研究与开发

基于 DPI 的 LTE 网络用户行为感知系统的设计与实现

王　建，张治中，骆云龙

（重庆邮电大学通信网与测试技术重点实验室 重庆 400065）

摘　要：针对电信运营商越发迫切的智能管道需求，提出了一种具有自学习功能的移动互联网用户行为感知系统的解决方案。本方案针对传统监测系统用户感知度低、统计能力不足等缺点，对现行 LTE 网络 S1 接口用户面协议进行分析，并结合当前互联网主流的行为分析技术——深度分组检测（DPI）技术和聚类爬虫技术的优势，实现了以协议解码、业务呼叫/事务详细记录（xDR）合成为基础的 LTE 网络用户行为精准分析。本系统经现网数据验证，能达到既定目标，对满足智能管道的需求具有一定的指导价值。

关键词：LTE 网络；S1 接口；深度分组检测；聚类爬虫；用户行为感知

doi: 10.3969/j.issn.1000-0801.2014.07.012

Design and Implementation of DPI-Based User's Behavior Perception System in LTE Network

Wang Jian, Zhang Zhizhong, Luo Yunlong

(Key Laboratory on Communication Network and Testing Technology,
Chongqing University of Post and Telecommunications, Chongqing 400065, China)

Abstract: Based on telecom operators increasingly urgent demand for intelligent pipeline, a kind of implementing scheme of user behavior perception system with self-learning function was proposed, which analyzed the protocols of user plane on S1 interface in long term evolution (LTE) network. Aiming at the traditional monitoring system with poor user perception and insufficient statistical capacity, the proposed system achieved user's behavior refined analysis which combined protocol decoding and services call/transaction detail records (xDR) synthesis method with the advantages of current prevailing internet behavior analysis technologies—deep packet inspection (DPI) technique and focused crawler technology. With the verification of existing network data, this system not only achieves the goal, but also has the guiding value of meeting the demand for intelligent pipeline.

Key words: LTE network, S1 interface, deep packet inspection, focused crawler, user's behavior perception

1　引言

随着国内 4G 牌照的相继发放，各大运营商加速了 LTE 网络的部署。基于 LTE 技术特征的高带宽、高质量的宽带网络业务服务为广大客户提供高速率、低时延的优质体验的同时，也给运营商网络运营支撑和管理带来了巨大压力[1]。借鉴移动互联网营销模式，结合用户自身的业务需求、上网习惯等综合考虑，实现对业务的精细识别和准确统计，增强客户感知，打造移动互联网智能管道成为了运营商当务之急。

传统的业务面监测技术基础是基于 IP 五元组的业务识别技术和针对上层协议消息头的 xDR（call/transaction

*　国家高技术研究发展计划（"863"计划）基金资助项目（No.2014AA01A706），2013 年重庆高校创新团队基金资助项目

detail record,xDR)合成技术。一方面,基于IP五元组的识别技术只能识别出常见的应用层协议,如HTTP、SMTP、FTP等,基于P2P技术的应用也往往采用隐藏或假冒端口号来躲避监管,另一方面,基于上层协议消息头的xDR合成技术无法细粒度地反映出用户的行为偏好。随着智能手机和数据业务的爆发增长,新一代客户感知网络监测体系具有广阔的发展需求和市场空间。在国内,华为推出的SmartCare是一个旨在服务于高端细分市场的网络监测解决方案,该方案帮助运营商解决端到端的业务质量提升和保障、用户体验和满意度提升及保障的问题[2]。面对这一发展趋势,中兴通讯提出一种基于云计算架构的用户体验应用解决方案——ZSmart CEMS。该方案采用大数据分布式以及并行处理架构,实现计算资源的动态调配以及利用率。通过DPI(deep packet inspection,DPI)技术识别业务类型,与业务质量、用户数据、终端信息相关联,实现对现网业务的实时监测[3]。

但是这些产品往往价格昂贵,或是应用范围局限,后台维护成本较高,无法满足LTE网络多业务大流量的监测需求。因此,实现一套完整且价格低廉的用户感知系统仍然是任重道远。

基于这一出发点,本文重点研究了LTE网络S1接口业务中用户行为习惯的精确感知与监测,在传统检测系统基础上,运用DPI技术和网络爬虫技术,实现了一套具有自学习功能的用户行为识别与分析系统。

2 系统设计

DPI技术对业务识别的应用已较为成熟,目前在一些网络入侵检测系统如BRO[4]和Snort[5]中得到了运用。对于业务流的识别来说,DPI是一种重要的识别手段,即通过对业务流中的特定数据报文中的"指纹"加以检测,并与特征信息库匹配,能确定业务流承载的具体信息。

