doi: 10.3979/j.issn.1673-825X.2014.05.010

基于 DPI 技术 LTE-S1 接口流量识别系统的设计与实现

杨丰瑞[1],吴　辉[2],张治中[2]
(1. 重庆重邮信科(集团)股份有限公司,重庆 400065; 2. 重庆邮电大学 通信网与测试技术重点实验室,重庆 400065)

摘　要: 针对长期演进(long time evolution,LTE)网络,基于数据解码技术、爬虫技术和 DPI 技术提出了一种流量识别系统的设计与实现方案。以图书类 APP 应用流量识别为例,说明深度封包检测(deep package inspection,DPI)技术在 S1 接口流量识别系统中的设计实现过程。采用黑盒测试方法,对系统进行性能测试,通过测试结果得出结论:该流量识别系统可以实现对用户行为的监测,对细化流量经营有着重要的推广意义。

关键词: 长期演进(LTE);S1 接口;深度封包检测(DPI)
中图分类号: TN929.5;TP391　　**文献标识码:** A　　**文章编号:** 1673-825X(2014)05-0622-04

Design and realization of the traffic recognition system in LTE-S1 interface based on DPI technology

YANG Fengrui[1],WU Hui[2],ZHANG Zhizhong[2]
(1. Chongqing Chongyou Information Technology (Group) Co,Ltd, Chongqing 400065,P. R. China; 2. Key Laboratory on Communication Network and Testing Technology, Chongqing University of Posts and Telecommunications, Chongqing 40065, P. R. China)

Abstract: A scheme of design and realization of the traffic recognition system which was based on data decode technology, crawler technology and DPI technology was proposed for LTE(long time evolution) network. Taking recognizing the flow rate of book APP as an example,the design and realization of DPI(deep package inspection) was expounded. In the traffic recognition system,the data decoding technology,the crawler technology and the DPI technology were used mainly. The performance test was executed by the method of black-box test. According to the result of the test, conclusion was made that the application of this kind of DPI technology can monitor subscribers´behavior in LTE network, and it is of great significance to promote for detailed flow rate management.

Key words: long time evolution (LTE); S1 interface; deep package inspection(DPI)

0　引　言

随着移动互联网的高速发展,当今社会已经完全进入移动互联网时代。调研结果表明,电信运营商并没有因网络流量使用量的飞速增加而增多收益。一方面,从 2008 年到 2013 年,网络流量每年都在成倍增长,到 2013 年网络流量已经增长到 2008 年的 50 - 60 倍,其中数据业务流量已经占到总流量的 95%,但给运营商带来的收入不足总收入的 50%;另一方面,全球运营商每 GByte 产生的平均营收已从 5 600 美元降至 11 美元,即每 MByte 的营收仅有 0.01 美元,而中国市场的形势更加严峻。因此,电信运营商之前的

收稿日期: 2014-01-14　**修订日期:** 2014-09-25　**通讯作者:** 吴辉　751569801@qq.com
基金项目: 2013 年重庆市高校创新团队建设计划(KJTD201312);国家科技重大专项(新一代宽带无线移动通信网)(2012ZX03001021-004)
Foundation Items: The Patronage project of The ChongQing city college innovation team building plan in 2013(KJTD201312); The major special project of national science and technology (The next generation wireless mobile communication network) (2012ZX03001021-004)

粗经营模式并非长久之计[1]。随着长期演进(long time evolution，LTE)应用规模的扩大，语音业务作为最大的电路域交换业务，将会被 VoLTE(voice of LTE)取代[2]，本文提出一种对 LTE 网络流量识别的方法，来实现流量精细化管理。

1　流量识别系统构架

本系统的实现过程可分为 3 个步骤：LTE-S1 接口数据解码、第 1 次深度封包检测(deep package inspection，DPI)和第 2 次 DPI。运用爬虫技术来获取特征库，流量识别系统构架如图 1 所示。

