基于DPI和大数据分析的宽带家庭画像

刘 超[1],刘馨璐[1],王 攀[2],张丽娜[1]
(1.江苏大学 电气信息工程学院,江苏 镇江 212013;
2.南京邮电大学 物联网学院,江苏 南京 210003)

摘 要:随着大数据时代的到来,固网宽带下家庭用户的行为分析成为运营商行业面临的新一轮的机遇与挑战,对用户的手机号码和终端类型实施准确提取则是重中之重。文中采用Hadoop框架,应用DPI深度挖掘和分布式爬虫等技术,对用户行为进行分析,从而快速准确地获取用户的手机号码和终端类型,将家庭用户的行为偏好相关联,最终塑造固网宽带下的家庭画像,从整体上洞悉用户的需求,做到精准营销,改善用户体验质量。实验结果表明,手机号码提取准确率达86%以上,终端识别准确率达90%以上。在识别率与准确率分析的基础上,对信息输出表做进一步分析,包含单个IP接入用户数、手机用户归属运营商及用户手机型号。为运营商提供更加丰富、准确、完善的固网宽带下的家庭画像。

关键词:DPI;Hadoop;号码提取;终端识别;家庭画像
中图分类号:TN91 **文献标识码**:A **文章编号**:1673-629X(2018)12-0162-05
doi:10.3969/j.issn.1673-629X.2018.12.034

Broadband Family Portrait Based on DPI and Big Data Analysis

LIU Chao[1],LIU Xin-lu[1],WANG Pan[2],ZHANG Li-na[1]
(1.School of Electrical and Information Engineering,Jiangsu University,Zhenjiang 212013,China;
2.School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:As the period of big data is coming,the behavior analysis of family users in fixed-line broadband is an opportunity and challenge for the carrier industry. Especially,extracting the user's phone number and terminal type is a top priority. Based on the Hadoop framework,we adopt the DPI technology,such as depth mining and distributed crawler,to rapidly and exactly analyze the user behavior and obtain the user's mobile phone number and terminal type. By correlating the behavioral preference of family users,the family portrait under fixed network broadband is finally shaped. The user demand is understood from the whole,achieving precise marketing and improving quality of user experience. The experiment shows that the extraction accuracy of the mobile phone number is more than 86%,and the accuracy of terminal recognition is more than 90%. Based on the recognition rate and the accuracy analysis,the information output table is further analyzed,which contains individual IP access users,cell phone users belonging to the carrier and the user's cell phone model,to provide the operator with the rich,accurate and complete family portrait under the fixed-line broadband.

Key words:DPI;Hadoop;number extraction;terminal identification;family portrait

0 引 言

近年来,以海量数据处理为目标的大数据技术成为新的研究热点。所谓"大数据",是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集合[1]。伴随着Facebook、Google、微博、APP等网络服务的蓬勃发展,对网络用户行为的分析和研究引起了众多研究者的兴趣。现代生活中,网络行为成为人们日常生活的主要成分,其中蕴含了许多用户社交关系、用户日常行为习惯以及个人兴趣喜好等诸多有价值的信息[2]。但仅仅分析每个用户的个体需求是远远不够的,在这个高速发展的时代,每个家庭都是社会的一部分,把家庭看作一个单独的整体来分析家庭的整体需求,将家庭用户的行为偏好相关联,完善成一幅家庭画像,从整体上洞悉用户的需求,强化客户关怀,做到精准营销,将会从另一个层面改善用户的体验质量,增加运营商的业务效率。

收稿日期:2018-01-15 **修回日期**:2018-05-23 **网络出版时间**:2018-07-04
基金项目:江苏省六大人才高峰项目(XXRJ-012);江苏省博士后基金(1402095C);江苏大学高级人才启动项目(1291140025)
作者简介:刘 超(1977-),男,硕导,副教授,研究方向为无线通信与计算机通信网;刘馨璐(1993-),女,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.20180703.1513.042.html

1　相关研究

国外的各大企业纷纷提出大数据的规划和政策,以推动大数据的发展。目前,Google、Facebook 等企业正在应用大数据技术来发展云端服务和社交软件。亚马逊公司很早就对用户的浏览信息实施数据分析,根据用户的浏览信息等数据,推算出用户的行为偏好,从而对用户实施精准推送[3]。相对于国外较成熟的大数据分析技术,国内目前还处于发展初期,对应的市场规模较小[4]。阿里巴巴会根据用户网购时浏览的商品信息和停留时间,交易行为可以进一步了解消费者的喜好,从而为用户推荐感兴趣的产品。

