



采用大数据技术的移动 DPI 关联算法探索及实现

刘孝颂, 马怡安

(中国电信股份有限公司上海研究院, 上海 200122)

摘要: 针对近年来 4G 渗透率不断提高所带来的用户质疑流量使用情况的问题, 通过深入的研究提出了一种利用移动 DPI 的详细话单与计费话单关联的算法, 采用大数据平台以及相关处理技术对算法进行了实现, 并展示了现网运行评估的效果。

关键词: 大数据; 计费话单; 深度分组检测; 数据流; LTE; eHRPD

中图分类号: TN915

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017343

Research and implementation of mobile DPI association algorithm with big data technology

LIU Xiaosong, MA Yi'an

Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: In order to solve the problem of traffic usage caused by the increasing penetration rate of 4G in recent years, an algorithm that uses mobile DPI detailed bill and billing bill was proposed, and big data platform was adopted. The related processing technology implements the algorithm and shows the effect of the current network operation evaluation.

Key words: big data, data detail report, deep packet inspection, data flow, LTE, eHRPD

1 引言

近年来, 移动通信技术的快速发展, 使得运营商对网络的升级速度也大大加快。用户一方面感受到了极速上网带来的便捷, 另一方面也开始关心自己的流量、流向。根据艾媒咨询的调查报告, 60.6%的 iOS 用户和 53.3%的 Android 用户怀疑移动运营商存在偷流量行为, 且用户认为 App 恶意偷流量。对于运营商而言, 随之而来的是大量的投诉或查询。传统上处理该类型投诉问题,

运营商只能依赖计费话单来应对用户投诉, 由于计费话单并没有详细的流量使用情况, 且时间粒度也较大, 无法让用户得到满意的答复。

对此, 运营商迫切需要通过移动 DPI (深度分组解析) 用户详细话单应对投诉。图 1 为运营商 4G 网络通用架构, 从图 1 可以发现, DPI 可以采集到用户使用的流量及对应的详细流向情况, 计费网关则能统计颗粒度较粗的 DDR (计费话单), 若能把每个用户的每条计费话单中包含的详细话单在 DPI 中找到, 并与计费话单关联起来,

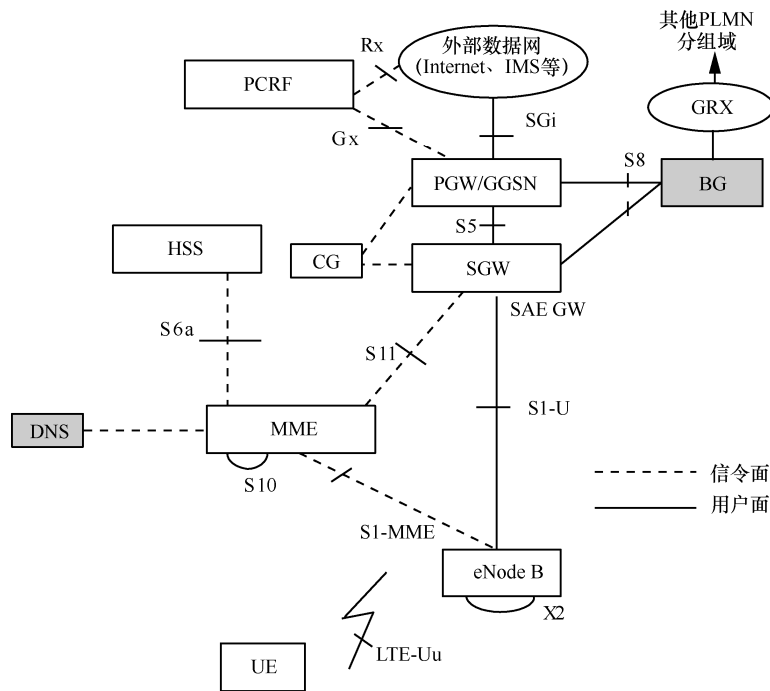


图1 运营商4G网络通用架构

则可以比较充分地应对用户投诉并告知用户其详细的流量使用情况。

但在实际操作过程中,存在两个关键问题:如何设计出一种合理的关联匹配算法;如何有效地在现网中处理海量的数据。本文将先从DDR和DPI详细话单产生的原理出发,通过数据寻找计费话单和详单关联中的问题,从而最终提出关联匹配算法或提出提升关联准确率的建议或措施。

