基于 Hadoop 的固网宽带终端识别技术研究和实现

范孟可,王 攀
(南京邮电大学 物联网学院,江苏 南京 210003)

摘 要:随着大数据时代的来临,大数据在各个行业应用越来越广泛。大数据在运营商行业的应用也很普遍,但同时也遇到了很多技术问题,其中家庭画像的塑造是运营商大数据的一个核心问题。如何提取和识别固网宽带下的终端类型是一个有待解决的问题。不像移动网,固网宽带由于没有信令通道,所以不携带任何准确的终端信息,因而对固网下的终端类型识别比较困难。传统方法都是采用解析和匹配 HTTP GET 报文中的 UA 字段进行识别。但这种方法由于 UA 的非标准化,以及终端数量和种类众多的缘故而导致终端类型的识别准确率低下。文中采用 Hadoop 框架,利用 Hive 中 UDF 的方法,结合分布式爬虫获取终端库,可以更加快速准确地识别出用户上网终端信息。实验结果表明,终端识别准确率可以达到 92% 以上,相比传统方法有了大幅提升。

关键词:终端识别;Hadoop;User Defined Function (UDF);分布式爬虫;固网宽带;大数据运营
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2017)11-0171-05
doi:10.3969/j.issn.1673-629X.2017.11.037

Research and Implementation of Terminal Identification Technology of Fixed-line Broadband Based on Hadoop

FAN Meng-ke, WANG Pan
(School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: With the coming of the era of big data, big data is more and more widely applied in various industries, which is also done in operators industry, but many technical problems are found simultaneously, of which family portraits of shaping is a core for operators of large data. How to extract and identify the terminal type of fixed-line broadband is a problem needed to be solved. Unlike mobile network, fixed-line broadband don't take any accurate terminal information due to lack of signaling channel, so it is hard to conduct terminal type identification in fixed-line. The traditional method adopts UA fields of HTTP GET message parsing and matching for identification, but it is low in identification accuracy because of UA non-standardized and the large amounts of terminal number and varieties. Based on the Hadoop framework, the UDF of Hive is used, and combined with the distributed crawler for obtainment of terminal library, the user terminal information online is identified more quickly and accurately. According to the experiment, the accuracy of terminal identification can reach above 92%, a substantial increase compared with the traditional method.

Key words: terminal identification; Hadoop; User Defined Function (UDF); distributed crawler; fixed-line broadband; big data operations

0 引 言

当今,随着计算机技术的发展,大数据被应用到生活中的各行各业。大数据已经是行业的趋势,当今时代也是“大数据”[1]时代。

传统的电信运营商还只是把数据简单地保存起来,没有发挥数据的价值。而随着信息技术的快速发展,运营商开始意识到数据对企业日常的管理和营销的支撑具有重大意义。因此,运营商建立了一些企业信息化系统为公司的经营决策[2]和资源配置提供帮助。这些系统包括企业的管理系统、运营支撑系统、市场营销支撑系统等。相对于互联网商,电信运营商的最大优势是它拥有用户的全流量数据,是用户数据[3]的第一接口。电信运营商拥有用户身份信息、网络状态、终端、业务识别、位置、社交关系、消费信用等信息,而且这些信息具有很大的商业价值。另外手机、PAD等移动终端是移动互联网时代必不可少的物品。电信运营商需要对用户使用的移动终端做深入研究,以此来提升用户体验,提高自己的用户量。因此进行终端

收稿日期:2016-11-27 **修回日期**:2017-03-29 **网络出版时间**:2017-08-01
基金项目:2015 江苏省产学研前瞻性联合研究项目(BY2015011-02)
作者简介:范孟可(1990-),男,硕士研究生,研究方向为大数据分析技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170801.1552.048.html
万方数据

型号识别以及终端功能配置识别具有很大的意义。

(1)不仅能够减轻无线网络中数据业务负担,而且能提高服务质量。固网宽带自建家庭 WiFi 是一个很好的解决方案。统计有多少移动终端支持 WiFi 功能对于运营商来说是需要解决的问题。