有学者对用户行为开展了许多有意义的分析与研究,并取得了大量极具影响力的研究成果[5]。刘海等[6]基于 4C 理论构建了"用户画像"数据库,通过对数据库的挖掘来进行消费者群体细分。在此基础上,从营销的角度构建了精准营销细分模型,重构消费者的需求、精准定位消费者群体。应晓敏等[7]提出了一种面向个性化服务的客户端细粒度用户兴趣建模方法,并且将用户兴趣不再简单地分为用户感兴趣的类和用户不感兴趣的类,而是按照人们通常对兴趣的理解划分为不同的兴趣类。宋竹等[8]提取了电信数据中手机通话与上网的基本特征,对通话和上网行为的频率分布进行曲线拟合,通过对通话和上网时间的归一化,定义了用户的使用偏好。

可以看出,目前的研究并没有涉及对家庭中的手机号码和终端类型做精准提取和分析,尤其是绝大多数的分析和研究仅仅针对个体用户,而非家庭用户。对于运营商而言,仅仅分析个体用户的行为特点是不够全面的,在宽带家庭账号下,根据整个家庭的日常上网情况可以分析整个家庭的行为习惯。可以分析出该家庭账号下用户总数及年龄结构层次、网络接入设备、手机品牌型号以及其他终端设备,根据分析结果可以得到一幅家庭画像,如图 1 所示。

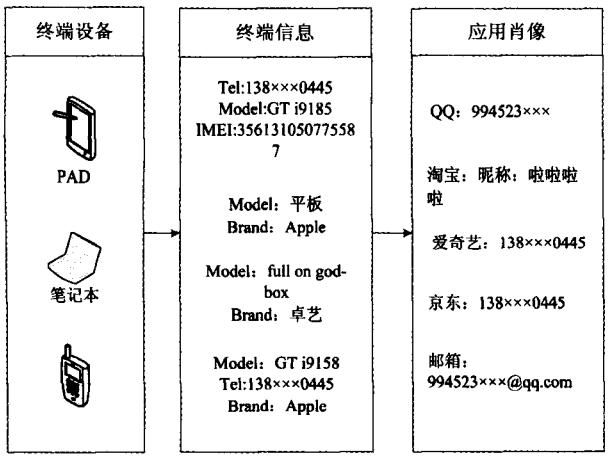

图 1　家庭画像

通过对整个家庭数据流量的分析处理,能够从整体上把握家庭所有用户的需求,从而借助互联网推送平台等方式更加精准地给所有家庭用户推荐更合适的产品和服务。

文中在利用运营商合法获取的数据基础之上,采用 DPI(deep packet inspection,深度分组检测)、Hadoop 框架、分布式爬虫等技术,提取家庭宽带下手机号码并对不用的用户终端进行识别,最终构建出反映家庭用户特征和行为兴趣的家庭画像。

2　相关技术

2.1　DPI 技术

DPI 是相对普通报文检测的一种全新的检测技术,即对第七层应用层的内容进行深度分析,从而根据应用层的净荷特征识别其应用类型或内容[9]。DPI 技术的核心点在于维护一个高准确性、高实时性的应用特征库,从而保障应用特征识别的准确性、实时性,进而保障运营商对应用的管控准确性和实时性。

2.2　Hadoop 框架

Hadoop 是一个能够对大量数据进行分布式处理的软件框架[10]。它以一种可靠、有效、可伸缩的方式进行数据处理,具有高可靠性、高扩展性、高效性、高容错性、低成本等优点。HDFS 和 MapReduce 是 Hadoop 框架的核心设计。HDFS 为海量的数据提供了存储,MapReduce 为海量的数据提供了计算。

Hive 是基于 Hadoop 的数据仓库工具,可以将结构化的数据文件映射为一张数据库表,并提供简单的 SQL 查询功能,能够将 SQL 语句转换为 MapReduce 任务进行运行。