2 流量匹配关键问题

2.1 关联衡量标准

对于客服或用户而言,最关心的是自己流量使用流向情况,如某一时间段的流量轨迹。故最直接的关联算法就是在一天时间内,针对每条计费话单的时间范围,找与之对应的DPI话单,关联完成后,统计关联后计费话单和DPI话单的流量误差并作为评判标准。若误差较大,则说明系统或关联算法等存在问题;反之则说明计费话单和DPI流量匹配率较高,可以以此应对用户投诉或供用户查询。

衡量标准1:话单级匹配率,即计费话单流量误差满足要求的比例。计算式如下:

$$P_r = \frac{\sum_i N_i}{N_{\text{DDR}}} \quad (1)$$

$$N_i = \begin{cases} 1, & |error_i| \leq \delta \\ 0, & \text{其他} \end{cases} \quad (2)$$

$$error_i = \frac{F_{\text{DDR}} - F_{\text{DPI}}}{F_{\text{DDR}}} \quad (3)$$

其中, P_r 为话单级匹配率; N_{DDR} 表示计费话单总数量; N_i 表示当第 i 条DDR的流量误差 $error_i$ 的绝对值小于一个阈值时 N_i 取值为1,否则为0;流量误差 $error_i$ 表示单条DDR流量和阈值相关联的DPI话单流量的误差率;最后, δ 是一个可设置的误差容忍度阈值,本文将其初始化为20%。

衡量标准1的意义在于展示了一天所有单条DDR的整体匹配情况,颗粒度较细,但无法反映单一用户的情况。

衡量标准2:单用户级匹配率,即一天时间内用户的DPI总流量和DDR总流量的流量误差。计算式如下:

$$P_u = \frac{\sum_k N_{m,k}}{N_{m,DDR}} \quad (4)$$

$$N_{m,k} = \begin{cases} 1, & |error_{m,k}| \leq \delta \\ 0, & \text{其他} \end{cases} \quad (5)$$

其中, P_u 为用户级匹配率; $N_{m,DDR}$ 为某用户 m 单日 DDR 总数; $N_{m,k}$ 表示用户 m 第 k 条 DDR 的流量误差 $error_k$ 的绝对值小于一个阈值时 $N_{m,k}$ 取值为 1, 否则为 0; $error_k$ 的计算式同式 (3)。

衡量标准 2 的意义在于展示了所有用户一天的计费话单的整体匹配情况, 能直观地显示用户整体的流量匹配误差, 但无法定位到单条话单的匹配情况。

本文的目标之一就是根据现网的情况, 设计出一套能合适、合理地应用于现网的计费话单与 DPI 话单关联的算法, 并以上述两个衡量标准作为算法优劣的评估标准。

2.2 问题及解决方案

2.2.1 现有统计粒度差异

DPI 的生成频度远高于 DDR, 约为 100:1 的关系。4G 话单按照 1 h 或 50 MB 进行分割, 由于这种话单切割规则, 在进行匹配的时候, 会出现 DDR 统计周期内的流量匹配与 DPI 的流量正好交错, 造成匹配失败。如果将统计周期由目前的 DDR 颗粒度进行放大, 则提高匹配率以进行验证。计费话单与详单时间切片示意如图 2 所示。

2.2.2 小流量话单的权重影响

小流量话单对于匹配率的权重影响较大,

DDR/DPI 在生成过程中的差异、造成的流量差异, 在小流量话单匹配过程中非常敏感, 对匹配率的影响权重较大。

情况 1

- DDR flow is: 82215
- DPI flow is: 65452 including 6 records
- DDR > DPI, error is: 20.3892233777