(2)对于传统市场营销,往往通过区域划分来反映相应的用户商业价值。但随着互联网的发展,区域划分并不能准确代表用户的商业价值[4]。现在提出通过用户终端、用户上网行为和用户订阅的业务来分析用户潜在的商业价值。

因此,为了满足以上类似的要求,研究互联网中用户的移动终端类型成为重要课题。然而,在这个大数据时代,人们的生活中出现了很多新型移动互联网业务,比如微信等即时通信、支付宝等手机支付、百度地图等导航服务和直播平台等。而且当今移动互联网用户量巨大,在网络方面 4G 的普及,电信百兆宽带的提倡,大大提高了网络速率。在智能终端[5]方面,硬件处理性能和软件功能都有大幅提高。这些因素导致网络流量数据[6]空前巨大,这些海量的用户数据,对于数据处理能力也提出了更高的要求,传统的计算方式已经不能满足当前数据量的要求。然而云计算模型 Hadoop 框架使对海量的数据处理[7]和挖掘成为可能。Hadoop 的出现为人们提供了一个可靠的共享存储和处理分析系统,使人们在存储和分析大数据时更加高效,其分布式文件系统 HDFS 可以实现数据的分布式存储。

基于此,文中采用 Hadoop 框架,对用户终端进行识别。提出了自己的识别方法并通过实验进行验证。

1 用户终端识别的方法

对于移动用户终端,最早是根据 HTTP 报文的 User-Agent(UA)报文头获取终端性能[8]信息。

早期,互联网是基于文本的,用户是通过敲命令的方式访问互联网。后来开发出了浏览互联的工具,这些工具称为用户代理,即 User-Agent。通过对 User-Agent[9]进行解析,可以获得用户终端的浏览器、操作系统、字符集和终端型号等。各种各样的 HTTP 请求报文[10]格式和字符集在不断变换,网络开发者无法应对,因此,标准化组织提出了两个规范:

(1)万维网协会提出了复合配置/偏好设置(CC/PP)标准化,即移动终端需要采用统一标准的格式向网络服务上传移动终端配置信息。

(2)开放移动联盟提出了一个 CC/PP 详细字典,即 User Agent Profile (UAProff)。UAProff 可以用来表示移动终端信息,而且在不好的无线网络中,网络服务根据终端性能,可以高效地在终端显示内容。一个移动终端如果遵循了 UAProff 标准,当它向服务端发送 HTTP 请求时,请求报文中会包含终端信息的 XML 文件 URL。服务端获取 URL 后,读取 XML 文件,得到终端信息。

虽然发布了以上两个标准,但是市场的移动终端类型很多,很多都没有遵循以上标准,致使终端识别没有得到好的解决。这时开源项目 WURFL(Wireless Universal Resource File)提出了另一种解决方案。UA 字段中含有很多用户移动终端的信息,WURFL 就是基于 UA 自身内容进行终端识别。WURFL 首先将 UA 的内容与包含终端信息的配置文件进行文本匹配,这样 Web 服务器就可以识别终端的型号和品牌。这种方式突破了 UA 格式的限制,然而仍有缺陷,因为它识别的终端信息需要依赖自身的资料库。当今的用户终端市场上,移动终端日益更新,但是 WURFL 资料库中终端型号及终端信息的更新速度远远赶不上用户终端的脚步,因此这种终端识别的准确率不是很高。

另外,传统上对 User-Agent 获取终端信息时,采用基于字符串匹配的方法。这种方法就是对 User-Agent 获取能够代表终端类的字符串,然后与一个内容很大的机器词典中的词条进行逐一匹配或者按照某种算法进行匹配,如果配到了,则获取到终端类型。该种方法实现简单,但是对于海量的用户数据来说具有很大的缺点。首先大容量的机器词典一般存于文件或者数据库中,占用很大的资源,而且匹配时对于机器词典具有很大的依赖性,不同词典会导致不同的结果。其次,当数据量很大时,终端匹配的效率非常低。当数据量超过千万级别时,数据库的性能会直线下降。上述方法几乎不能完成海量数据的用户终端型号识别。