2.3　网络爬虫 WebMagic 技术

WebMagic 项目代码分为核心和扩展两部分[11]。核心部分(webmagic-core)是一个精简的、模块化的爬虫实现,而扩展部分则包括一些便利的、实用性的功能。

WebMagic 的结构分为 Downloader、PageProcessor、Scheduler、Pipeline 四大组件,Downloader 负责从互联网上下载页面,以便后续处理;PageProcessor 负责解析页面,抽取有用信息以及发现新的链接;Scheduler 负责管理带抓取的 URL 以及去重工作;Pipeline 负责抽取结果的处理,包括计算、持久化到文件或数据库等。

3　基于 DPI 和大数据分析的宽带家庭画像塑造方法

塑造一个完善并全面的家庭画像,首先需要确定家庭的唯一标识,即宽带账号,因为每个家庭的宽带账号是唯一的。家庭中用户的手机号码和终端类型是家

庭画像的关键属性。其中终端类型包括终端的品牌、型号及上市时间等。通过数据包提取到以上数据后，必须对其进行去噪处理，以确保提取出的信息是真实、有效、完整的。总体技术路线包括确定家庭画像唯一标识，确定家庭画像属性、号码和终端提取和去除噪声数据。通过分析得到，在塑造家庭画像的过程中，手机号码的提取和终端类型的识别尤为重要。

3.1 用户手机号码的提取

为了提升用户手机号码提取的准确率，文中采用 Hyperscan 进行匹配。采集家庭宽带下的网络流量，采用 DPI 中净荷特征匹配技术对采集到的数据进行清洗，过滤掉无关流量后，再利用 Hyperscan 高速匹配，提取出数据包中疑似手机号的关键字。在获取大量关键字后，将通过 DPI 处理后的数据和关键字导入 Hadoop，对数据分类存储，进行数据匹配，最终提取出较为准确的用户手机号码。

3.2 用户终端的识别

移动用户终端的识别起初是根据 HTTP 报文的 User-Agent 报文头获取终端性能信息。对 UA 解析获取终端信息时，通常采用的是基于字符串匹配的方法。该方法实现较简单。随着用户数据的迅猛增加，终端匹配效率逐渐降低。文中采用一种改进的用户终端的识别方法，首先对 UA 进行分词，然后采用正则表达式过滤掉不代表用户终端信息的字符串，最后通过正则表达式获取特定位置的字符串。家庭宽带下的用户使用终端类型较多，有手机、平板、PC、电视机、盒子等，通过统计不同终端类型，写出不同的正则表达式进行匹配，从而得到一个正则表达式的配置文件。同时采用分布式爬虫 WebMagic 获取电商上各种终端型号的相关信息作终端库信息。最终根据 Hadoop/Hive 分布式快速处理大数据量的特点对用户终端进行准确识别。

4 技术方案

反映家庭画像的主要元素是家庭宽带下对用户的手机号码的提取和终端类型的识别。主要由数据采集、数据清洗、数据提取及数据挖掘与分析共四个部分组成。

4.1 数据采集层

家庭宽带用户的 HTTP 上行流量从分流平台以千兆电口形式实时输出到高速采集服务器；对于已经建成固网宽带 DPI 大数据平台的运营商，无需配置数据采集服务器，将 DPI 日志文件直接输送到数据清洗系统，即可完成数据采集工作。

该系统数据流量的采集采用 Libpcap[12]。Libpcap 采用基于网卡的原理捕获数据包，支持所有基于 Unix 的操作系统，能够快速采集和过滤网络流量。Libpcap 可以根据用户已经设定好的过滤规则对数据进行逐一匹配，匹配成功则放入内核缓冲区，并传递给用户缓冲区，匹配失败则直接丢弃。

4.2 数据清洗层

为了获取用户的真实点击量，保证数据挖掘的准确性和高效性，在数据分析前必须对数据进行清洗，过滤掉非用户点击的流量，如图片流量、脚本流量、广告以及框架等无效数据。

利用 DPI 数据清洗系统，去除采集到的流量数据中的大量冗余信息，再将数据传递给 Hadoop 分析平台，以保证所获取数据的准确性和分析的高效性。数据清洗首先过滤非 TCP/IP 或者非 Http/get 流量，然后过滤后缀为 jpg、gif、css 等图片和脚本流量，再过滤带有指定特征字符串如广告、框架类型的流量，最后过滤自刷新页面和存储过滤后剩下的数据。