在当前移动互联网(GPRS/EDGE/HSDPA)环境下,大部分业务由HTTP承载。伴随着3G、4G网络以及智能终端的发展,自有协议也登上了移动互联网的大舞台。

- 对于以HTTP承载的报文一般采用HTTP报头的URL关键字段与user-agent组合识别的方式进行识别。通过对URL特征字符串的提取,识别出移动互联网用户具体的操作。如用户在iTunes手机客户端下载"铁路12306"订票软件,user-agent为"iTunes-iPhone/5.1.1",host为:"a*.phobos.apple.com",用户浏览该应用特征字符串为"mzl.ihlxdfje.*.jpg",用户下载该应用的特征字符串为"mzl.ihlxdfje.*.jpg? downloadKey="("*"表示通配符,匹配多个字符)。
- 对于自有协议(如微信、QQ、米聊等即时通信软件)和部分P2P应用中采用非密文传输的报文,需要对传输层及以上的数据进行分析,提取出具体业务的特征指纹,这些指纹可能是指定的端口号、指定的字符串以及比特序列。如对微信的协议报文总结归类提取比特序列特征,其中"0xed 0xed"是报文结尾特征字段,"0x02"表示用户发送文字,"0x09"表示用户发送图片,"0x38/0x39"表示摇一摇功能。

特征库的信息健全程度是衡量整个系统业务识别能力的标准。针对HTTP承载的报文,需要结合网络爬虫,从PC端获取业务信息。对于自有协议的报文,则可以通过Wireshark工具分析业务报文,手动录入"指纹"信息到特征信息库。

如图1所示,S1接口是LTE的网络中演进的通用陆地无线接入网(evolved universal terrestrial radio access network,E-UTRAN)和演进分组核心网(evolved packet core,EPC)的通信桥梁,是LTE网络的核心节点。S1接口作为一个逻辑接口,它分为控制面和业务面两个部分,其

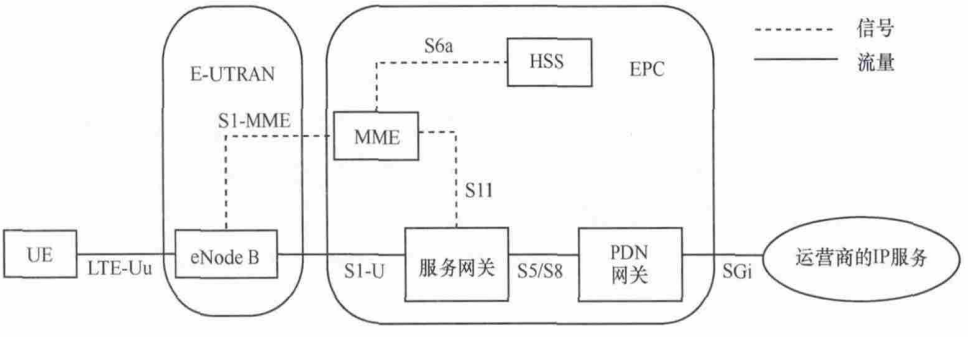

图1 LTE网络架构

中 S1-MME 接口是信令面接口，用于传送会话和移动性管理等控制信息；S1-U 是业务面接口，用于传输用户数据业务[6]。本方案选取业务面接口——S1-U 接口作为数据采集点，通过对该接口的协议栈解码和 xDR 合成获取用户某次操作的特征标识。

用户行为感知系统主要包括数据采集模块、S1 接口协议栈解析模块、数据存储模块、特征分析/网络爬虫模块、DPI 业务识别模块和可视化业务统计模块。总体实现框架如图 2 所示。

各模块功能设计如下：

(1)数据采集模块实现对现网 S1 接口数据实时采集，将数据交给协议栈解析模块；

(2)协议栈解析模块实现对 S1-U 接口协议栈解码、xDR 合成，提供基础数据，并为 DPI 模块接口提供关键字段；

(3)数据存储模块采用关系型数据库，对用户基础数据、特征数据、用户行为数据等进行关联存储；

(4)网络爬虫模块根据一定的业务主题对 Web 网站的数据进行内容级别上的爬取、提炼，并将数据信息写入数据仓库；

(5)实时 DPI 和二次 DPI 模块是保障系统性能的关键，实时 DPI 主要用于完成业务类型识别，二次 DPI 模块将完成业务内容的识别，获取行为标识的内容信息；

