



专题:大数据

基于大数据的 IPTV 精准营销方案研究与应用

于洪涌, 闻剑峰, 蔡鑫, 邱晨旭

(中国电信股份有限公司上海研究院, 上海 200122)

摘要:分析了大数据技术在 IPTV 精准营销中的应用,因地制宜地制定了“Hadoop 大数据平台+爬虫技术+建模工具”的框架方案,并对方案中的关键点和整个数据分析建模过程在现网中的应用进行了深入分析。在 IPTV 精准营销中取得了较好的效果,具有较高的应用价值。

关键词:大数据技术;数据建模;IPTV;精准营销

中图分类号:F274

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016316

Research and application of IPTV precision marketing based on big data

YU Hongyong, WEN Jianfeng, CAI Xin, QIU Chenxu

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

Abstract: The big data technology in IPTV precision marketing was analyzed. According to local conditions, the framework solution of “Hadoop big data platform + Web crawler + modeling tool” was developed. The key points of the solution and the whole data modeling process were analyzed. Good results were achieved in IPTV precision marketing, and higher application value was gained.

Key words: big data technology, data modeling, IPTV, precision marketing

1 引言

随着 2015 年国务院“65 号文”全面开放运营商进入 IPTV 领域和运营商网络升级, IPTV 迎来一个发展良机。来自流媒体网的数据, 2016 年 1-6 月, IPTV 用户净增 1 991.7 万户, 总数达到 6 581.2 万户。然而, 在运营商业务体系中, IPTV 主要起到宽带业务填充和增加用户黏性的作用, 直接收入贡献不突出。IPTV 互联网特性提供了点播等不同于传统电视的业务模式, 为 IPTV 业务增收另辟蹊径。同时, 随着视频业务在网络流量中比重日益增加以及政策放开带来的竞争, 发展新用户也是 IPTV 业务当前的发展重点。

在当前 IPTV 业务营销中, 最大问题是准确定位“向谁”营销“什么业务”, 传统方式指向性不够, 导致营销效率低下, 且容易引起用户反感。另一方面, 电信运营商拥有丰富的用户数据, 从数据中掘金, 是当前大势所趋。现在数据具有体量大、类型多样、速度快、价值密度低 4 个特征, 为适应新形式下的数据挖掘, 大数据技术应运而生。

2 需求分析

在 IPTV 业务实际运营中, 发现存在以下问题。

(1) 订购用户少

以某运营商省公司 2016 年 5 月数据为例, 增值业务



用户数仅占全部 IPTV 用户数的 10%左右,增值业务收入占 IPTV 业务收入的 1/4。

(2)基于用户惰性行为的分析

增值业务订购有点播、包月、包年等不同时效的产品包,从用户行为看,少用户不停点播不同内容的,应该更多地发展包年用户。

(3)不同内容产品包的用户差距明显

以某运营商省公司的数据为例,订购最多的“影院高清包年”用户是“英超高清包年”的上百倍。

(4)高清拉动

随着高清电视的普及和带宽的提升,用户对高清视频需求增长明显。

(5)积分订购和账单订购用户退订行为差距大

积分订购用户在业务到期后退订明显,账单订购用户退订相对平缓。

针对以上分析,发展增值业务用户需要考虑:用户的视频需求,包括观看时间、内容、是否高清敏感等;用户的支付习惯,包括消费敏感度、积分支付行为等。对于发展新用户,首先在宽带用户中发展,由于目标用户还没有使用 IPTV,应基于运营商 ODS(operating data store,运营数据仓储)数据和网络 DPI(deep packet inspection,深度分组检测)数据对用户进行分析,向有视频倾向的宽带用户推荐 IPTV 业务,同时推荐用户感兴趣的增值业务。

3 方案框架

3.1 基于 Hadoop 大数据平台实现数据汇聚

本文方案用到的数据比较复杂:数据体量大,仅 IPTV 业务数据就包括十几个表,每天大于 5 GB 的裸数据(某运营商省公司 IPTV 业务数据),而 DPI 数据更是体量大、价值密度低;数据类型多,包括 ODS 传统数据库数据、IPTV 日志类业务数据、DPI 数据和网上爬取的数据等;不同类型的和来源的数据需要进行用户和格式的统一;涉及的数据大多来自现网数据,包括 IPTV 业务数据、用户套餐信息等,为避免影响现网业务,有必要新建数据平台进行数据保存处理。