通过 DPI 技术深度挖掘数据包，提取相关信息后判断数据包的协议类型，进行首次过滤，去除非 TCP/IP 和非 HTTP/GET 流量。然后在剩余的数据包中对应用层进行解析，进行再次过滤，丢弃无效数据，例如 uri 后缀为 jpg、gif、png 等图片、脚本及框架类型的流量和自刷新页面等，这些数据中不包含用户的相关信息，最后存储二次过滤后剩余的有效数据。

4.3 数据提取层

将经过 DPI 清洗后的数据结果导入 Hadoop 平台的 Hive 数据库中。借助 Hive 提供的 SQL 快捷接口可以方便用户在插入和查询数据时书写代码，快速处理海量数据。

清洗过后的完整数据包基本上都包含 uri、UA、host 等字段。手机号码多来自同一个数据包的 host 和 UA 字段，而终端类型则存在于 UA 字段中。数据提取过程包括 Http 字段提取、AAA 账号匹配、统一解码和特征字符串匹配，然后输入到手机号码报文特征库或者终端信息库。

通过采用 DPI 深度报文监测技术和 Hyperscan 高速匹配技术过滤清洗后，记录结果包含时间戳信息、用户 IP、宽带账号、手机号、手机关键字、cookie 终端缓存数据、host 主机名、UA 用户代理等内容项。对采集到的报文做关键信息提取后，再利用特征字符串匹配的方法提取准号码清单。利用 WebMagic 爬虫框架对终端信息进行爬取，生成终端型号库，爬取结果部分数据如表 1 所示。通过对用户数据含有 UA 字段进行分析，找出最常出现的 UA 字符串，根据这些 UA 字符串编写正则表达式生成正则表达式库。编写 MapReduce 代码通过正则表达式库去 UA 字段提取出 UA 中的终端型号，测试通过后打包成 jar 包，通过 Hadoop 集群中

的 Hadoop jar 命令提取出数据中所有 UA 字段中的终端。

表 1　终端类型爬虫信息库

品牌	终端类型	终端型号
OPPO	手机	Oppo r9
华为	手机	HUAwei c5730
三星	平板电脑	Sm t231
苹果	iPad	iPad Air2
乐视	智能电视	Letv x50 air

4.4　数据挖掘与分析

为了获取更加准确的信息,需要对清洗后的数据进行分析验证。

首先通过号码正则表达式提取出所有的手机号码,通过号码出现的天数和频率,以及号码所对应的终端数量,找到该账号下出现频率和天数较高的以及号码对应终端数较少的识别为该账号下的手机号码。其次,通过爬虫获取到如中介、商户、热线等号码,进行"伪号码"过滤,去除非真实用户的手机号码。

对剩余的数据再进行决策树分析[13-14],通过对某一手机号码的归属地、出现频次以及出现的时间段进行分析,判别号码清单中挖掘出的手机号码是否真实活跃在其出现的家庭宽带下[15]。具体决策过程如图 2 所示。

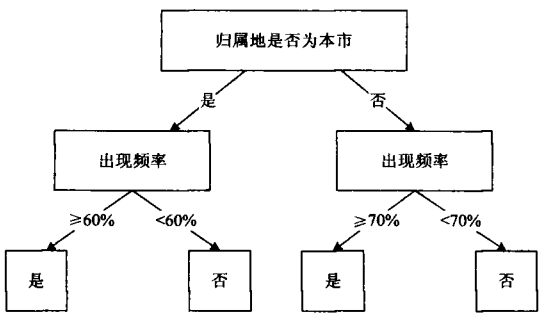

图 2　决策树分析图

5　实验与结果分析

5.1　实验环境

为了验证手机号码和终端信息获取的准确性,以固网宽带下的家庭用户为基础搭建实验环境,利用该系统获取到的信息和实际用户信息进行对比,通过号码提取与终端识别的准确率来判断信息获取的准确性。实验环境结构图包括数据存储器、Hadoop 处理服务器、采集服务器、家庭路由器和家庭用户等部分,其中包含对用户网络数据的采集,DPI 数据清洗和 Hadoop 数据分析。