情况 2

- DDR flow is: 626215
- DPI flow is: 559864 including 113 records
- error is: 10.595562227

观察上述两种情况, 情况 1 中, 计费话单流量为 82 215 byte, 而详单流量为 65 452 byte, 两者相差 16 763 byte, 但计费话单流量误差为 20.39%, 根据之前的评判标准, 该条话单是影响匹配率的“问题话单”。但观察情况 2 可以发现, 计费话单流量为 626 215 byte, 而详单流量为 559 864 byte, 两者相差 66 349 byte, 计费话单流量误差为 10.60%, 根据之前的评判标准, 该条话单是满足要求的。不过很明显, 情况 2 对用户更敏感, 本文应该更关注情况 2 的原因, 而非情况 1。

从图 3 可以发现, 计费话单中小流量话单占比很大。如 10 KB 以下的小流量话单占比超过 26%, 但其流量占比却仅有 0.034%。而超过 1 MB 的计费话单占比不到 20%, 却拥有 85%左右的流量占比。

通过上述分析可以得出一个结论, 即应该适当降低对小流量计费话单的匹配率阈值, 而较大

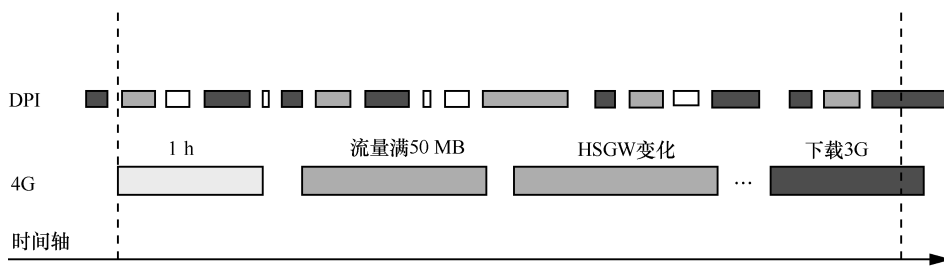


图2 计费话单与详单时间切片示意

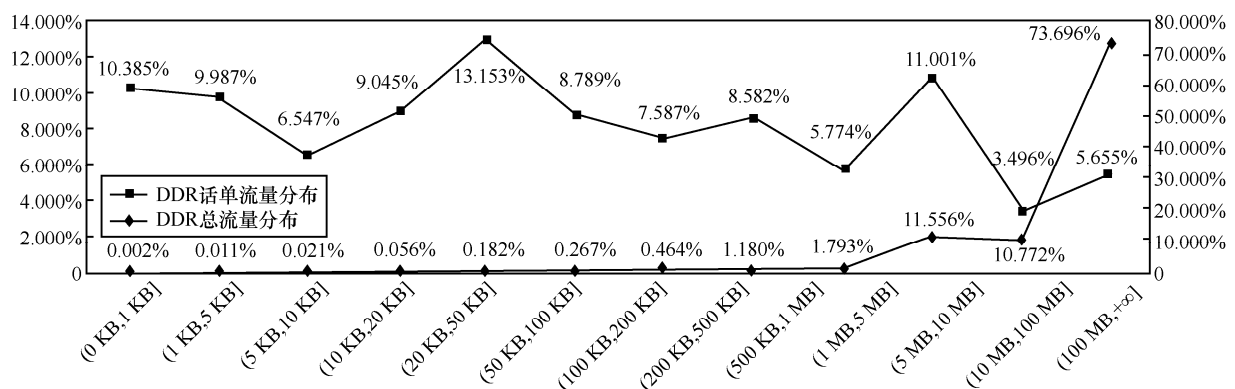


图3 计费话单条数及流量占比分布

流量的计费话单匹配率阈值无需放宽，因为无论从用户敏感度还是流量占比都说明了小流量计费话单并非很重要。

对此，本文提出了一个新的流量匹配标准，见表1。

表1 新的流量匹配标准

序号	DDR 流量分类条件	匹配标准
1	小于 100 KB	存在 DPI 即匹配，≠100%
2	100~200 KB	流量误差 100%之内，即[-100%,100%]
3	200~500 KB	流量误差 50%之内，即[-50%,50%]
4	大于 500 KB	流量误差 20%之内，即[-20%,20%]