文中是对 User-Agent 进行分析,获取相应用户终端信息。对 User-Agent 进行分词[11],采用正则表达式[12],首先过滤掉不代表用户终端信息的字符串,然后特定位置的字符串通过正则表达式获取。因为各种移动终端较多,比如手机、平板、PC 等,因此通过统计不同终端类型,写出不同的正则表达式进行匹配,得到一个正则表达式的配置文件。然后利用分布式爬虫获取电商上各种终端型号的相关信息作终端库信息。利用 Hadoop/hive 分布式快速处理大数据[13]量的特点进行终端匹配。为了使开发方便快捷,使用 hive 中的 UDF 功能,对用户终端类型进行识别。

2 相关技术

2.1 Hadoop 简介

Hadoop[14]是 Apache 组织管理的一个开源项目,是对于 Google 云计算理论 Big Table、MapReduce、GFS 的软件实现。Hadoop 可以让用户在不了解底层细节

万方数据

的情况下开发 MapReduce 程序,并可以运算和存储在硬件配置较低的商用集群上。Hadoop 主要包含两个核心组件,即分布式文件系统 HDFS 和分布式计算模型 MapReduce。HDFS[15] 是 Hadoop 的分布式文件系统,包含两种节点:namenode(管理者)和 datanode(工作者)。namenode 负责管理整个文件系统的命名空间,维护文件系统的树及树内的所有文件目录,并将这些元数据保存在 namenode 的本地磁盘上。namenode 也同时记录每个块及各个块的数据节点信息。HDFS 文件系统实际存储是由 datanode 负责,根据需要存储和检索的数据块,定期向 namenode 发送它们所存储块的列表,从而与 namenode 进行交互。MapReduce 是一个编程框架模型,可以进行稳定、高效、超大数据量的分布式分析计算。MapReduce 的执行主要包含 Map 和 Reduce 两个过程,当 Map 过程结束后还会进行 Shuffle/Sort 过程,负责对 Map 产生的输出进行排序和把 Map 输出传递给 Reduce。

2.2　Hive/UDF

Hive[16-17] 是构建在 Hadoop 上的数据仓库平台,其目的是让 Hadoop 上的数据操作与传统的 SQL 相结合,让开发更简单。Hive 可以在 HDFS 上构建数据仓库来存储结构化的数据,这些数据来源于 HDFS 的原始数据。Hive 拥有类似 SQL 的 HiveQL 查询语言,可以进行存储、查询、变换数据、分析数据等操作。通过解析 HiveQL[18] 语句,在底层被转换成 MapReduce 操作。更方便的是,Hive 提供自定义函数,即 UDF(User Defined Function)。虽然 Hive 中内置许多函数,但通常并不能满足用户的需求,因此 Hive 提供了自定义函数的开发。用户根据自己的需求编写相应的函数,从而更方便地对数据进行处理。在使用时,可以将自定义的函数打成 jar 包,在 Hive 会话中添加自定义 jar 文件,然后创建函数,继而使用。另外,也可以将自定义函数写到 Hive 的内置函数中,使之成为默认函数,这样就不需要在使用时重新创建。通过上述内容不难发现,使用 Hive 中的 HiveQL 语句比编写 MapReduce 代码更简单,而且提供 UDF 功能,更减少了开发的代码。

2.3　网络爬虫-WebMagic 爬虫框架

网络通常类比成一个蜘蛛网,每个节点就是一个网站,蜘蛛丝就是网站的链接,联系着各个网站。网络爬虫的基本原理就是通过网页中的链接地址来找到下一个网页,通常从网站主页面开始,读取网页 HTML 内容,通过解析 HTML 内容获取到想要的内容和其中的链接地址,然后从这个链接地址跳转到下一个网页,再进行解析。就这样一直不断地重复下去,直到把这个网站上的所有网页都分析完为止。

WebMagic 是一个开源的 Java 垂直爬虫框架,目标是简化爬虫的开发流程,让开发者注重逻辑功能开发。其主要功能包括:页面下载、链接提取、URL 管理和内容分析与持久化。WebMagic 中的各组件如下:

