



基于异构关联的大数据价值密度提升方法

汪少敏, 王铮

(中国电信股份有限公司上海研究院, 上海 200122)

摘要: 电信大数据通常分散存储在 DPI、OIDD、CRM 等多个系统中, 且格式、表述和规则在各系统中互不相同; 因而, 同一对象在不同系统中的多类数据很难被有效识别及完整利用, 大数据分析的样本规模和特征维度严重受限, 导致分析结果可信度和准确率下降。提出了电信大数据的异构关联方法与实现架构, 并进行了方法的流程举例和验证, 从用户维度实现了多系统间的数据融合, 优化了诸如用户画像等应用的数据样本空间, 从而大幅提升电信大数据价值密度。

关键词: 大数据; 电信大数据; 多源异构; 异构关联

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017341

Method of improving big data value density based on heterogeneous association

WANG Shaomin, WANG Zheng

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

Abstract: The big data resources possessed by telecom operators are usually distributed in many different systems, such as DPI、OIDD、CRM. Moreover, the formulation, interpretation and rules of the big data are not always the same in different systems. Therefore, it is difficult to identify and utilize the same object's multi-type data in different systems. Big data analysis' sample size and dimension are limited, with the decreasing of analysis results' reality and accuracy. The methods, architectures and implementation examples of big data's heterogeneous association were presented. The data fusion in user-dimension from different systems could optimize the data sample space of applications, such as user portrait. Thus, the value of carrier's big data density was greatly improved.

Key words: big data, telecom service big data, multi-source and heterogeneous, heterogeneous association

1 引言

大数据已在各行业开展广泛应用, 其中电信行业由于其天然的数据基础和应用需求, 是大数据应用的重点领域。电信大数据是指基于运营商丰富的大数据资源进行数据采集、数据处理、数

据挖掘分析及应用^[1]。这些数据包括: 互联网及移动互联网的用户行为数据、用户位置数据、用户电信业务数据、网络信令数据等。随着大数据与人工智能技术的飞速发展, 作为蕴含巨大社会价值和商业价值的电信大数据, 已被运营商列为重点应用课题^[2]。



然而,由于运营商的架构特点,运营商的这些数据分散存储在不同的系统中,例如,用户的上网行为数据、位置数据和信令数据、业务信息数据分别存储在 DPI (deep packet inspect, 深度报文识别) 系统、OIDD (open information of dynamic data, 开放信息动态数据) 系统、ODMS (operation data management system, 运营数据管理系统) 中。其造成了运营商丰富的大数据多源异构的现状。运营商对这些数据进行挖掘处理时,多采用系统内部分析处理的方式,不能进行多系统间数据融合、交叉分析^[3]。在大数据应用,特别是用户画像方面,数据的准确性、全面性都大打折扣,使得数据价值密度受到限制。解决电信大数据多源异构问题与提升电信大数据价值密度的需求越来越迫切。

本文分析了电信大数据多源异构问题,提出了电信大数据异构关联方法和异构关联实现架构,并进行了方法的流程举例和验证。该方法基于运营商最有价值且数量庞大的用户数据,根据不同来源数据的业务逻辑,通过可关联的字段,实现了多数据源的数据之间以用户维度的关联匹配规则。以自然人识别 ID 为主键将不同数据源、不同业务逻辑的数据进行串联,实现不同数据间的内容匹配。本文提出的方法能有效解决电信大数据多源异构所造成的同一用户的多类数据无法关联、数据分析维度及样本规模降低等问题,不仅能扩大用户维度的数据一次性挖掘分析可涉及的数据范围,还可以实现分散在不同系统中的数据源在用户维度的紧耦合,从而实现电信业务数据的收敛和交叉融合,使数据信息更完整,挖掘价值更大。

2 电信大数据的多源异构问题

电信大数据包含的数据种类繁多,从数据载体角度,分为用户数据、网络数据和运维数据。用户数据包括个人用户和行业用户的信息、业务及行为等数据。网络数据包括电信 3G、4G 移动