2.2.3 计费话单边界影响

由于统计标准对时间的要求是每天 0:00—24:00，所以根据 DDR 与 DPI 话单产生规则，必定有部分 DPI 话单由于时间的切分而无法统计。

对此，本文需要根据 DDR 的时间跨度，对不同的 DDR 分别进行分析。计费话单边界影响分类如图4所示。

表2展示的某天计费话单中各类型话单的总条数和总流量情况。本文选取A类话单和C类话单进行关联匹配。因为DPI话单的时间范围为0:00—24:00，而B类话单和D类话单可以算作后一天的A类话单，故不应该加入当天的统计。

表2 不同类型计费话单条数及流量

话单	总条数/条	总流量/byte
A 类型话单	649 698	22 047 540 059 021
B 类型话单	1 984	1 243 607 026 411
C 类型话单	60 747 368	133 582 486 154 002
D 类型话单	368 670	5 988 048 615 736
总计	61 767 720	162 861 681 855 170

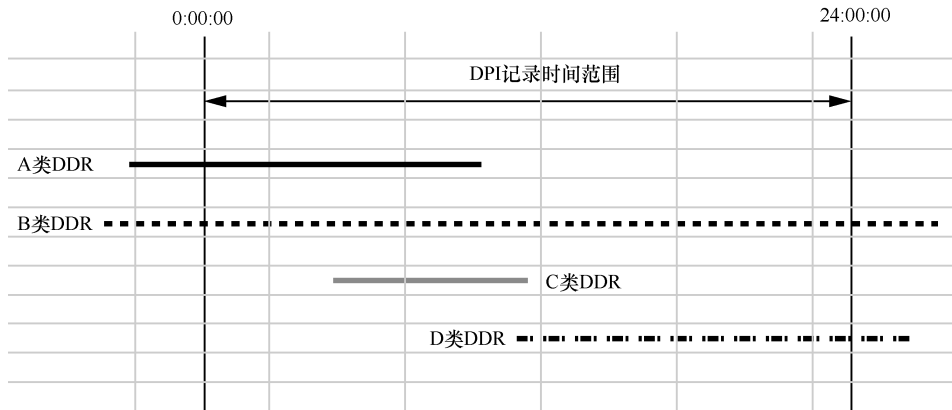


图4 计费话单边界影响分类

3 采用大数据技术的算法设计与实现

3.1 海量数据处理技术

以 300 万用户规模一天产生的数据为例。计费话单在合并前大约有 9 700 万条,合并后大约有 7 100 万条,290 TB 的总流量。详单大约有 127 亿条,310 TB 的总流量。在处理过程中势必要用到大数据处理技术。

本文实验验证环境拥有 40 个节点,总计 3 TB 内存,720Vcore,480 TB 存储空间的集群。在实现过程中本文使用了 MapReduce 和 Spark 两种技术进行处理。其中 MapReduce 处理一天的数据需要总计 19 h,而 Spark 处理一天的数据仅需 45 min (使用 1.5 TB 内存,300Vcore)。很明显,Spark 技术相比于 MapReduce 技术,在单点数据倾斜、中间结果缓存等方面拥有更大的技术优势。最终体现在对海量数据的处理效率上。