(6)可视化业务统计模块根据不同维度、需求进行识别结果统计，并提供可视化界面显示。

3 系统实现

3.1 S1-U 接口协议栈解析模块

S1-U 接口协议栈解析采用模块化设计思想。子功能模块包括原始数据分组、协议栈解码、xDR 合成和实时 DPI 识别。协议栈解码模块针对各层协议，自下往上逐层对协议进行解码，并将解码结果保存到链表结构体中。xDR 合成应用散列索引和超时检测机制将同一呼叫流程的消息关联在一起，完成上下行流量统计、呼损统计等功能[7]。该模块解码结果和 TCP 分组或 UDP 分组的负载净荷作为 DPI 识别模块的数据。图 3 是 S1-U 接口协议栈解析实现框架。

其中，GTP-U(GPRS tunnel protocol-user plane, GPRS 通道协议—用户面)是一组基于 IP，用于 GSM、UMTS 和 LTE 网络以支持 GPRS 的用户面通信协议。它以 GTP 为基础，并采用一种通用的隧道封装方法，能对各种类型的分组数据进行透明封装传输。

在感知系统中，对自有协议的识别需要将传输层以上的载荷码流实时处理，故将这部分协议的识别整合到实时的解码模块中。为了进一步满足实时 DPI 模块高性能要

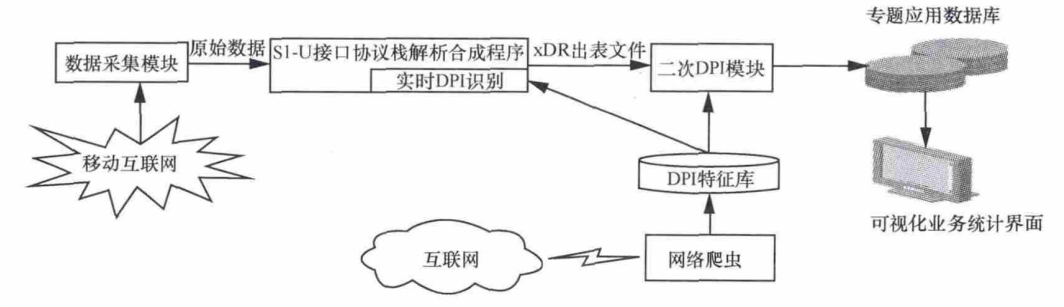

图 2 用户行为感知系统整体框架

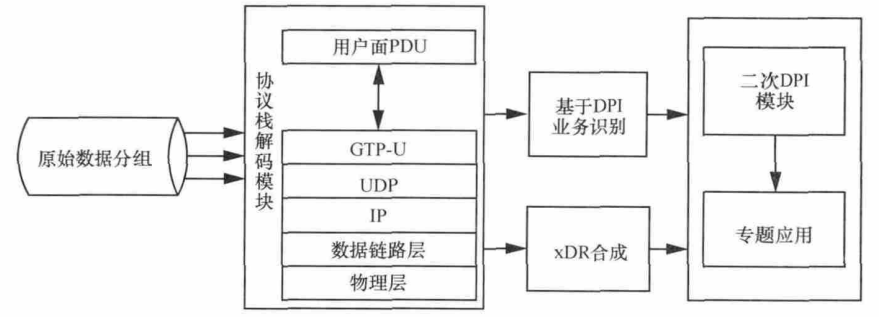

图 3 S1-U 接口协议栈解析实现框架

求，该模块的设计充分利用分级思想，采用二维嵌套链表结构，其中二维表只保存特征首字符，当命中特征后，再转入链表结构匹配其他特征字符。常用的业务识别方法包括IP/端口号关联识别技术、流量识别技术和比特位识别技术等，参考文献[8]和参考文献[9]已经对业务识别方法做了深入研究。

出于LTE网络流量大的考虑，本方案提出一种自学习方法，即根据已识别的记录学习五元组和三元组信息，并对该类记录采用超时淘汰机制，对后来的待识别记录优先匹配五元组或三元组索引信息。该方法不仅能大幅度减轻服务器压力，同时也提升了分组识别速率和识别成功率。实时DPI识别流程如图4所示。

(1)模块检测到一条xDR记录，若该报文承载协议不是HTTP，依次根据五元组、三元组、host、user-agent和比特流特征匹配特征库，直到成功识别业务转到步骤(2)。否则识别下一条记录。