网络及宽带网络产生的数据,如信令数据等。运维数据包括电信网络运维过程中产生的数据,如设备日志数据等。其中,较为有价值且常被用来分析挖掘的数据是用户数据和网络数据中的位置数据。这些数据多为结构化数据^[4],易于分析处理,可用于分析用户兴趣偏好、用户行为追踪等,从而为电信提供基于用户画像的策略分析、精准营销和客户关怀,以减少客户流失、增加市场收入、提升客户感知及忠诚度。并且,这些数据的数据量巨大,例如,省级 4G 网络 DPI 数据,每天的数据增量为 TB 级别。巨大的数据量为大数据分析的准确性提供了基础。所以,电信大数据,特别是用户数据和位置数据,有很高的分析价值^[5]。

然而,电信大数据分散存储在电信网络中的不同系统上。如图 1 所示,用户的宽带上网行为数据存储在宽带 DPI 设备中;移动上网行为数据存储在移动 DPI 设备中;用户的业务信息数据存在 ODMS 中;用户的计费 and 基本信息数据存储在 CRM (customer relationship management, 客户关系管理) 系统中;位置数据和信令数据存储在 OIDD 系统中。这些系统对数据的解释、数据的表述、数据的格式均不相同,系统间数据互不关联,相互割裂。这形成了电信大数据多源异构的现状。

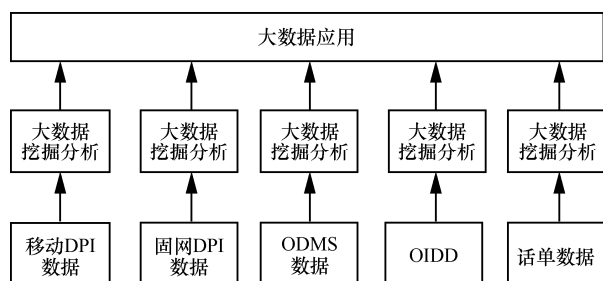


图 1 电信大数据多源异构现状

电信大数据的多源异构造成了丰富的大数据资源被分散,价值密度降低,体现在以下方面。

(1) 同一用户的多类数据无法关联

由于数据的多源异构,同一用户的多种数据

存储在不同系统中,例如用户的移动 DPI 数据、固网宽带 DPI 数据、用户话单数据分散在 3 个独立系统中。这些系统中的数据相互独立、互不关联,数据规范不一致,导致这些数据无法对应到同一用户,从而无法结合这 3 种数据分析用户行为特征,不能绘制出较为完整的用户画像,使得数据价值受到限制。

(2) 数据分析维度降低

进行大数据挖掘分析时,只能挖掘某单一系统的数据,不能结合多系统的数据进行分析挖掘,造成数据分析的维度降低,反映事物特性的特征减少,从而导致大数据挖掘分析的准确性、全面性降低。

(3) 数据分析的样本规模被限制

由于大数据分析是基于数据样本学习,所以数据样本数量越大,数据分析结果的准确率越高。当只能通过单一系统数据进行数据分析挖掘时,分析样本的数据量和丰富程度被限制。所以,电信大数据的多源异构特点,限制了数据分析的样本规模,降低了分析结果的准确率。

3 电信大数据异构关联方法

为了提升电信大数据价值密度,解决电信大数据多源异构所造成的问题,本文提出电信大数据异构关联方法。该方法在数据汇聚后的数据处理层实现,对采集的数据清洗后,对所有数据进行自然人识别和标识,标识后的数据可实现多系统数据间的关联分析,从而实现多源异构数据的关联和拼接,解决多源异构造成的同一用户的多类数据无法关联、数据分析维度降低等问题。

本文提出电信大数据异构关联方法主要包括以下 3 步,如图 2 所示。

步骤 1 面向身份信息的关键字提取。对各系统的数据进行字段分析,提取其中反映用户身份的 ID 信息,形成该条数据的关键字,如手机号码、宽带账号、各社交 UID、MAC 地址等。提