3.2 优化后的匹配算法

针对所有的计费话单,先筛选 4G 的计费话单,随后对计费话单进行预处理。预处理规则为将同一号码时间存在交叉、包含、连接关系的记录合并。随后进行话单级匹配关联:将 DPI 与 DDR 进行记录级关联,关联条件为 DPI 起始时间在 DDR 起止时间之间(一条 DDR 对应多条 DPI,一条 DPI 只对应唯一一条 DDR)。最后按 DDR 汇总:按照 DDR 的话单 ID 对关联后的记录进行聚合操作,并进行话单级和用户级的统计操作。

具体的计算规则如下。

(1) 统计 DPI 话单预处理前后总条数、总流量以及 DDR 预处理前后总条数、总流量。

(2) 针对每条 DDR 的总流量设置分段的匹配度标准,假设该条 DDR 流量为 F_1 ,与之关联的多条 DPI 流量总和为 F_2 , $(F_1 - F_2)/F_1$ 为匹配度。当 $F_1 \leq 100$ KB,则不考虑匹配度的值,认为这些话单均符合要求并分别记录条数;当 $F_1 > 100$ KB & $F_1 \leq 200$ KB,则认为匹配度在 $[-100\%, 100\%]$

内均符合要求,分别记录条数;当 $F_1 > 200$ KB & $F_1 \leq 500$ KB,则认为匹配度在 $[-50\%, 50\%]$ 内均符合要求,分别记录条数;当 $F_1 > 500$ KB,则认为匹配度在 $[-20\%, 20\%]$ 内均符合要求,分别记录条数。

(3) 统计剩余的 DDR 和 DPI 话单条数与流量。

(4) 统计每个用户当天的 DDR 总流量与 DPI 总流量。

(5) 统计 DDR 和 DPI 话单中用户总数、关联的总用户数,并针对关联的总用户数,进行流量误差分布统计。即对每个用户,假设该用户 DDR 一天总流量为 F_1 ,与之关联的多条 DPI 话单流量总和为 F_2 , $(F_1 - F_2)/F_1$ 为流量误差。输出为流量误差的用户数情况分布。

3.3 优化前后对比

本文选取了直接匹配和使用优化后算法匹配的前后共 7 天的数据进行评估,评估效果如图 5 所示。可以发现在对 DDR 进行预处理、筛选了小流量话单并考虑边界影响等因素后,无论是话单级匹配率还是用户级匹配率均得到明显提升。话单级匹配率稳定在 90% 上下,而用户级匹配率接近 95%。后者在实际的应用中价值更明显,因为大部分用户一般对某日的总流量表示异议,使用优化后的匹配算法可以解释将近 95% 用户的实际问题,之前的困境得到了大大的改善。

本算法已应用于某运营商客服系统,当投诉用户对流量使用情况有异议后,客服人员可以使用该用户手机号码进行查询(需用户授权),并将查询结果展示给用户,以此解决用户疑问。

4 结束语

本文针对运营商现网运营中关于流量使用的投诉问题,在深入研究计费话单和 DPI 话单生成的机制后,通过分析得到了一种新的计费话单和详单的关联匹配算法,并设计开发了相应的稽核验证程序,通过现网数据进行验证,取到良好的效果。