(2)识别成功，判断该条记录是否已存在五元组或三元组特征，如果不存在，则学习该匹配特征，并更新特征库。

(3)查看该业务是否存在子功能特征，如果存在，则提取子功能比特识别码匹配负载比特流，成功则记录该子功能名称，否则该条记录识别结束。

3.2 特征分析/网络爬虫模块

前面提到针对不同的业务，其特征结构或策略可能都不同，因此必须对业务特征进行分析。对于自有协议的报

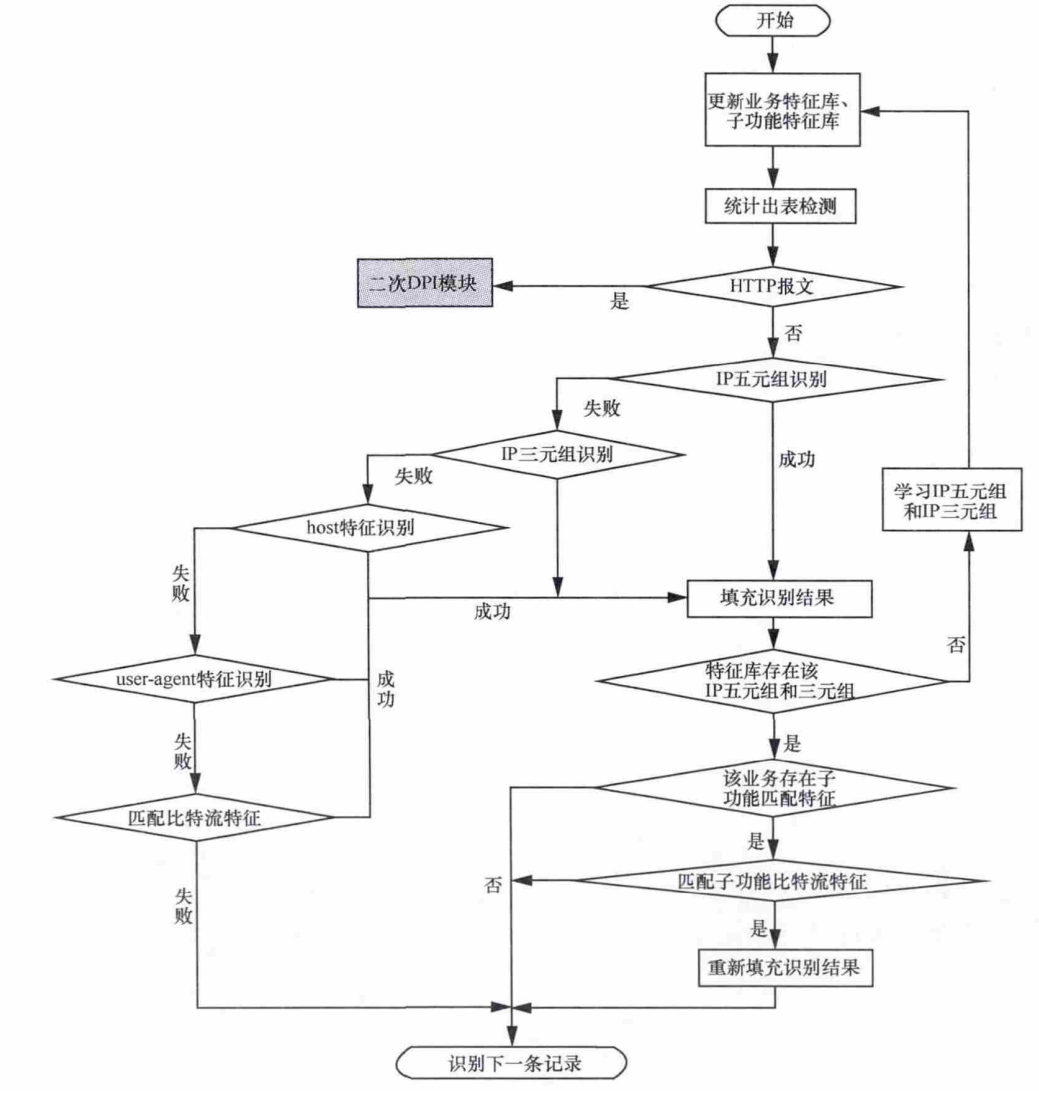

图4 实时DPI识别流程

文，运用 Wireshark 抓取分组进行分析，遍历应用所有的子功能，确保信息完整性，确定该功能的“指纹”，并将“指纹”和关联信息插入到 DPI 特征库；对 HTTP 承载的报文采用聚类爬虫及技术从 PC 端提取相关的特征信息保存到信息特征库。

聚类爬虫是一个按照既定的规则在传统 PC 互联网端自动获取网页信息的程序。它通过对内容特征进行分析，提取既定的网页元数据，为专题分析提供特定主题的数据资源[10]。爬虫程序是在 PC 端获取信息。因此，对于移动互联网端的业务请求链接，必须能够找到一种映射关系与特征库中的数据进行关联，并将这种移动互联网与 PC 互联网中信息关联的特征字符串作为 DPI 特征信息库的特征 key 值。图 5 是爬虫工作流程。