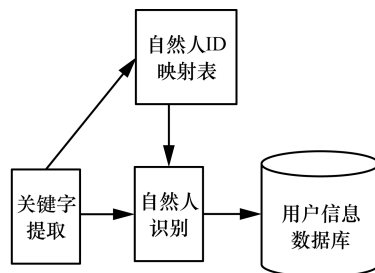


图2 电信大数据异构关联关键技术

取的数据关键字用于下一步查找自然人 ID 映射表,同时也可用于生成和维护自然人 ID 映射表。

步骤 2 基于自然人 ID 映射表的数据识别。本文提出了自然人识别 ID 和自然人 ID 映射表。一个自然人可以拥有手机号码、宽带账号、社交网络身份 ID 等多种用户 ID,所以需要在所有用户标识 ID 之上,建立一套全网统一的自然人识别 ID,自然人识别 ID 是不同系统数据间用户的唯一性标识。通过自然人识别 ID 这个唯一标识,将不同数据源的数据进行关联。自然人 ID 映射表保存用户的各种 ID,包括自然人识别 ID、手机号码、宽带账号、各社交 UID 等。将数据提取的关键字在自然人 ID 映射表中查找匹配的 ID (手机号码、宽带账号和各社交 UID 等),匹配 ID 对应的自然人识别 ID 即为该条数据的自然人识别 ID。

步骤 3 对数据进行自然人标签标识。通过对所有系统的数据加上自然人识别 ID 标签的方式进行数据的自然人标识。不同数据源的数据加上全网统一的自然人识别 ID 之后,在进一步的数据挖掘分析时,可通过自然人识别 ID 进行关联分析,从而实现多源异构数据的关联融合。例如,移动 DPI 数据、固网 DPI 数据和话单数据加上自然人识别 ID 后,可通过自然人识别 ID 区别出同一用户的移动 DPI 数据、固网 DPI 数据和话单数据,这样可以结合用户的移动上网行为和固网上网行为以及电话呼叫行为等多种类型的数据,更全面地绘制出该用户的用户画像。



电信大数据异构关联关键技术有以下几种。

(1) 面向身份信息的关键字提取

本文提出的电信大数据异构关联方法中,通过对数据关键字的提取和比对,识别数据所对应的自然人,所以,关键字和用户身份强相关,如手机号码、宽带账号、用户社交 UID 等。由于不同来源的数据格式和数据内容互不相同,所以不同来源的数据具有不同的关键字,表 1 列举了固网 DPI 数据、移动 DPI 数据、话单数据、ODMS 数据、OIDD 数据和 ODS 数据的关键字。

表 1 电信大数据异构关联关键字举例

数据类型	关键字
固网宽带 DPI 数据	宽带账号; cookie 信息; 用户社交 UID (微博 UID、QQ 空间 UID 等)
移动 DPI 数据	手机号码; 用户社交 UID (微博 UID、QQ 空间 UID 等)
话单数据	手机号码 (主叫、被叫)
ODMS 数据	手机号码
OIDD 数据	手机号码
ODS 数据	手机号码; 宽带账号

关键字可以通过解析数据中的字段获得,例如:解析固网 DPI 数据的 AD 账号字段,可获得用户的宽带账号关键字;解析宽带和移动 DPI 数据,从 HTTP 业务用户访问记录中的 DestinationURL 字段中提取微博、腾讯社交网络 UID,可获得用户的社交 UID 关键字;解析话单数据中的主叫号码、被叫号码字段,可获得手机号码关键字。

当从数据中提取关键字后,在自然人 ID 映射表中查找关键字,若在自然人 ID 映射表中找到了关键字匹配的条目,则从自然人 ID 映射表中获得此关键字的自然人识别 ID,即此条数据的自然人识别 ID,从而实现此条数据的自然人识别;若没有找到匹配条目,将更新自然人 ID 映射表中的条目。所以,提取的关键字有两种用途:一是用于生成和维护自然人 ID 映射表,二是有效识别数据

的自然人身份。

(2) 自然人 ID 映射表

本文提出了自然人识别 ID 和自然人 ID 映射表。通过数据中的关键字查找自然人 ID 映射表,能够获得此条数据的自然人识别 ID。所以,自然人 ID 映射表保存了用户的各种 ID 和映射关系,包括自然人识别 ID、手机号码、宽带账号以及各社交 UID 等。自然人 ID 映射表的键值 ID 为自然人识别 ID,它为全网统一的标注和识别该用户的 ID。自然人 ID 映射表还包含了用户自然人识别 ID 和用户其他各种 ID 的映射关系。自然人 ID 映射表举例如图 3 所示。