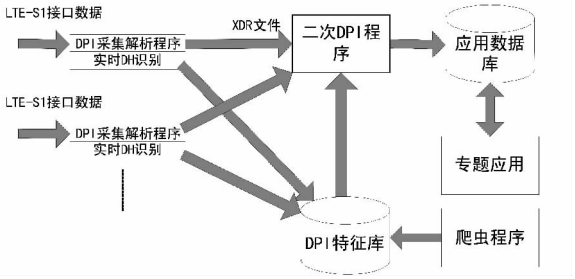

图 1　流量识别系统构架

Fig. 1 Architecture of the flow recognition system

2　第 2 次 DPI 特征库建立过程

以建立阅读应用特征库为例，说明整个特征库建立过程。

步骤 1　在数据库中，建立 MATCHRULE，BOOKINFO 和 WEBSITEID 特征表。

MATCHRULE 表用于匹配用户阅读行为以及提取所阅读书籍的 BOOKID，其包含 PREFIX，REGUX，BEHAVIOR，PREFIX_REGUX 和 WEBSITEID 5 个字段。PREFIX 字段用于存放图书介绍 URL 以及图书阅读 URL 的前缀，REGUX 字段是获取 BOOKID 号的正则表达式，BEHAVIOR 字段用于存放与 PREFIX 匹配的用户行为信息，PREFIX_REGUX 字段用于存放介绍和阅读图书 URL 的正则表达式，其与底层解码获取的 URL 匹配成功即可实现 URL 与特征库的关联，WEBSITEID 用于记录图书网站 ID 号。BOOKINFO 表用于存放网页爬虫爬取的书籍信息，包含 WEBSITEID，BOOKID(书本 ID)，AUTHOR(书本作者)，BOOKNAME(书籍名称)以及 TYPE(书籍类型)字段。WEBSITEID 表用于存放 BOOKINFO 表中取各个网站图书信息的 SQL 语句。BOOKINFO 表，MATCHRULE 表和 WEBSITEID 表分别如图 2 - 图 4 所示。

WEBSITEID	BOOKID	BOOKNAME	TYPE	AUTHOR
1	2833101	傲天神命	东方玄幻	凌云大少
1	3007730	暗手之血煞	东方玄幻	任远
1	3057947	暗影刺客	东方玄幻	梦的那个角落

图 2　BOOKINFO 表

Fig. 2 Bookinfo table

PREFIX	REGUX	WEBSITEID	BEHAVIOR	PREFIX_REGUX
http://m.qidian.com/book/showboo	(?<=bookid=)\d+	1	浏览介绍	http://m.qidian.com/book/showbook.aspx\?bookid
http://m.qidian.com/Book/BookReader.aspx?bookid	(?<=bookid=)\d+	1	阅读	http://m.qidian.com/Book/BookReader.aspx\?bookid
http://wap.cmread.com/rbc/	(?<=rbc/)\d+	2	浏览介绍	http://wap.cmread.com/rbc/
http://client.cmread.com/cmread/portalapi?	(?<=contentId=)\d+	2	阅读	http://client.cmread.com/cmread/portalapi\?
http://m.jjwxc.net/book2	(?<=book2/)\d+	3	浏览介绍	http://m.jjwxc.net/book2/\d+/[^\d]*
http://m.jjwxc.net/book2	(?<=book2/)\d+	3	阅读	http://m.jjwxc.net/book2/\d+/\d+/
http://qqreader.3g.qq.com/book/onlinereadlog?	(?<=&id=)\d+	4	阅读	http://qqreader.3g.qq.com/book/onlinereadlog\?
http://qqreader.3g.qq.com/android/book?	(?<=&id=)\d+	4	浏览介绍	http://qqreader.3g.qq.com/android/book\?
http://panda.sj.91.com/Service/NovelPay.aspx?	(?<=&id=)\d+	5	阅读	http://panda.sj.91.com/Service/NovelPay.aspx\?
http://panda3g.sj.91.com/Service	(?<=&id=)\d+	5	浏览介绍	http://panda3g.sj.91.com/Service/Api.ashx\?