(1)Downloader 组件:采用 Apache HttpClient 对页面进行下载,获取页面的 HTML 以便后续处理。

(2)PageProcessor 组件:采用基于 XPath 和 CSS 的选择器,对网页的 HTML 内容进行解析,得到想要的信息。另外还可以从中获取新的页面链接。

(3)Schedule 组件:主要负责对获取的 URL 进行管理并去掉重复的 URL。开发人员不仅可以使用基于 JDK 的内存队列来管理 URL,也可以通过 Redis 对 URL 进行分布式管理。

(4)Pipeline 组件:对爬取结果进行自定义,获取想要的数据格式,并且可以将结果保存到文件中或数据库中,以实现数据的永久保存。

3　技术方案

基于 Hadoop 的电信运营商海量数据处理需要做到:

(1)不仅能够海量读取、存储多种结构的用户数据,还能对已有的数据仓库进行监控和管理。

(2)能够对大规模的数据进行高效处理运算,具有高数据吞吐量的功能。

(3)为了提高市场评估和网络运营能力,需要具有高拓展性,支持低成本数据挖掘,同时兼容多种应用。

(4)实时性的数据分析要求不高。

基于以上的数据特性和目标,根据 Hadoop 和 Hive 的特性,对电信宽带下的用户终端信息进行挖掘,可以更快地处理海量数据,以及更方便地对用户终端信息进行挖掘。同时编写 UDF 将源数据进行预处理,从而得到人们期望处理的数据格式。Hive 可以方便地插入用户编写的处理代码并在查询中调用它们。由此可以看出,利用 hive 执行任务的效率低于直接用 MapReduce 程序执行任务的效率,但是 Hive 给广大用户提供了最宝贵的 SQL 接口,并且避免了编写繁琐的 MapReduce 程序。为此,做出以下技术方案,如图 1 所示。

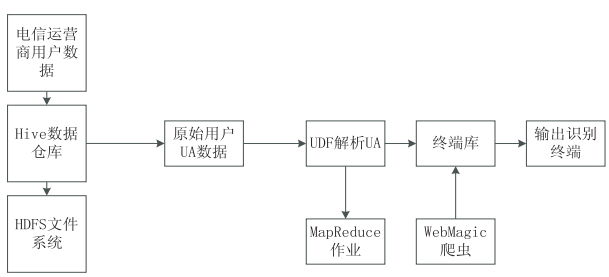

图 1　技术方案

因为用户数据中包含用户终端信息的字段,不是

万方数据

规则的,无法直接通过 HQL 语句特定的函数进行处理,所以采用 UDF 对用户终端的字段进行处理。通过对用户数据中含有 UA 的字段进行分析,写出能获取到用户终端类型的正则表达式,写成配置文件,作为正则匹配。UDF 函数逻辑如图 2 所示。

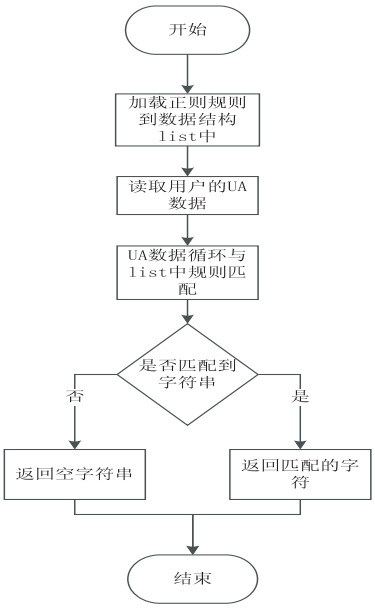

图 2 UDF 函数逻辑

4 实 验

4.1 实验环境

CDH 的 Hadoop 集群,一个 namenode 节点,五个 datanode 节点。

namenode 配置:CPU 的型号为 Intel 2650,内存为 16 G。

datanode 配置:CPU 为 Intel 2650,内存 32 G。

4.2 实验数据

(1)用户上网源数据。

为了验证该技术方案,搭建了实验测试环境,包括 100 台终端(PC/PAD/PHONE/盒子),采集了 12 个月的数据,共计 1 000 万条。数据记录包含字段有用户宽带账号、用户终端信息、用户访问的 URL。解析用户的 UA 就是在包含用户终端信息的字段中。这个字段的部分数据如图 3 所示。