5.2　实验结果

选取 1 000 个友好家庭用户,采取问卷调查等方式事先采集家庭用户的基本数据,包含家庭的人口情况、手机号码及使用的终端类型等。通过与运营商合作,利用该系统采取分光方式获取用户的上网流量数据。对获取的数据进行清洗、提取、分析后可以得到信息输出表,包含用户宽带账号、手机号码、终端品牌、终端型号、上市时间、QQ 号、用户使用邮箱账号等信息。

对以上信息进行整理与分析,可以获得手机号码和终端类型的识别率曲线图,如图 3 所示。

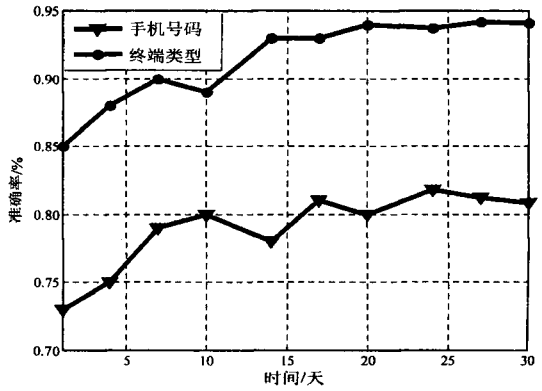

图 3　手机号码和终端类型的识别率曲线

终端类型的特征信息较为单一准确,而手机特征关键字包含的类型和数量远大于终端类型的特征信息,导致终端类型的识别率高于手机号码提取的识别率。长期观察后可以发现两者的识别率均有所提高,其中手机号码的识别率达到 84% 左右,终端类型的识别率则达到 92% 左右。

参照问卷调查的结果,与信息输出表进行比对,可以进一步获得手机号码提取和终端类型识别的准确性曲线图,如图 4 所示。

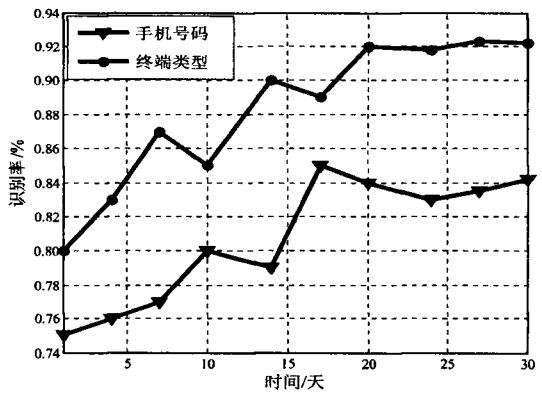

图 4　手机号码和终端类型的准确率曲线

由于识别出的手机号码中有部分号码非该家庭用户的固有号码,导致终端类型识别的准确率仍然高于手机号码提取的准确率。随着时间的递增,两者的准确率均逐渐上升并趋于稳定,手机号码识别的准确率维持在 80% 左右,而终端类型的准确率则达到 95% 左右。