图 3　MATCHRULE 表

Fig. 3　MATCHRULE table

WEBSITEID	SQL
1	select * from bookinfo where websiteid=1 and bookid=
2	select * from bookinfo where websiteid=2 and bookid=
3	select * from bookinfo where websiteid=3 and bookid=
4	select * from bookinfo where websiteid=4 and bookid=
5	select * from bookinfo where websiteid=5 and bookid=

图 4　WEBSITEID 表

Fig. 4 WEBSITEID table

步骤 2　网页爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载页面，是搜索引擎的重要组成[3]。采用爬虫程序，定时获取书籍信息，以及更新图书信息，爬取起点小说伪代码如下。

```
Protected override void TasksManager( )
{…
for ( int i = 0; i < 15; i + + )
{ for ( int j = 0; j < 27; j + + )
{ string url = " http: //all. qidian. com/book/bookStore. aspx?
ChannelId = " + typeids [i] + " &SubCategoryId = - 1&Tag =
all&Size" + " = - 1&Action = - 1& Or de rId = 6&P = " +
subType [j] + " &update = - 1&Vip = - 1&Boutique = -
1&SignStatus = - 1";
CrawleEntity ce = newCrawleEntity( url, typeids [i], subType
[j]);
manager. Push( ce) ;}}}
    private WebPage GetPage( string url) {…}
private bool GetBookMessage( string webtext) {…}
```

TasksManager( ) 实现起点小说所有图书列表 URL 的获取，GetPage( ) 实现获取 URL 对应的 HT-

ML 信息,GetBookMessage( ) 实现提取 HTML 中的小说信息,并存入数据库。

3 流量识别系统的设计与实现过程

3.1 S1 接口数据解码

S1 接口数据解码,提取各层协议的字段信息,为 DPI 处理提供基础。本文中,解码方案采用提取关键信息字段的简单解码。解码过程采用从接口协议栈底到顶的逐层解码思想,提取本层协议关键字段后,根据上层协议标识,调用相应的协议解码器,将协议数据单元(protocal data unit,PDU) 递交上层协议解码接口,以此递归直到没有上层协议[4-5]。

3.2 第 1 次 DPI

运用 DPI 技术,设备可以检查有效荷载或包头,搜索违约协议、病毒、垃圾邮件、入侵[6]或判断业务类型。第 1 次 DPI 过程是为了对及时通信、阅读、微博、导航、视频、音乐、应用商店、游戏、支付、动漫、邮箱、P2P 业务、VoIP 业务、彩信、浏览下载财经、安全杀毒和其他业务 18 个大类业务进行区分,实现步骤如下。

步骤 1 一个数据流量可以定义为一个拥有相同 5 元组的数据包的集合,即: 源端口地址、目的端口地址、源 IP 地址、目的 IP 地址以及协议类型[7]。将数据包按照 5 元组重新整合,便可得到整个数据流的内容,常见的应用层协议端口号如下: HTTP 端口号为 TCP 80,8080,8086; FTP 端口号为 TCP 21; POP3 端口号为 TCP 110; SMTP 端口号为 TCP 25; RTSP 端口号为 TCP 554; DNS 端口号为 UDP 53; WSP 端口号为 UDP 9200; WTP 端口号为 UDP 9201,9203。

步骤 2 通过手机拨测 18 种业务类型的主要应用,建立相应特征库,并分析上层协议携带的特征字段,来标识业务类型。就 HTTP 数据而言,主要是统计 HTTP 层 URL,HOST 和 IP 五元组信息,从而实现具体业务与其特征 URL 的对应。

步骤 3 将从 PDU 提取的 URL 与特征库中的 URL 正则表达式进行逐次匹配,直到匹配成功,则判断该 PDU 是与该正则表达式对应的业务类型。正则表达式作为一种匹配选择语言,在包的扫描应用当中,可以替换特定的字符串形式,之所以正则表达式能得到广泛的应用,在于其强大而灵活的表达形式[8]。

通过第 1 次 DPI 可以实现用户业务大类型的判断,然而对流量识别不够细化。

3.3 第 2 次 DPI

通过之前的方案,建立了阅读业务特征库以及获取了包含 URL 的 XDR 信息,本次操作主要实现底层解码 URL 与特征库的对应,即第 2 次 DPI 的实现,程序流图如图 5 所示。