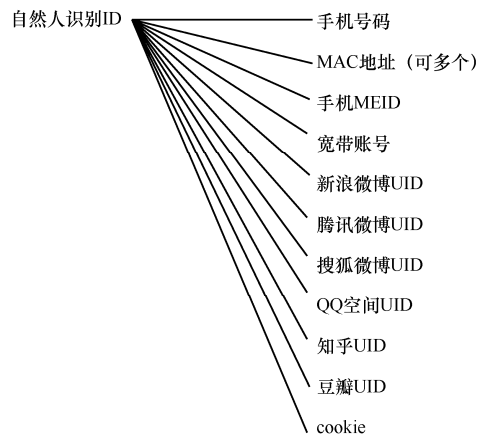


图 3 自然人 ID 映射表举例

① 自然人 ID 映射表的生成

自然人 ID 映射表中的内容通过从移动 DPI 数据、固网 DPI 数据、ODS 数据和终端自注册数据等数据中提取的关键字生成和更新。例如:通过提取 ODS 数据的关键字,获取自然人的手机号码、宽带账号并分配自然人识别 ID,写入自然 ID 映射表;通过终端自注册平台数据的关键字获取手机 MAC 地址、MEID;通过移动 DPI 数据的关键字获取微博 UID、QQ 空间 UID 等社交 UID。

自然人 ID 映射表的生成可分为两步:先通过上述方法分析存量的电信大数据(已有未识别数据),初步生成自然人 ID 映射表中的内容;再对

新采集的数据采用上述方法进行实时分析,从而不断增加自然人ID映射表的内容。

② 自然人ID映射表的维护

已生成的自然人ID映射表需要不断地更新维护。自然人ID映射表的更新维护同样通过对电信大数据的关键字分析,从数据中提取的关键字在已生成的ID映射表中进行查找匹配,若查找的关键字和自然人ID映射表中的字段匹配成功,则得到了该条数据的自然人识别ID;若在自然人ID映射表中没有相匹配的ID,则说明此条数据的关键字为新的ID信息,应更新到自然人ID映射表中。通过数据关键字查找匹配自然人ID映射表的方式,实现自然人ID映射表在应用过程中的更新维护。

(3) 数据的自然人标签标识

数据经过关键字提取后,根据获得的关键字,查找自然人ID映射表,找到和关键字一致的ID,从而获得该条数据的自然人识别ID,然后通过对数据增加标签的形式,将该自然人识别ID标注到该条数据。通过对每条数据标注自然人识别ID,达到不同数据相互关联的目的。

一条数据可能存在多个关键词的情况,如移动DPI数据可能解析出手机号码、新浪微博UID等关键字。这种情况下应按一定的优先级顺序,比对关键字和自然人ID映射表中的ID。根据各关键字和自然人的关联程度的强弱不同,可设置关键字查找比对的优先级顺序如下:手机号码>MEID号码>MAC地址>cookie>新浪微博UID>社交UID>宽带账号。当关键字匹配优先级高的ID时,使用优先级高的ID所对应的自然人识别ID标识数据。