采用 Hadoop 技术搭建大数据平台,集群规模为“2+16+1”(名称节点+数据节点+接口机),50 TB 存储空间(能满足同时开展 3 个省公司的 IPTV 大数据分析的数据存储),部署了 Sqoop、Flume 等数据同步工具,Big、Hive、Impala、Spark 等数据查询分析工具。

3.2 基于 CRISP-DM 的数据挖掘流程

CRISP-DM (cross-industry standard process for data mining,跨行业数据挖掘标准流程)提供了一个数据挖掘生命周期的全面评述,在数据挖掘中被广泛采用。其流程如图 1 所示。

CRISP-DM 模型流程包括图 1 中的 6 个步骤^[1],说明如下。

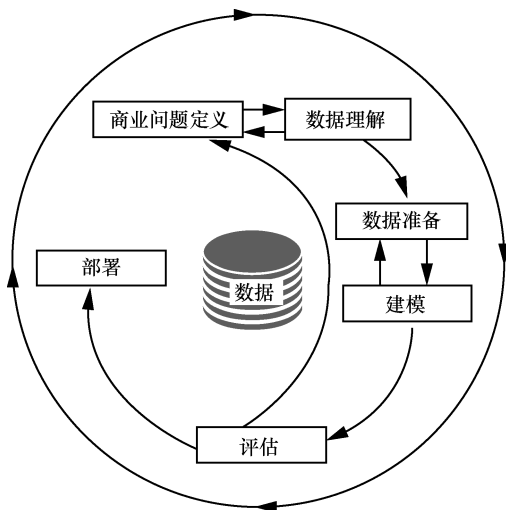


图 1 CRISP-DM 模型处理流程示意

(1)商业问题定义

主要从项目目标和业务角度理解需求,并制定初步计划;为发展 IPTV 增值业务和 IPTV 新用户提供数据挖掘服务,精准定位需求用户,其中发展 IPTV 增值业务包括发展增值业务新用户、增值业务升级(点播/包月转包年)、业务升级(标清转高清/4K(即 4K 分辨率,ultra HD 标准),宽带升级)等;发展新用户主要在宽带用户中发展 IPTV 用户。

(2)数据理解

从数据收集开始进行数据探索,发现数据内部属性;需要的数据包括 IPTV 业务数据、用户电信画像数据(用户套餐信息、家庭成员结构、支付习惯等)、用户视频画像数据等。

(3)数据准备

数据准备阶段包括从未处理数据中构造最终数据集的所有活动。基于业务目标,对汇集的数据进行分析,缩小数据范围,并对不同来源数据进行对接统一;Hive 提供的 Hive SQL 非常适合熟悉 SQL 语言的人使用,本项目使用 Hive 的 UDF(user defined function,用户定义函数)进行数据提取,为数据建模提供宽表数据。

(4) 建模

对数据进行建模分析;建模可以使用 R 语言、Python 言语、Mahout 编程实现,但对实现者要求高,实现周期长,管理复杂;数据探索结果显示,经过数据分析提取,最终用于数据挖掘的数据量并不多(以某运营商省公司的“包月转包年”模型为例,经过分析后获取的宽表数据每个月数据量在 50~70 MB),这非常适合使用成熟的建模工具进行处理,本文采用支持 CRISP-DM 模型的现成建模工具。

(5) 评估

检查构造模型的步骤,确保模型可以完成业务目标;对建模的算法、效果进行分析比较,并基于测试分区进行检查。

(6) 部署

将模型获得的知识进行应用;模型导出营销用户清单及对应的推荐业务,营销部门根据该清单进行营销,并反馈营销效果。

3.3 总体框架方案

基于上述分析,框架方案可如图 2 所示,“IPTV 精准营销大数据平台”基于 Hadoop 技术搭建,汇聚了数据分析所需的数据,同时部署了爬虫程序用于爬取视频相关数据,构建用户视频画像(详见第 4.2 节),还有数据的预处理分析也在“IPTV 精准营销大数据平台”上进行,经过处理向建模工具输出宽表数据;建模工具实现数据的建模分析,向营销部门输出营销用户清单。营销由专门的营销部门实现,是方案成果的使用部门和价值体现,同时还需要根据营销结果进行数据分析和建模的调整。