(2)识别用户终端 UA 的正则表达式文件。

这个文件中的正则表达式观察用户上网的源数据中包含用户终端信息的字段,写出能匹配出用户 UA 的正则表达式。

4.3 实验步骤

(1)创建用户终端表。

```
create tablet_user_terminal_info (
username string,
ua string
)
row format delimited
fields terminated by '\t';
```

dalvik/1.6.0(linux; u; android 4.4.2; pe-tl10 build/huaweipe-tl10)
dalvik/1.6.0(linux; u; android 4.4.2; huawei p7-l05 build/huaweip7-l05)
dalvik/1.6.0(linux; u; android 4.4.4; coolpad 8675-a build/ktu84p)
dalvik/2.1.0(linux; u; android 5.1; m1 metal build/lmy47i)
dalvik/2.1.0(linux; u; android 5.0.2; vivo y51a build/lrx22g)
qiyivideo/7.5(ios;com.qiyi.iphone;ios9.3.2;iphone6,2)
dalvik/2.1.0(linux; u; android 5.0.2; x600 build/abxcnop5501304131s)

图 3 源数据的 UA 字段

(2)将编写解析 UA 的 udf 程序打成的 jar 包导到 Hive 的环境变量中。

```
addjar /home/shkd/20160418/udf/datamining-1.0-SNAPSHOT-jar-with-dependencies.jar;
```

(3)创造临时函数 uaparse,解析源数据中包含用户终端 UA 的信息的字段。

```
create temporary functionuaparse as 'cn.com.runtrend.datamining.udf.UAParserUDF'
```

(4)在 Hive 脚本调用创建的临时函数,经解析结构存入 t_user_terminal_info。

```
insert intot_user_terminal_info select username,uaparse(ua) from t_etlr_userinfo;
```

其中,t_etlr_userinfo 为用户上网源数据表。

(5)将解析 UA 中能代表用户终端类型的字段与终端库进行匹配,获取最终的终端类型。

4.4 实验结果

t_user_terminal_info 的部分内容见表 1。

表 1 t_user_terminal_info 的部分内容

Username	Ua
090010740425	huaweipe-tl10
090011380427	huaweip7-l05
090011458310	coolpad 8675-a

将解析的 UA 与终端库进行匹配,最后对类型进行汇总,得到识别的不同终端类型数量,如图 4 所示。

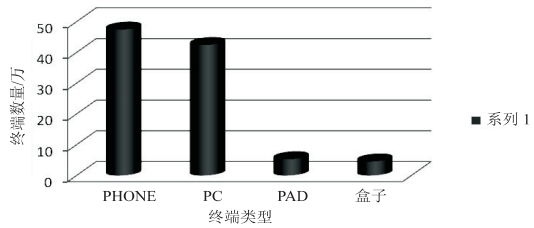

图 4 识别的终端类型数量

万方数据

4.5　实验评估

对于其他用户终端,采用准确率进行评估。记各种终端的总样本数为 n,准确识别的数量为 m,识别终端的准确率为 $c = m / n * 100\%$。对于上面实验得到的识别准确率如图 5 所示。

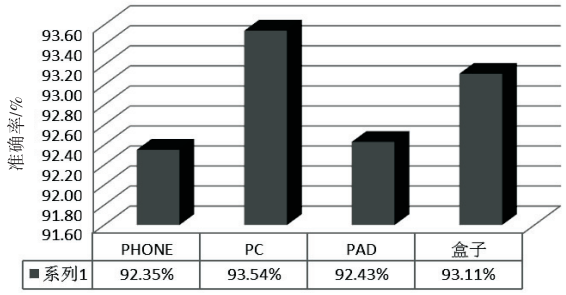

图 5　识别的终端准确率

最终识别的准确率并不是非常高,这其中存在一些主要原因:

