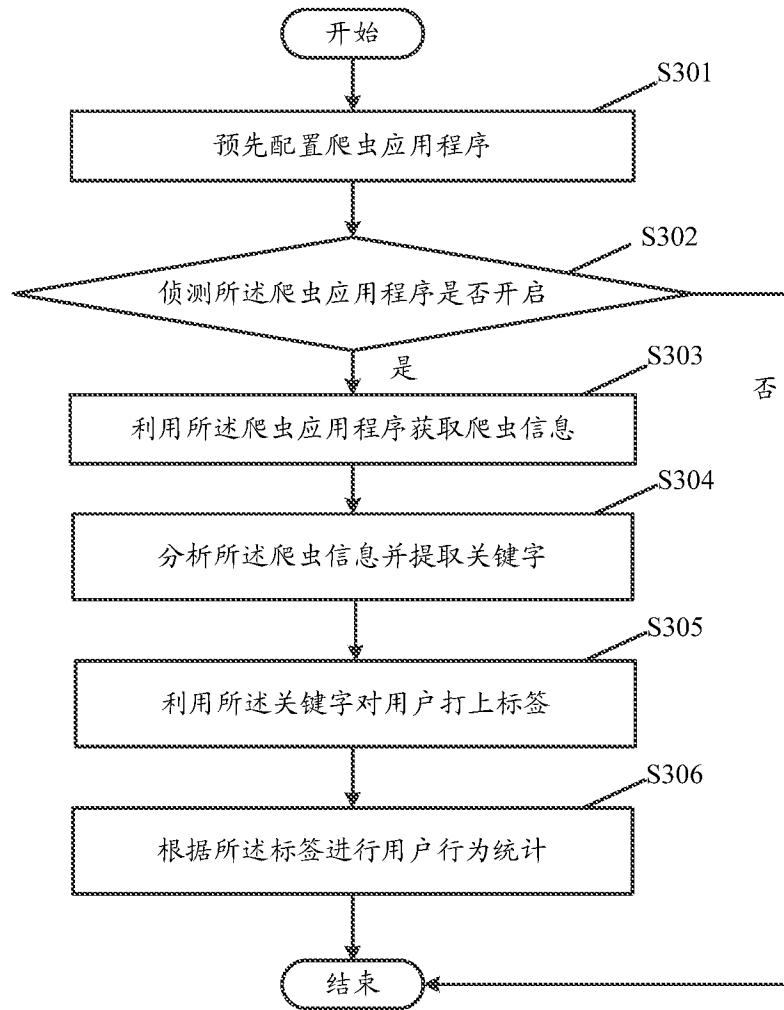


图 3

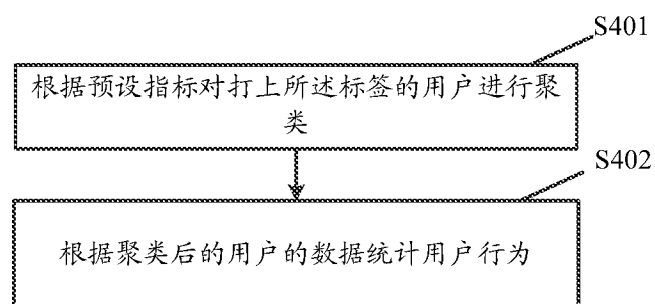


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/089707

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT; CNABS; CNKI; TWTXT; TWABS; USTXT; VEN; EPTXT; WOTXT: 爬虫, 提取, 用户行为, 分析, 关键字, 聚类, 统计, 用户画像, crawler, get, obtain, user behavior, analysis, keyword, clustering, statistic, user profile

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 107870986 A (PING AN TECH SHENZHEN CO., LTD.) 03 April 2018 (2018-04-03) entire document	1-20
X	CN 106844588 A (SHANGHAI FEIXUN COMMUNICATION CO., LTD.) 13 June 2017 (2017-06-13) description, paragraphs 44-64 and 81-106	1-20
A	CN 106940705 A (SHANGHAI ZHANGMEN TECHNOLOGY CO., LTD.) 11 July 2017 (2017-07-11) entire document	1-20
A	US 2011258198 A1 (MICROSOFT CORP.) 20 October 2011 (2011-10-20) entire document	1-20

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

08 August 2018

Date of mailing of the international search report

22 August 2018

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/089707

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107870986	A	03 April 2018	None			
CN	106844588	A	13 June 2017	None			
CN	106940705	A	11 July 2017	None			
US	2011258198	A1	20 October 2011	US	8244701	B2	14 August 2012

国际检索报告

国际申请号

PCT/CN2018/089707

A. 主题的分类

G06F 17/30(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNTXT;CNABS;CNKI;TWTXT;TWABS;USTXT;VEN;EPTXT;WOTXT: 爬虫, 提取, 用户行为, 分析, 关键字, 聚类, 统计, 用户画像, crawler, get, obtain, user behavior, analysis, keyword, clustering, statistic, user profile

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 107870986 A (平安科技深圳有限公司) 2018年 4月 3日 (2018 - 04 - 03) 全文	1-20
X	CN 106844588 A (上海斐讯数据通信技术有限公司) 2017年 6月 13日 (2017 - 06 - 13) 说明书第44-64、81-106段	1-20
A	CN 106940705 A (上海掌门科技有限公司) 2017年 7月 11日 (2017 - 07 - 11) 全文	1-20
A	US 2011258198 A1 (MICROSOFT CORP) 2011年 10月 20日 (2011 - 10 - 20) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2018年 8月 8日

国际检索报告邮寄日期

2018年 8月 22日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

杨牛

电话号码 86-(20)-28950388

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/089707

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	107870986	A	2018年 4月 3日	无			
CN	106844588	A	2017年 6月 13日	无			
CN	106940705	A	2017年 7月 11日	无			
US	2011258198	A1	2011年 10月 20日	US	8244701	B2	2012年 8月 14日

表 PCT/ISA/210 (同族专利附件) (2015年1月)