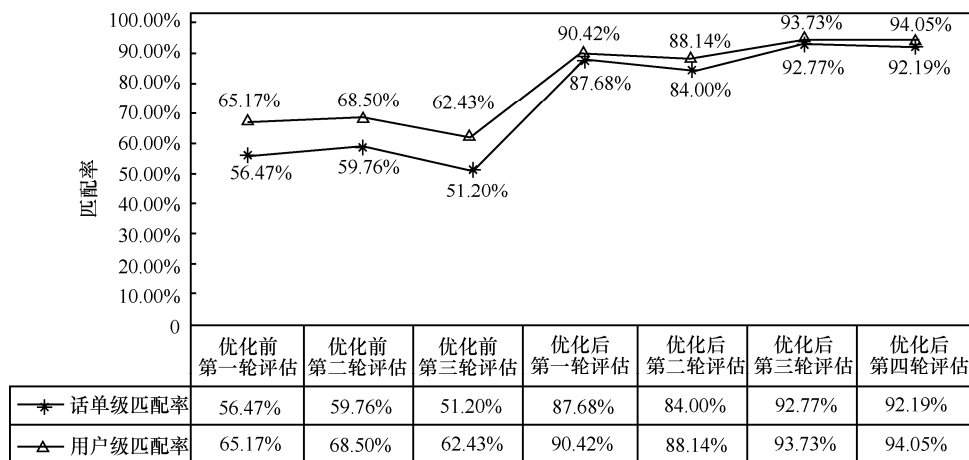


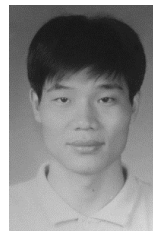
图5 算法优化前后话单级与用户级匹配率

大数据处理结合数据挖掘,可在将来的研究中,了解用户使用流量的规律,并在用户流量异常的时候做到提前提醒和告警,或者针对用户流量的使用情况给用户推荐更合适的套餐等。

参考文献:

- [1] POIKSELKA M, MAYER G, KHARTABIL H, 等. IMS: 移动领域的 IP 多媒体概念和服务[M]. 赵鹏, 周胜, 望玉梅, 译. 北京: 机械工业出版社, 2005.
POIKSELKA M, MAYER G, KHARTABIL H, et al. IMS: IP Multimedia Concepts and services in the mobile field[M]. Translated by ZHAO P, ZHOU S, WANG Y M. Beijing: China Machine Press, 2005.
- [2] 胡乐明, 曹磊, 陈洁. IMS 技术原理及应用[M]. 北京: 电子工业出版社, 2006.
HU L M, CAO L, CHEN J. The principle and application of IMS technology[M]. Beijing: Publishing House of Electronics Industry, 2006.
- [3] 3GPP. Telecommunication management; charging management; charging data records (CDR) file format and transfer: TS32.297[S].2012.
- [4] 3GPP. Telecommunication management; Charging management; charging data record (CDR) parameter description: TS32.298[S]. 2012.
- [5] 3GPP. Telecommunication management; charging management; charging architecture and principles: TS32.240[S]. 2012.
- [6] 魏永, 周云峰, 郭利超. OpenDPI 报文识别分析[J]. 计算机工程, 2011(S1): 98-100.
WEI Y, ZHOU Y F, GUO L C. Analysis of OpenDPI message recognition[J]. Computer Engineering, 2011(S1): 98-100.
- [7] 黄晓武. 基于 DPI 技术的网络流控策略[J]. 电脑知识与技术, 2011, 07(2X):1260-1261.
HUANG X W. Network flow control strategy based on DPI technology[J]. Computer Knowledge and Technology, 2011, 07(2X):1260-1261.
- [8] 于娟, 袁春蕾. 网络流量实时监控分析系统的设计与实现[J]. 物联网技术, 2013(2):71-73.
YU J, YUAN C L. Design and implementation of real-time network traffic monitoring and analysis system[J]. Internet of Things Technologies, 2013(2):71-73.
- [9] 邓博. 基于数据挖掘技术构建电信 4G 客户预测模型的研究[D]. 兰州: 兰州大学, 2015.
DENG B. Research on the construction of telecom 4G customer prediction model based on data mining technology[D]. Lanzhou: Lanzhou University, 2015.

[作者简介]



刘孝颂 (1977-), 男, 中国电信股份有限公司上海研究院网络数据系统架构师、中级工程师, 主要从事网络数据分析/挖掘工作以及基于网络数据的应用方面的工作, 主要研究方向为 LTE 网络数据采集方法以及各接口之间的信令消息关联方法、LTE 网络控制面信令消息解读分析等。



马怡安 (1987-), 男, 中国电信股份有限公司上海研究院中级工程师、大数据智慧运营研发, 主要从事大数据架构下数据分析及挖掘以及基于网络数据自主开发各类应用和质量检测工具等方面的工作。