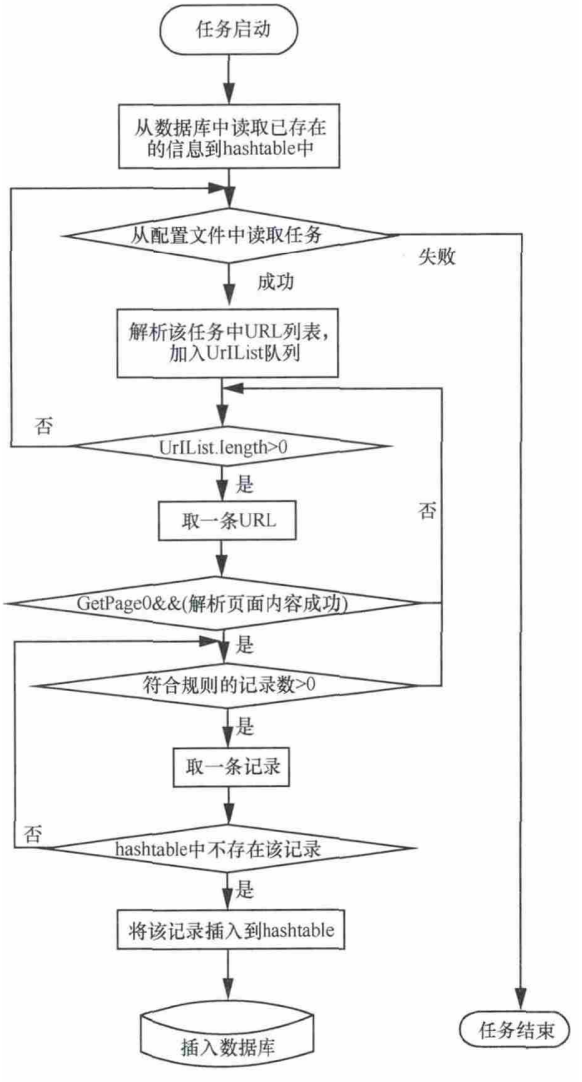

图 5 聚类爬虫工作流程

爬虫程序采用多线程多任务的设计方式，通过 UI 控制界面对多个爬虫任务进行管理，定时启动爬虫任务。每个爬虫任务负责完成某一个网站的信息采集任务。为了防止重复提取信息，爬虫任务启动时应该将已爬取的信息 key 值映射到散列表中，并且每次成功抓取一条记录都应在散列字典中记录。对那些实时更新的网站，爬虫程序需要定时自动重启策略。

3.3 DPI 业务识别模块

高效的匹配算法是衡量 DPI 识别引擎性能优劣的关键。综合比较各匹配算法性能[11~13]，本系统选择基于确定的有限状态机(deterministic finite automaton，DFA)引擎的正则表达式识别业务特征，并选择散列字典作为数据在内存中的存储容器。DFA 引擎每次匹配都能得到一个确定的状态，它不要求回溯，具有线性时间复杂度。散列算法是一种高效的信息索引算法，理论时间复杂度接近 $O(1)$。二次 DPI 识别流程如图 6 所示。

二次 DPI 模块工作流程如下。

(1)程序启动时导入特征数据库到内存中，并在系统闲时定期更新，其中 DPIdataID 作为域名字典和 DPI 信息字典关联标识。

(2)根据 host 信息查询域名字典，获取业务名称(如“AppStore”)、DPIdataID 和行为识别表达式，如果为空则登记此 host。

(3)由步骤(2)，根据行为识别表达式匹配行为关键字如“浏览”、“下载”等，若匹配失败，则记录应用类型，识别下一条 URL；若匹配成功，则查看是否存在多级信息，如果存在则循环匹配多级 DPI 查询字典，直到没有 ChildID 为止，获取该业务标识的 key(“指纹”)匹配表达式。

(4)根据 key 匹配表达式匹配 URL，如果获取 key 值成功，则根据 key 值查询此 DPIdataID 关联的信息字典，获取该记录具体信息，否则，只记录应用类型。

(5)key 值映射信息成功，则记录本次 HTTP 请求对应的信息，如“AppStore—下载 -12306 铁路客户端—版本:×××—开发商:×××”，否则只记录本次用户动作，如“AppStore-下载”。