4 电信大数据异构关联的实现架构

本文提出了电信大数据异构关联方法应用到实际网络中的两种方式实现架构:一种是关联后再存储,另一种是存储后再关联。

4.1 方式1: 关联后再存储

电信大数据关联异构关联实现架构方式1如图4所示。

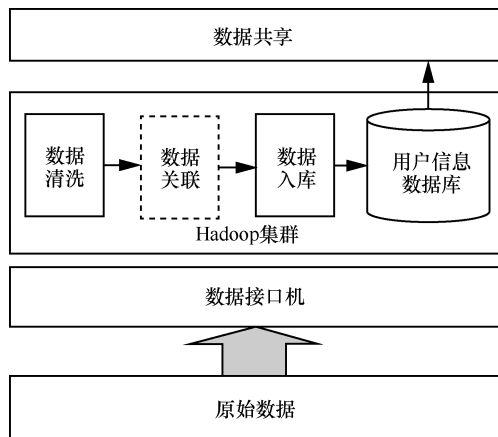


图4 电信大数据关联异构关联实现架构方式1

关联后再存储的实现方式为:原始数据经过清洗后,逐一对每条数据进行自然人识别,并加上自然人识别ID的标签。然后,将加上自然人识别ID后的数据存入数据库,数据库中所有的数据均是标注了自然人识别ID的已关联数据。上层应用使用数据时,直接分析挖掘已关联数据。

这种方式的特点在于:

- 在数据清洗阶段即完成每条数据的关联,这将增加数据清洗阶段的工作量和数据存入数据库的时长;
- 因为数据入库前要逐条解析数据并关联,所以对系统性能有一定要求;
- 存储后的数据皆为已关联数据,使用方便,可实时取用。

4.2 方式2: 存储后再关联

电信大数据关联异构关联实现架构方式2如图5所示。

存储后再关联的实现方式为:原始数据不做关联即存入数据库。当数据应用请求需要关联数据时,数据关联模块对数据库中的被请求数据进行自然人识别和加自然人识别ID标签进行数据关联。数据关联完成后,应用从数据库中取得需

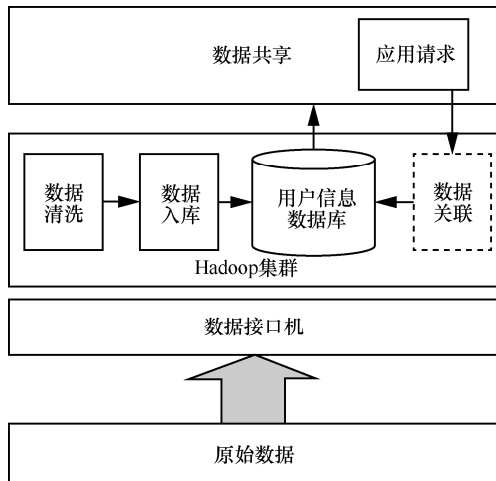


图5 电信大数据关联异构关联实现架构方式2

要的已关联数据。库中关联后的数据保留标签，供下次应用请求时取用，避免重复关联。

这种方式的特点在于：

- 不影响清洗入库流程，不影响数据存入数据库的时长；
- 数据关联按需实现，不需要对数据逐条解析和关联，对系统性能要求较方式1较低；
- 应用不能实时取用已关联数据，需先请求关联，等待关联完成后取得关联数据，通过分步请求方式获得关联数据。

5 流程举例和验证

根据本文中的电信大数据异构关联方法，以移动DPI数据自然人识别和标注为例，验证异构关联的可行性和效果，如图6所示，对某一条的移动DPI数据自然人识别和标注的流程如下：对清洗后的移

动DPI数据进行关键字提取，得到关键字——手机号码；在自然人ID映射表中查找该手机号码，得到该手机号码映射的自然人识别ID；在清洗后的数据中加入自然人识别ID标签，标注自然人；标注后的该条移动DPI数据入库。根据上述流程，以中国电信多个省市一天的4G移动DPI数据为数据源，对这些数据逐条进行识别和关联验证。验证环境为Linux上的Hadoop系统。数据源情况见表2。

表2 电信业务大数据异构关联验证数据源情况

数据类型	4G 移动 DPI
数据源	北京、广东、辽宁、山东、上海、陕西、四川 (上海三天数据，其他省市一天数据)
数据量	2.8 TB
数据条目数	77.76 亿

验证结果如图7所示。验证生成的自然人ID映射表，包括条目9 374 328条。原始数据77.76亿条，可标识数据77.76亿条，标识率为100%。由于移动DPI数据的关键字为手机号码，所以标识率较高。

6 结束语

运营商作为数据密集型企业，有丰富的大数据资源。然而这些数据资源分布在多个相互独立的系统中，存在多源异构情况，数据相互独立，互不关联，所以数据价值未被充分挖掘。本文提出了基于异构关联的大数据处理方法，可以实现不同数据间的内容匹配，从而提升数据价值密度，为后续高质量的数据挖掘打下基础。