4 数据汇聚和预处理

4.1 运营商数据汇聚

本文用到的运营商数据包括 IPTV 业务数据、用户套

餐信息、家庭标签、支付标签、ODS 数据、DPI 数据等,处理的方式各不相同:IPTV 业务数据目前没有汇聚到运营商大数据平台,需要与 IPTV 业务平台对接,定期采集数据到“IPTV 精准营销大数据平台”;用户套餐信息、家庭标签、支付标签等数据从运营商 ODS 系统中查询获取,将结果保存到“IPTV 精准营销大数据平台”;DPI 数据已经保存到运营商大数据平台,利用其租户空间进行数据初步分析,获取用户视屏标签数据,结果汇聚到“IPTV 精准营销大数据平台”。

4.2 通过数据爬取构建用户视频画像

数据爬取需求主要包括:节目信息爬取和建立视频标签的数据爬取。

IPTV 业务平台只有节目的名称,没有视频节目标签数据,如节目分类、主演、地区等;需要通过爬虫爬取节目信息,为用户 IPTV 播放记录打上视频标签。输入输出样例如下:

输入样例:00000050000000010000000011024895|爱回家_6_26

输出样例:00000050000000010000000011024895|爱回家_6_26|标题=爱·回家|年份=2012|分类=电视剧|主演=刘丹/徐荣/黎诺懿/郭少芸/朱慧敏/林漪琪|地区=香港|导演=徐遇安|类型=剧情/搞笑/时装/家庭|编剧=冼翠贞

对于没有使用 IPTV 业务的宽带用户,可通过其 DPI 数据获取视频观看记录,构建其视频画像,虽然运营商大数据平台已有包括视频标签的完整标签体系,但指向性不够,兼顾目标和效率,本文选取乐视、优酷、爱奇艺等几个规模大、更接近电视屏的视频网站,基于用户对这些视频网站的访问行为构建用户的视频画像,为此本文通过爬虫爬取了明星库、节目库、基于主要视频网站的视频分类和标签信息,构建了用于刻画用户画像的视频标签系统。

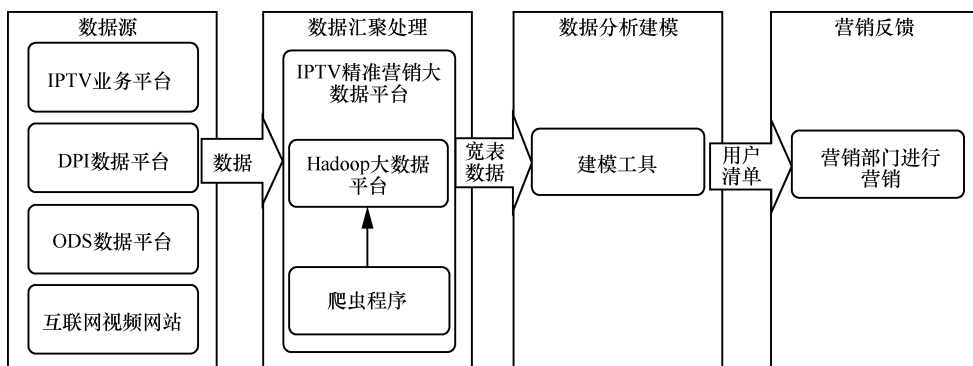


图 2 IPTV 精准营销框架方案流程



爬虫程序基于开源项目 Scrapy 自主开发,并分布式部署在“IPTV 精准营销大数据平台”,爬取的数据自动保存到“IPTV 精准营销大数据平台”。其中爬取的节目信息直接通过自主开发的程序为用户 IPTV 观看行为数据打视频标签;爬取的明星库、节目库、视频分类信息构建了用户视频标签系统,并运营商大数据平台租户空间中对用户 DPI 数据打视频标签,并将最终打上视频标签的数据汇聚到“IPTV 精准营销大数据平台”。