(1)爬虫获取的终端库可能并不完整,导致有些识别的终端不能被匹配到;

(2)正则匹配的配置文件中,正则匹配表达式并不完善,从而导致识别的终端出错。

因此,正则表达式的配置文件和爬虫的终端标准库需要不断更新,才能提高识别的准确率。

5　结束语

通过实验发现,基于 Hadoop/Hive 集群可以实现对家庭固网宽带下用户终端信息的识别。Hadoop 集群具有高可靠性、高拓展性、高容错性。为分析固网宽带用户提供了一种非常好的技术手段。利用 Hadoop 平台更高效,对电信宽带用户能精准挖掘有用信息,实现流量变现。另外,Hive 提供 SQL 接口,利用 SQL 可以更方便地使用 Hadoop,而且 Hive 还提供自定义函数,避免了复杂的 MapReduce 程序的编写,让开发更简单。

参考文献:

[1] 张　引,陈　敏,廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展,2013,50:216-233.

[2] 冯明丽,陈志彬. 基于电信运营商的大数据解决方案分析[J]. 通信与信息技术,2013(5):36-40.

[3] Baghel S K, Keshav K, Manepalli V R. An investigation into traffic analysis for diverse data applications on smartphones [C]//National conference on communications. [s. l.]: IEEE,2012:1-5.

[4] 陈　勇. 大数据及其商业价值[J]. 通信与信息技术,2013(1):59-60.

[5] 王研昊,马媛媛,杨　明,等. 基于隐性标识符的零权限 Android 智能终端识别[J]. 东南大学学报:自然科学版,2015,45(6):1046-1050.

[6] Sagiroglu S, Sinanc D. Big data: a review[C]//International conference on collaboration technologies and systems. [s. l.]: IEEE,2013:42-47.

[7] 程　莹,张云勇,徐　雷,等. 基于 Hadoop 及关系型数据库的海量数据分析研究[J]. 电信科学,2010,26(11):47-50.

[8] 梁其峰. WLAN 终端识别技术研究[J]. 科技传播,2013(18):186-188.

[9] Wu T, Xu Z, Ni L, et al. Towards a media fragment URI aware user agent[C]//Web information system and application conference. [s. l.]: IEEE,2014:37-42.

[10] La V H, Fuentes R, Cavalli A R. Network monitoring using MMT: an application based on the user-agent field in HTTP headers[C]//International conference on advanced information networking and applications. [s. l.]: [s. n.],2016:147-154.

[11] 何　莘,王琬芜. 自然语言检索中的中文分词技术研究进展及应用[J]. 情报科学,2008,26(5):787-791.

[12] 徐　乾,鄂跃鹏,葛敬国,等. 深度包检测中一种高效的正则表达式压缩算法[J]. 软件学报,2009,20(8):2214-2226.

[13] Pal A, Agrawal S. An experimental approach towards big data for analyzing memory utilization on a hadoop cluster using HDFS and MapReduce[C]//International conference on networks & soft computing. [s. l.]: [s. n.],2014:442-447.

[14] White T. Hadoop 权威指南[M]. 北京:清华大学出版社,2015.

[15] 刘　鹏,黄宜华,陈卫卫. 实战 Hadoop 开启通向云计算的捷径[M]. 北京:电子工业出版社,2011.

[16] 谢　恒,王　梅,乐嘉锦,等. 基于 hive 的计算结果特征提取与重用策略[J]. 计算机研究与发展,2015,52(9):2014-2024.

[17] Ganesh S, Binu A. Statistical analysis to determine the performance of multiple beneficiaries of educational sector using Hadoop-Hive[C]//International conference on data science & engineering. [s. l.]: IEEE,2014:32-37.

[18] Bhardwaj A, Vanraj, Kumar A, et al. Big data emerging technologies: a casestudy with analyzing twitter data using apache hive[C]//International conference on recent advances in engineering & computational sciences. [s. l.]: IEEE,2015:1-6.

万方数据