(6)将识别到的信息与用户 IMSI 关联，并将本次记录插入应用专题数据库中。

4 结果统计

本系统将现网 S1-U 接口采集到的测试数据，经解码/

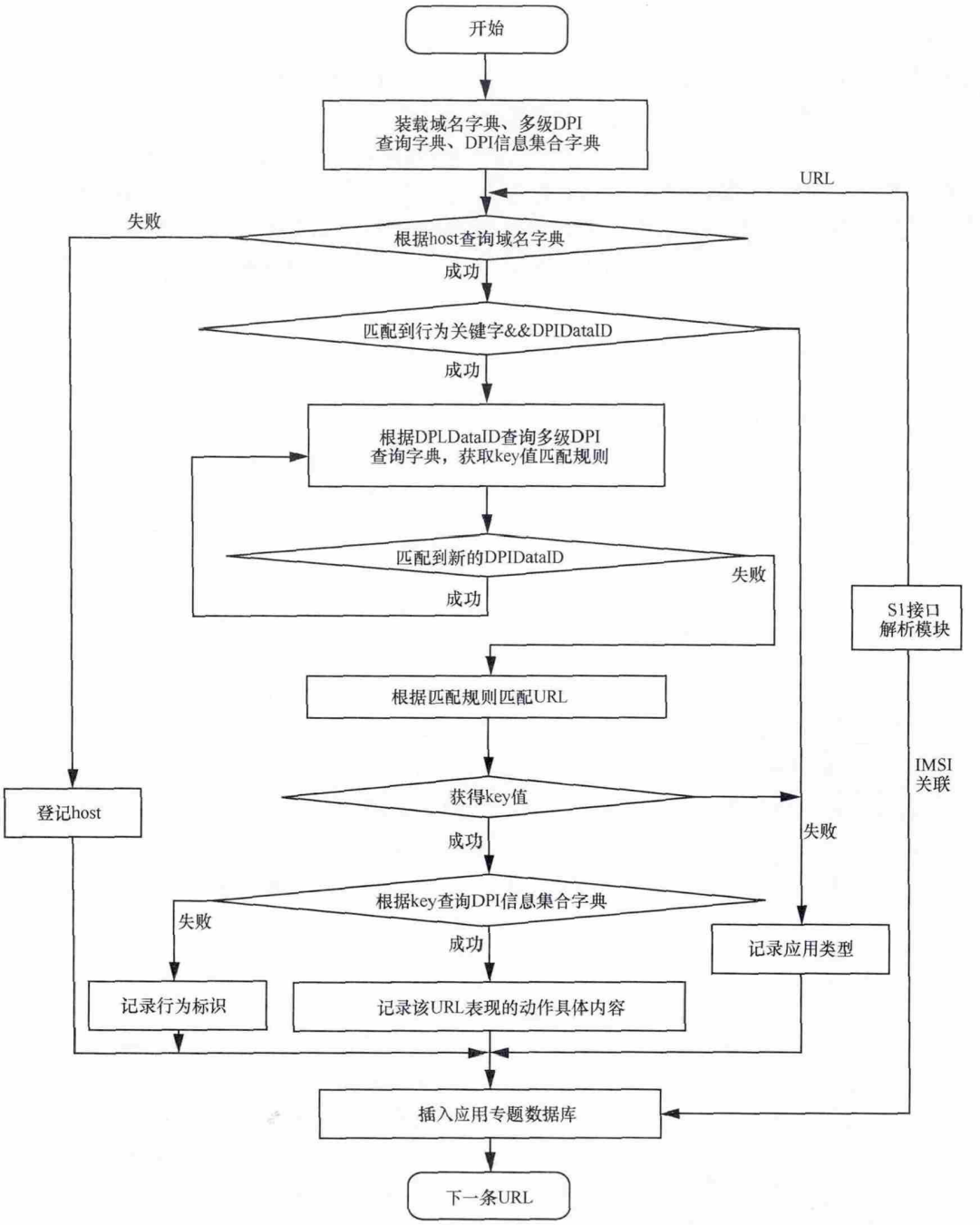

图 6　二次 DPI 识别模块流程

合成和实时 DPI 业务识别处理后，再转给二次 DPI 匹配模块进行业务内容识别，最后交付给统计模块，按不同专题或维度进行业务统计。

图 7 是业务识别结果出表的一段截图，其中 B、C 字段分别代表主类型和子类型（如主类型“1”表示应用商城，子类型“3”表示苹果 Appstore）；D、E、F、G 分别表示服务端 IP、服务端端口号、客户端 IP、客户端端口号；字段 H 和字段 I 分别表示 host 和 URL；K 列表示 user-agent 信息；L 和 M 分别表示上下行流量。