4.3 数据预处理

预处理包括从用户维度对数据进行统一处理,并根据业务逻辑进行数据抽取,形成统一的宽表数据,供建模分析。预处理体现了建模的人工智慧——需要确定数据抽取的范围、粒度,体现了数据挖掘的业务逻辑,其成果将直接影响建模的有效性。具体包括:数据的统一,基于用户宽带账号、IPTV 账号对应的用户标识进行数据的统一;数据提取,根据业务分析确定业务处理逻辑,并从原始数据中提取、汇总、比较等,生成最终用于建模的数据。“汇聚的数据—数据分析—建模分析”间的关系就如同“经初步处理的棉麻原料—布匹—成衣”的过程。

以 IPTV 包月转包年模型为例,经过数据分析后形成的宽表数据包括以下方面。

(1) 用户基础信息

包括用户宽带套餐、积分数据、IPTV 业务状态、IPTV 机顶盒信息(厂商、型号、清晰度、数量等)等,这些信息基本上可以直接通过数据抽取得到。

(2) 用户 IPTV 增值业务使用信息

包括按不同支付方式汇总近 3 年订购金额、次数,在订产品包个数和总金额等;用户在订/退订各年包/半年包汇总数据;用户在订、退订的最长时间产品包信息;用户点播的高清/非高清电影、电视剧等不同种类节目汇总的次数、天数、时长、节目数等;这些数据要从汇聚的数据经过计算、汇总得到;用户观看直播节目汇总信息。

(3) 用户其他画像信息

基于用户住宅区域、套餐信息和账单支付形成的用户支付画像,基于用户工作日宽带网络、IPTV 使用情况,家庭 Wi-Fi 使用、网站访问搜索和购物信息等形成的家庭人口画像,这些已有运营商大数据平台生成,可直接抽取用户标签数据使用。

还有的需要进行额外的数据比较,如通过比较用户 IPTV 清晰度和带宽信息(加上用户高清视频标签),推荐

用户进行带宽升级、IPTV 清晰度升级或两者均升级等;这需求比较 IPTV 标清/高清/4K 需要的带宽信息、用户套餐的带宽信息、用户 IPTV 机顶盒清晰度信息,定义用户带宽和 IPTV 满足度的新字段。

经过预处理的数据优化了数据结构,大大降低了数据量,更易于建模分析。提交给建模的数据包括观察数据和预测数据,观察数据用于建模,得到“实现业务目标”的用户模型,然后用这些用户模型对预测数据分析,得出“下月可能实现业务目标”的用户清单。

5 建模分析

建模过程输入宽表数据,输出营销月用户清单。整个分析过程包括:根据观察数据建立用户模型;根据用户模型对预测数据进行分析,得出下月营销目标用户,并根据用户的视频画像进行聚类,向用户推荐合适的产品包。本部分以“包月转包年”模型为例描述建模分析过程。

5.1 数据建模分析

具体的建模分析过程可分为数据预处理和数据建模两部分。

其中数据预处理主要是根据建模要求,对宽表数据做进一步处理,使其更符合建模分析的需要,以“包月转包年”模型为例,数据预处理包括以下方面。

(1) 字段处理

将宽表数据根据建模要求进行数据统计、新字段定义,如统计用户产品月龄、在订/退订产品包信息,定义“次月是否新订年包”等;基于营销时机的处理,如选择最近订购或退订过包月业务的用户。

(2) 区别不同订购方式

业务分析中,发现用户账单支付和积分支付行为差距明显,在建模前将两类数据区别开,分别进行建模。

(3) 数据分区

建模中将数据按训练数据和测试数据进行分区。

(4) 数据平衡

整体用户中订购用户相对稀疏,以某月数据为例,订购用户和未订购用户之比接近 1:60,在建模前需要对订购用户和未订购用户进行数据平衡。

数据建模主要是选择合适的算法对数据进行分析,并选择最优模型用于用户预测。“包月转包年”模型建模选择了“自动分类器”中的 C5、判别、CHAID、C&R 树、Quest 等算法进行建模,其中账单支付的各算法建模结果比较如