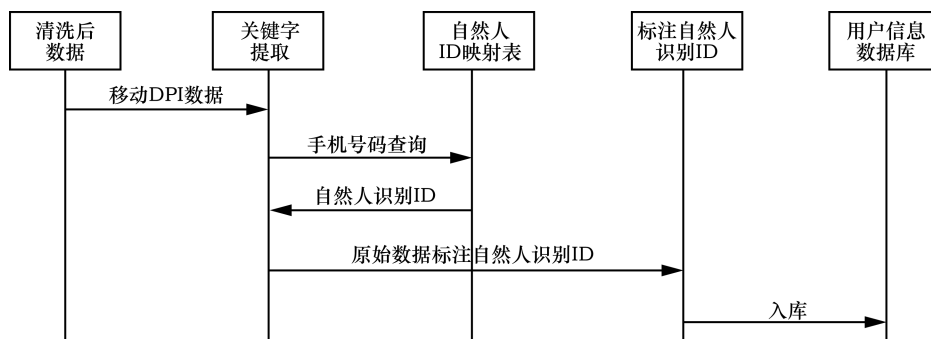


图6 移动DPI数据自然人识别和标注流程举例

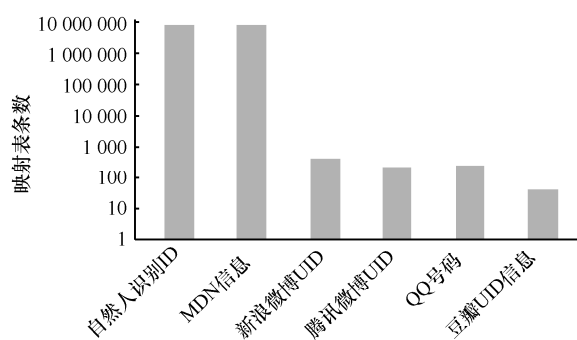


图7 移动DPI数据自然人ID映射表验证结果

参考文献:

- [1] 李秋静, 叶云. 电信大数据解决方案及实践[J]. 中兴通讯技术, 2013, 19(4): 39-41.
LI Q J, YE Y. Telco big-data solution and experience[J]. ZTE Technology Journal, 2013, 19(4): 39-41.
- [2] 童晓渝, 张云勇, 房秉毅, 等. 大数据时代电信运营商的机遇[J]. 通信信息技术, 2013(1): 5-9.
TONG X Y, ZHANG Y Y, FANG B Y, et al. Opportunities and strategies to adopt big data for telecom operators[J]. Information and Communications Technologies, 2013(1):5-9.
- [3] 韩晶, 张智江, 王健全, 等. 面向统一运营的电信运营商大数据战略[J]. 电信科学, 2014, 30(11): 154-158.
HAN J, ZHANG Z J, WANG J Q, et al. The unified opera-

tion-oriented big data strategy for telecom operators [J]. Telecommunications Science, 2014, 30(11): 154-158.

- [4] 沈雷明, 别志铭. 基于电信大数据的数据建模平台研究[J]. 电信科学, 2014, 30(6): 138-141.

SHEN L M, BIE Z M. Research on data modeling platform based on big data of telecom[J]. Telecommunications Science, 2014, 30(6):138-147

- [5] 靳丹, 张磊, 王洪军, 等. 基于Hadoop的大数据清洗框架设计与应用[J]. 网络新媒体技术, 2015(9):33-38.

JIN D, ZHANG L, WANG H J, et al. Design and application of Hadoop based data cleaning framework[J]. Journal of Network New Media, 2015(2):5-10.

[作者简介]



汪少敏(1983-), 女, 中国电信股份有限公司上海研究院工程师, 主要研究方向为大数据架构、数据挖掘分析和人工智能技术。

王铮(1973-), 男, 中国电信股份有限公司上海研究院工程师, 人工智能交互团队负责人, 主要研究方向为大数据架构、数据挖掘分析和人工智能技术。