在识别率与准确率分析的基础上,对信息输出表做进一步分析,包含单个 IP 接入用户数和用户手机型号等,具体分析结果如图 5、图 6 所示。

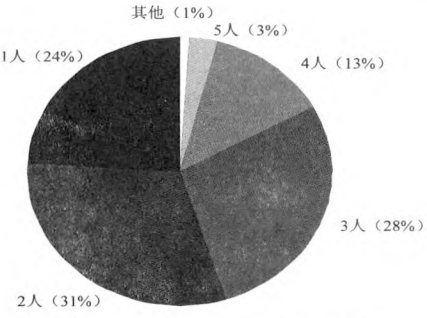

图 5 单 IP 接入用户数分析

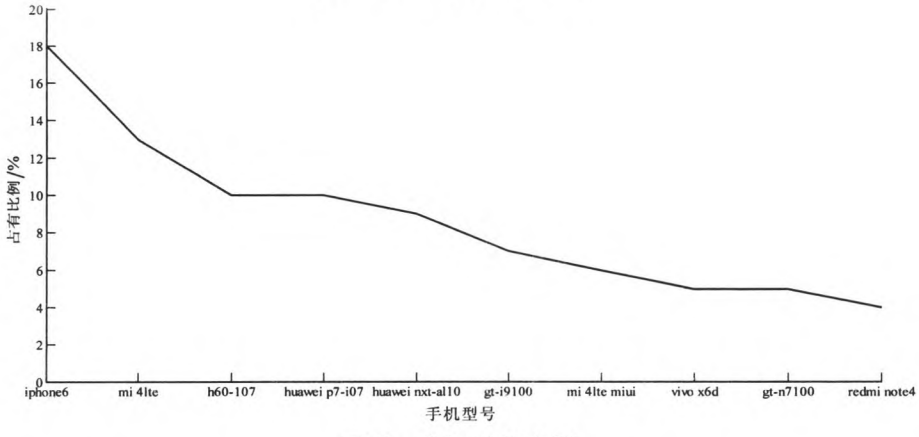

图 6 手机型号分析

分析结果表明，单个 IP 下接入人数以 2 人居多，其次是 3 人和 1 人，即在统计的绝大部分单个家庭用户中，使用 2 部手机的情况较多，同时使用苹果手机的用户较多，其次是小米和华为。

6 结束语

利用 DPI 深度报文检测技术、Hyperscan 高速匹配、Hadoop 和 WebMagic 爬虫技术能够以较高的识别率和准确率快速精准地识别家庭宽带下用户的手机号码和终端类型，高效地构建固网宽带下的家庭画像。下一步工作将会对用户信息进行全方位提取，包括接入终端信息、用户行为偏好等，并对以上信息进行行为建模、深度挖掘和知识发现，具体分析家庭每个用户的网络行为习惯和兴趣爱好，从整体上洞悉用户的需求、强化客户关怀，为运营商提供更加丰富、准确、完善的固网宽带下的家庭画像。

参考文献：

[1] MCAFEE A, BRYNJOLFSSON E. Big data: the management revolution[J]. Harvard Business Review, 2012, 90(10): 60-66.

[2] GARG S, SARJE A K, PEDDOJU S K. Improved detection of P2P Botnets through network behavior analysis[C]//International conference on security in computer networks and distributed systems. [s. l.]: [s. n.], 2014: 334-345.

[3] ACQUISTI A, BRANDIMARTE L, LOEWENSTEIN G. Privacy and human behavior in the age of information[J]. Science, 2015, 347(6221): 509-514.

[4] LIAO W, AL-KOFAHI K, MOULINIER I. Feature engineering and user behavior analysis: US, US9552420[P]. 2017.

[5] 崇 林. 基于运营商大数据的互联网海量用户行为分析系统设计与实现[D]. 南京: 南京邮电大学, 2016.

[6] 刘 海, 卢 慧, 阮金花, 等. 基于“用户画像”挖掘的精准营销细分模型研究[J]. 丝绸, 2015, 52(12): 37-42.

[7] 应晓敏, 刘 明, 窦文华. 一种面向个性化服务的客户端细粒度用户建模方法[J]. 计算机工程与科学, 2003, 25(6): 39-42.

[8] 宋 竹, 秦志光, 罗嘉庆, 等. 电信数据中用户行为特征测量与分析[J]. 电子科技大学学报, 2015, 44(6): 934-939.

[9] 吕锦扬. DPI 技术在移动数据网络分析的应用[J]. 电信技术, 2013(6): 72-75.

[10] WHITE T, CUTTING D. Hadoop: the definitive guide[M]. [s. l.]: O'Reilly Media Inc., 2011.

[11] LIH T M, CHOONG W K, CHEN C C, et al. MAGIC-web: a platform for untargeted and targeted N-linked glycoprotein identification[J]. Nucleic Acids Research, 2016, 44: W575-W580.

[12] 温曙光, 谢高岗. libpcap-MT: 一种多线程的通用数据包捕获库[J]. 计算机研究与发展, 2011, 48(5): 756-764.

[13] 李泓波, 彭三城, 白劲波, 等. 一类决策树 ID3 改进算法探究[J]. 软件导刊, 2016, 15(2): 30-32.

[14] 王小巍, 蒋玉明. 决策树 ID3 算法的分析与改进[J]. 计算机工程与设计, 2011, 32(9): 3069-3072.

[15] 路 翀, 徐 辉, 杨永春. 基于决策树分类算法的研究与应用[J]. 电子设计工程, 2016, 24(18): 1-3.