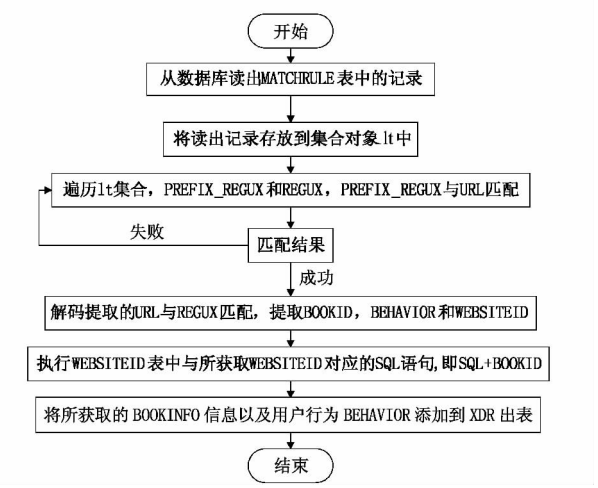

图 5 第 2 次 DPI 程序流图

Fig. 5 Flow chart of the second DPI

1) 从数据库中将 MATCHRULE 表的各条图书网站的匹配信息取出,加入到集合 *lt*,实现伪代码如下。

```
string sql = "select PREFIX_REGUX,REGUX,WEBSITEID from matchrule";
IDataReader idr = odb.ExecuteReader( sql, null); List < string [] > lt = new List < string [] >( );
while ( idr. Read( ) ) { string [] st = new string [3]; st [0] = idr [0] . ToString( ); st [1] = idr [1] . ToString( ); st [2] = idr [2] . ToString( ); lt. Add( st) ;}
```

2) 底层解码 URL 与 *lt* 集合中所有的 PREFIX_REGUX 进行遍历匹配,若匹配,则对 URL 中的 bookid 进行提取,关键实现代码如下。

```
for ( int i = 0; i < lt. Count; i + + ) { Match m = Regex. Match( url, lt [i] [0]) ; if ( m. Value ! = "") { m = Regex. Match( url, lt [i] [1]) ; string dpidataid = null; string bookid = null; if ( m. Value ! = "") { bookid = m. Value; } } }
```

3) 通过步骤 2,得到与 URL 匹配的消息所对应的 websiteid,执行 WEBSITEID 表中与获取的 websiteid 对应的 SQL 语句,从而求得 URL 所对应的图书信息以及用户行为,其关键代码如下。

```
sql = "select sql from WEBSITEID whereWEBSITEID = " + websiteid; idr = odb. ExecuteReader( sql, null) ;
```

```
while ( idr. Read ( ) ) { sql = idr [0] . ToString ( ) + bookid; idr = odb. ExecuteReader( sql, null) ; while ( idr. Read ( ) ) { bookname = idr [1] . ToString( ) ; booktype = idr [2] . ToString( ) ; author = idr [3] . ToString( ) ; break; } }
```

4　测试结果及分析

步骤 1　通过手机拨测方式，得到浏览各大图书网站的拨测数据，同时记录好拨测手机的移动设备国际辨识码(international mobile equipment identity，IMEI) **以及进行阅读应用的时间。拨测记录如图 6 所示。**

步骤 2　运行 DPI 流量识别系统，处理从 LTE－S1 接口取得的拨测数据，查看流量识别结果与拨测记录是否一致。测试结果记录如图 6 所示。

步骤 3　按照上述步骤重复，对不同数据进行反复测试，查看 DPI 结果是否正确，确保产品的功能可靠以及程序健壮。从测试情况来看，测试结果基本达到需求标准，能够实现用户阅读行为的详细记录。