排序方式(S): 使用 ☒ 升序 ☐ 降序 视图: 训练集

是否...	图形	模型	构建时间 (分钟)	最大 利润	最大利润 发生比率 (%)	增益(前 30%)	总体 精确性 (%)	使用的字段编 号	曲线下方 面积
☒		判别...	< 1	30.0	1	1.815	68.037	62	0.740
☐		C5 1	< 1	120.0	1	1.858	66.361	23	0.765
☐		CH...	< 1	3.378	0	1.782	64.685	10	0.731
☐		C&...	< 1	-14.706	0	1.493	59.856	13	0.678
☐		Que...	< 1	-49.528	0	1.49	59.697	14	0.656

(a) 训练集结果

排序方式(S): 使用 ☒ 升序 ☐ 降序 视图: 测试集

是否...	图形	模型	构建时间 (分钟)	最大 利润	最大利润 发生比率 (%)	增益(前 30%)	总体 精确性 (%)	使用的字段编 号	曲线下方 面积
☒		判别...	< 1	-825.0	0	1.974	66.346	62	0.719
☐		CH...	< 1	-923.788	0	1.841	60.4	10	0.700
☐		C5 1	< 1	-895.0	0	1.736	60.177	23	0.683
☐		C&...	< 1	-907.110	0	1.631	56.332	13	0.671
☐		Que...	< 1	-947.204	0	1.623	56.045	14	0.643

(b) 测试集结果

图 3 账单支付下各算法建模结果比较

图3所示,基于“总体精确性”指标选择最优算法(账单支付中选择判别算法),用于后续进行用户预测的模型。

积分订购建模中总体成功率更高,以选择的“C&R树”算法为例,训练数据集结果为 78.299%,测试数据集结果为 84.206%,这与业务分析是一致的;对于积分支付用户来说,只要用户喜欢这些增值业务(有视频观看偏向),手里又有积分,那订购的成功率还是较高的;对于账单支付用户来说,用户要真金白银的出钱考虑的因素就多了,况且还有乐视、优酷等其他互联网品牌可以选择。

5.2 导出营销用户清单

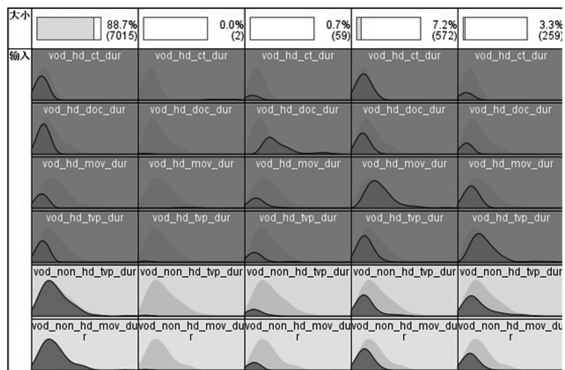
导出营销用户清单包括数据预处理、用户预测和推荐产品包选择。数据预处理包括字段处理、用户剔除(最近半年已营销用户等)和区分用户支付方式等。

数据预处理后,使用第 5.1 节中选定的数据模型分别对账单支付用户和积分用户进行预测,得到下个月“根据模型可能订购业务”的用户,并设置置信度阈值,选取营销目标用户。

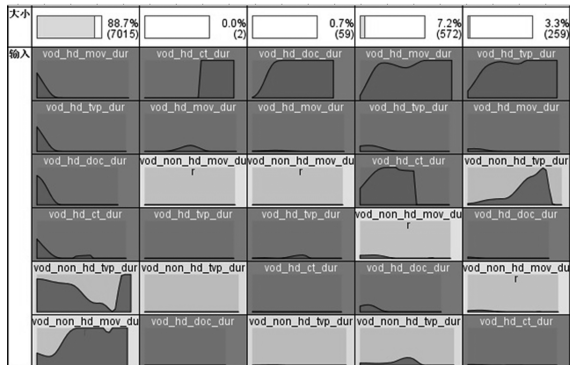
最终提交的营销清单还包括向用户推荐的产品包,严格意义上应该分析用户的视频观看行为得到用户的视频画

像,向用户推荐“用户最感兴趣”的节目,实际营销中营销的产品包往往是有限的——根据自身资源和统计“最热门”的产品来确定。本次营销根据 IPTV 平台的节目资源和以往用户订购统计信息,确定了包括“影院高清晰年包”“热剧年付”“全能看包年”“纪实高清半年包”“动漫高清半年包”等在内的十几个产品包,内容集中在电影、电视剧、动漫、纪实上。