将业务面出表数据做 IMSI 关联，即可确定某用户具体的上网信息。如图 8 所示，以苹果 AppStore 为例，统计用户访问 AppStore 的详细信息。如图 9 所示，将业务归类，统

A	B	C	D	E	F	G	H	I	K	L	M
14:22:37	1	3	223.082.244.236	49300	010.042.006.073	80	a1064.phobos.app	/us/r1000/027/Pur	iTunes-iPhone/5.1	490	15708
14:22:42	1	8	211.152.118.011	52256	010.167.078.023	80	push.wandoujia.c	/v3/list?status=0	android-16-480x8	1126	1130
14:23:24	7	2	074.125.128.018	59460	010.130.018.220	443	delta.taobao.com	/favicon.ico	################	3896	13485
14:23:45	1	3	23.201.102.88	61982	010.042.006.073	80	a2.mzstatic.com	/us/r30/Purple4/v	iTunes-iPhone/5.1	531	700
14:24:47	1	3	199.7.71.72	61946	010.042.006.073	80	EVSecure-ocsp.ve	/MFYwVKADAgEAME0w	securityd (unknov	330	818
14:26:44	1	3	23.201.102.75	61914	010.151.139.209	80	a1.mzstatic.com	/us/r30/Purple6/v	iTunes-iPhone/5.1	552	230
14:27:32	7	2	042.120.180.012	49215	010.130.236.125	80	a1d.taobao.com	/recommend.htm?t=	################	866	893
14:28:40	4	1	117.135.169.19	50504	010.151.139.209	80	short.weixin.qq.c	/cgi-bin/micromsg	Android QQMail MO	507	68
14:28:45	4	1	117.135.169.19	50506	010.151.139.209	80	short.weixin.qq.c	/cgi-bin/micromsg	Android QQMail MO	603	68
14:29:52	4	1	117.135.169.19	50505	010.151.139.209	80	short.weixin.qq.c	/cgi-bin/micromsg	Android QQMail MO	925	68
14:33:47	4	15	111.013.004.021	59415	010.130.018.220	80	c.renren.com	/o.jsp?d=13854905	Apache-HttpClient	1300	658
14:34:35	1	1	61.138.219.43	38069	010.167.161.223	80	u5.mm-img.com	/rs/res1/21/2013/12/16/a914/804/32		240	818
14:34:44	1	1	221.179.8.164	34926	010.167.161.223	80	mmota.10086.cn	/downloadApp?id=3	android-16-480x8	687	342

图 7　业务识别出表

时间	IMSI	类型	频道	类别1	名称	作者	上行流量	下行流量
2013/10/11 12:45:34	4600345(	54	应用商店	APPstore	商业	唯品会-正品名牌时尚折扣网		324	11785
2013/10/11 5:56:04	460034	43	应用商店	APPstore	娱乐	风云直播 - 最全最快的网络电视	xuan zhang	520	55679
2013/10/11 9:35:37	4600389	99	应用商店	APPstore	娱乐	QQ游戏大厅	Tencent Technology (Shenzhen) Company Limited	0	0
2013/10/11 16:51:38	4600434	483	应用商店	APPstore	商业	新概念商务英语		0	0
2013/10/11 15:56:45	4600456	32	应用商店	APPstore	商业	全球商业经典 HD	Inforgence	1298	0
2013/10/11 23:47:49	4600565(	33	应用商店	APPstore	商业	新概念商务英语		236	0
2013/10/11 5:23:12	4600594	74	应用商店	APPstore	娱乐	艺术签名设计	Sensky Ltd.	321	149
2013/10/11 13:44:27	46006577	84	应用商店	APPstore	商业	搜房帮	SouFun	299	1132
2013/10/11 7:22:15	4600788	43	应用商店	APPstore	商业	搜房帮	SouFun	235	0
2013/10/11 15:51:01	4600789	32	应用商店	APPstore	娱乐	QQ游戏大厅	Tencent Technology (Shenzhen) Company Limited	1314	3427
2013/10/11 14:54:23	46007965	53	应用商店	APPstore	娱乐	艺术签名设计	Sensky Ltd.	0	0
2013/10/11 17:13:58	4600798(	54	应用商店	APPstore	娱乐	艺术签名设计	Sensky Ltd.	211	313
2013/10/11 3:56:12	4600856	39	应用商店	APPstore	商业	唯品会-正品名牌时尚折扣网		0	0
2013/10/11 14:54:23	46008789	58	应用商店	APPstore	商业	全球商业经典 HD	Inforgence	155	255