拨测信息记录					测试信息记录			
拨测时间	IMEI	拨测内容	拨测结果	Start_time	M-TMSI	IMEI	Behavior	DpiInfo（Subtype：BookName：Author）
2014.1.6 9:33:23	861319020030534	手机熊猫阅读客户端浏览介绍	成功	2014.1.6 9:33:23	4294967295	861319020030534	浏览	言情:超级帅哥偶像团:夏千叶
2014.1.6 9: 35: 36	861319020030534	手机熊猫阅读客户端阅读小说	成功	2014.1.6 9: 35:	4294967295	861319020030534	阅读	言情:超级帅哥偶像团:夏千叶
2014.1.6 9:38:45	861319020030534	起点小说手机客户端浏览介绍	成功	2014.1.6 9:38:45	4294967295	861319020030534	浏览	西方奇幻：龙魂军团：莫鱼
2014.1.6 9: 39: 37	861319020030534	起点小说手机客户端阅读小说	成功	2014.1.6 9: 39:	4294967295	861319020030534	阅读	西方奇幻：龙魂军团：莫鱼
2014.1.6 9:50:22	861319020030534	QQ阅读手机客户浏览介绍	成功	2014.1.6 9:50:22	4294967295	861319020030534	浏览	鬼谷子的局6:中国历史:寒川子
2014.1.6 9: 51: 39	861319020030534	QQ阅读手机客户端阅读小说	成功	2014.1.6 9: 51:	4294967295	861319020030534	阅读	鬼谷子的局6:中国历史:寒川子
2014.1.6 9:55:42	861319020030534	晋江文学手机阅读客户端浏览介	成功	2014.1.6 9:55:42	4294967295	861319020030534	浏览	科幻：残梦：PM迷宫
2014.1.6 9: 57: 16	861319020030534	晋江文学手机阅读客户端阅读小	成功	2014.1.6 9: 57:	4294967295	861319020030534	阅读	科幻：残梦：PM迷宫
2014.1.6 9:58:10	861319020030534	CMRead手机客户端浏览介绍	成功	2014.1.6 9:58:10	4294967295	861319020030534	浏览	玄幻奇幻:霸刀凶猛:鹰刀
2014.1.6 9: 59: 48	861319020030534	CMRead手机客户端阅读小说	成功	2014.1.6 9: 59:	4294967295	861319020030534	阅读	玄幻奇幻:霸刀凶猛:鹰刀

图 6　DPI 结果表

Fig. 6 Results table of DPI

5　总　结

本论文进行的深度包检测，是基于 LTE－S1 接口数据解码实现的，通过解码技术与 DPI 技术的结合，实现了 LTE 网络流量识别的功能，对运营商流量进行差异化管理有重要的辅助作用，其他类型业务的 DPI 与阅读业务的 DPI 方法不尽相同，不再进行过多的论述。目前整个项目仍在进一步的研发和测试当中，预计后期会投入到各地运营商的使用。

参考文献：

[1] **李明. 华为解读流量经营：用户级控制成运营商胜负手** [EB/OL]. (2013-01-18) [2013-11-07]. http://www.educity.cn/it/538917.html.
LI Ming. The HUAWEI explain traffic management: The control of the users' level become the key to success [EB/OL]. (2013-01-18) [2013-11-07]. http://www.educity.cn/it/538917.html.

[2] ALBERTO Teković, IVAN Pešut, ZLATAN Morié. Voice Service in an LTE Network-CSFB [C]//ELMAR, 2013 55th International Symposium. USA: IEEE Press, 2013: 251-254.

[3] **周立柱，林玲. 聚焦爬虫技术研究综述** [J] **计算应用**，2005，25(9)：1965-1969.
ZHOU Lizhu, LIN Ling. Focusing on the research summary of crawler technology [J]. Computer Applications, 2005, 25(9): 1965-1969.

[4] **鲍宁海. TD-SCDMA 网络测试仪—IuPS 与 Gn 接口协议解码软件的设计与实现** [D]. **重庆：重庆邮电大学**，2007.
BAO Ninghai. TD-SCDMA Network Testing Instrument—On the Designand Implement of Protocol Decoding Software at IuPS & Gn Interfaces [D]. Chongqing: Chongqing University of Posts and Telecommunications, 2007.

[5] **魏辉，张治中. TD-SCDMA 网络测试仪中 SCCP 协议解码及上层 PDU 获取方案** [J]. **重庆邮电大学学报：自然科学版**，2007，19(1)：47-52.
WEI Hui, ZHANG Zhizhong. Decoding and upper layer PDU getting of SCCP in TD-SCDMA network analyzer [J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2007, 19(1): 47-52.