为此,采用 k -means 聚类算法,对用户视频观看行为进行聚类,并根据营销的产品包显示高清电影观看时长(vod_hd_mov_dur)、标清电影观看时长(vod_non_hd_mov_dur)、高清电视剧观看时长(vod_hd_tvp_dur)、标清电视剧观看时长(vod_non_hd_tvp_dur)、高清纪实观看时长(vod_hd_doc_dur)、高清动漫观看时长(vod_hd_ct_dur)在聚类中的分布,结果如图 4 所示,其中图 4(a)是按“总体重要性排序”的绝对分布,从中可看出对总体重要性而言,动漫、纪实排在前两位,这可以解释大多用户对电影、电视剧偏向的普遍性;图 4(b)是“聚类内重要性”的相对分布,基于图 4 向用户推荐产品包(其中聚类 2 只有两个值作为离群值概率),见表 1。



(a) 按总体重要性排序的绝对分布



(b) 按聚类内重要性的相对分布

图4 对用户观看内容的聚类结果

表1 向不同聚类用户推荐不同产品包

聚类	推荐产品包1	推荐产品包2
聚类1	高清影院年包	热剧年付
聚类2	高清纪实半年包	热剧高清半年包
聚类3	全能看	高清影院年包
聚类4	热剧年付	首映+看大片年付

5.3 模型评估优化

模型的评估优化包括部署前模型本身的评估和部署后根据营销反馈的优化。

模型本身的评估包括:算法和测试分区的验证、平衡因子的调整、字段的调整等。首先,通过选择不同的算法得出不同的算法模型,结合测试分区的验证,选择最优算法或算法组合;本例中通过选择“自动分类器”中不同算法,并结合测试分区的验证确定最优模型。然后通过调整平衡因子调整目标用户在整体数据中的比例,多次运行模型,确定平衡因子的最优范围。还有结合建模结果中各字段的重要性进行字段微调。

同时模型还要根据营销反馈的数据进行优化,可以结合建模将用户随机分成几组,调整算法、置信度、字段等,结合不同的营销效果进行模型的调优,这部分工作还在进

行中。

总之,数据挖掘中建立数据模型不是分析的结束,而是又一轮分析的开始,要经过各种手段的调整优化,不断提升数据挖掘的效果。

6 结束语

本次分析主要配合某运营商省公司的电销进行,从效果来看,营销成功率比原先提升将近一倍,效果最好的“包月转包年”营销成功率由原来的不到5%提升到12%,应该说取得了比较好的效果。

当然数据挖掘在营销过程中主要还是幕后的“军师”,本文主要通过数据建模得出下月有可能成为“IPTV用户”或“订购增值业务的用户”。而实际营销主要是向用户提供满足用户需求的、性价比可接受的产品,大数据分析在IPTV用户需求分析、产品提供等其他方面也可以发挥幕后军师的作用,全方位提升IPTV的营销效果。同时,大数据分析也可以扩大到其他更广的领域,就运营商内部而言,在终端换机、3G升4G、交叉营销等方面,已经利用大数据分析技术来有的放矢提高营销成功率。可以说,大数据技术在电信行业已经也将发挥更大的作用。

参考文献:

- [1] CRISP-DM 方法论[EB/OL]. [2016-11-04]. [http://wiki.mbalib.com/wiki/CRISP-DM 方法论](http://wiki.mbalib.com/wiki/CRISP-DM_方法论).

[作者简介]



于洪涌(1976-),男,中国电信股份有限公司上海研究院数据分析师,主要从事IPTV数据分析、智慧家庭数据分析等方面工作。

闻剑峰(1977-),男,中国电信股份有限公司上海研究院大数据智慧运营研发(高级)工程师,主要从事大数据基础架构研发工作。

蔡鑫(1975-),男,中国电信股份有限公司上海研究院高级工程师,主要研究方向为数据规划、数据分析、数据标准。

邱晨旭(1973-),男,中国电信股份有限公司上海研究院高级工程师,主要从事电信数据规划、项目管理等工作。