图 8　AppStore 用户访问详细信息

图 9　某一时段内全网用户上网偏好

计出某一时间段内全网用户上网偏好趋势。

5　结束语

本文提出了一种基于 LTE 网络的用户行为分析系统的解决方案，分析了智能网络趋势下用户行为感知的必要性，在传统协议栈解码的基础上，结合 DPI 技术和网络爬虫技术，实现了一种自学习功能的移动互联网用户行为感知系统。该方案弥补了传统业务识别方案的不足，大幅提升了业务识别的精确度和识别效率。经现网数据测试，本方案不仅能够对 LTE 网络用户行为进行有效的识别，而且能够针对不同的业务专题实施策略统计。本方案为完善移动互联网“智能管道”具有很高的实践价值。

参考文献

1　张海峰，张杰. TD-LTE 数据业务发展趋势. 互联网天地，2013(4)

2　Huawei SmartCare CEM 解决方案. http://www.huawei.com/cn/services/hw-u_256372.htm#.U5UM qI1h6jU, 2014-06-09

3　客户体验管理系统 CEMS. http://www.zte.com.cn/cn/solutions/anyservice/oss_bss/201212 /t2012124_372546. html, 2014-06-09

4　Paxson V. BRO: a system for detecting network intruders in real-time. Computer Networks, 1999, 31(23): 2435~2463

5　Roesch M. Snort: lightweight intrusion detection for networks. Proceedings of 13th Systems Administration Conference, Washington, USA, 1999

(下转第 120 页)

与每首被推荐歌曲之间的空间距离，继而按照这个距离对推荐歌曲进行重新排序。

参考文献

1 覃亮，王喜成. 层次分析法在制造业电子商务网站评价中的应用. 桂林电子工业学院学报，2006(1)：74~75

2 余力，刘鲁. 电子商务个性化推荐研究. 计算机集成制造系统，2004，10(10)

3 奉国和，梁晓婷. 协同过滤推荐研究综述. 图书情报工作，2011，55(16)

4 陈萌，杨成，王欢等. 交互式电视中个性化推荐系统的研究. 电视技术，2012，36(14)

5 徐淮杰，张二芬. 基于关联规则与奇异值分解的音乐推荐系统. 电子设计工程，2013，21(1)

[作者简介]

朱映波，男，博士，中国电信股份有限公司数字音乐运营中心高级工程师、总经理，在多媒体及互联网应用、电信增值业务领域有丰富的开发、管理和运营经验，主要研究方向为移动互联网音乐相关技术和业务。

刁建伟，男，中国电信股份有限公司数字音乐运营中心中级工程师、市场营销部总监，主要研究方向为互联网下音乐业务的发展。

康波，男，中国电信股份有限公司广东研究院中级经济师、客户研究经理，主要研究方向为数据挖掘、用户体验。

刘胜强，男，中国电信股份有限公司广东研究院中级经济师、消费者实验室主任，主要研究方向为用户体验、数据挖掘。

（收稿日期：2014-05-20）

（上接第 83 页）

6 3GPP TS36.401. E-UTRAN Architecture Description , 2010

7 李艳，张治中. LTE 网络 S1AP 监测方案的研究与实现. 电信科学，2013，29(1)：31~38

8 张应. 基于综合特征的 P2P 流量识别与控制系统研究. 复旦大学硕士学位论文，2009

9 Chen H, Hu Z, Ye Z, *et al*. A new model for P2P traffic identification based on DPI and DFI. Information Engineering and Computer Science, 2009(ICIECS 2009), Wuhan, China, 2009

10 刘建明. 垂直搜索引擎中的主题爬虫技术研究. 广东工业大学硕士学位论文，2013

11 Chaudhary A, Sardana A. Software based implementation methodologies for deep packet inspection. ICISA 2011, Jeju Island, Korean, 2011

12 Guo L, Wang Y, Yao Q, *et al*. A fast regular expression matching algorithm for deep packet inspection. ICITIS 2010, Beijing, China, 2010

13 Yu F, Chen Z, Diao Y, *et al*. Fast and memory-efficient regular expression matching for deep packet inspection. Proceedings of the 2006 ACM/IEEE Symposium on Architecture for Networking and Communications Systems, San Jose, California, USA , 2006

[作者简介]

王建，男，重庆邮电大学通信网与测试技术重点实验室硕士生，主要研究方向为宽带通信网测试技术。

张治中，男，重庆邮电大学通信网与测试技术重点实验室博士生导师、教授，主要研究方向为通信网测试技术、宽带信息网络、NGN 等。

骆云龙，男，重庆邮电大学通信网与测试技术重点实验室硕士生，主要研究方向为宽带通信网测试技术。

（收稿日期：2014-06-10）