[6] ANAT Bremler-Barr, SHI, RIT Tzur David, David Hay, Yaron Koral. Decompression-Free Inspection: DPI for Shared Dictionary Compression over HTTP [C]//INFOCOM, 2012 Proceedings IEEE. USA: IEEE Press, 2012: 1987-1995.

(下转第 726 页)

2008,24(6):80-83.
XU Xibao, YANG Guishan, ZHANG Jianming. Scenario Modeling of Urban Land Use Changes in Lanzhou with ANN-CA [J]. Geography and Geo-Information Science, 2008,24(6): 80-83.

[9] 井长青,张永福,杨晓东. 耦合神经网络与元胞自动机的城市土地利用动态演化模型 [J]. 干旱区研究,2010,27(6):854-860.
JING Changqing, ZHANG Yongfu, YANG Xiaodong. Approach of Dynamic Evolution Model of Urban Lang Use Based on the Integration of ANN and CA [J]. Arid Zone Research,2010,27(6): 854-860.

[10] 乔纪纲,邹春洋. 基于神经网络的元胞自动机与土地利用演化模拟—以广州是白云区为例 [J]. 测绘与空间地理信息,2012,35(6):17-20.
QIAO Jigang, ZOU Chunyang. The Simulation of Cell Automaton and Land Use Evolution Based on Neural Network: Taking Baiyun District of GuangZhou as a Case Study [J]. Geomatics & Spatial Information Technology, 2012,35(6):17-20.

[11] 曹敏,史照良. 基于遗传神经网络获取元胞自动机的转换规则 [J]. 测绘通报,2010,24(3):24-27.
CAO Min, SHI Zhaoliang. Transition Rule for GANN-CA [J]. Bulletin of Surveying and Mapping, 2010,24(3): 24-27.

[12] 白新萍. 基于-模型的滨海新区土地利用预测 [J]. 安徽农业科学,2011,39(12):7321-7324.
BAI Xinping. Forecast of Land Use Based on ANN-CA in Tianjing New Coastal Area [J]. Journal of Anhui Agri. Sci,2011,39(12): 7321-7324.

[13] 黎夏,叶嘉安. 地理模拟系统: 元胞自动机与多智能体 [M]. 北京: 科学出版社,2007.
LI Xia, YCH Anthony Gar-On. Geographical simulation system: cellular automata and Multi-Agent [M]. Beijing: science press,2007.

[14] WHITE H. Commentionist nonparametnc regression : Multilayer feed forward networks can learn arbitrary mapping [J]. Neural Networks,1990,3(6):47-51.

作者简介:

刘明皓(1970-),男,湖南安乡人,博士。主要研究方向为地理信息系统应用研究。E-mail: liumh@ cqupt. edu. cn。

李东鸿(1988-),女,陕西人,硕士研究生。主要研究方向为计算机应用技术。

(编辑: 魏琴芳)

(上接第625页)

[7] SAMRUAY Kaoprakhon, VASAKA Visoottiviseth. Classification of Audio and Video Trafficover HTTP Protocol [C]//Communications and Information Technology. 2009. ISCIT 2009. 9th International Symposium on. USA: IEEE Press,2009: 1534-1539.

[8] YU Fa, CHEN Zhifeng, DIAO Yanlei. Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection [C]//Architecture for Networking and Communications systems, 2006, ANCS 2006, ACM/IEEE Symposium on. USA: IEEE Press,2006: 93-102.

作者简介:

杨丰瑞(1963-),重庆人,男,教授,博士,主要研究方向为通信新技术应用。E-mail: yangfengrui@ cqcyit. com。

吴 辉(1989-),湖南衡阳人,男,硕士生,主要研究通信网测试以及通信协议。E-mail: 751569801@ qq. com。

张治中(1972-),湖北人,男,博士生导师。主要研究方向为第三代移动通信测试技术、宽带信息网络、NGN 网络等。E-mail: zhangzz@ cqupt. edu. cn。

(编辑: 田海江)