Modern Marketing 现代市场营销, 2022, 12(4), 81-88
Published Online November 2022 in Hans. http://www.hanspub.org/journal/mom
https://doi.org/10.12677/mom.2022.124010

Hans汉斯

新零售模式下烟草行业全产业链数据采集与集成方法研究

李中华，刘亚龙，杜　预，周家贤*，郭　梁，蒲雪松

云南中烟有限责任公司营销中心，云南　昆明

收稿日期：2022年9月25日；录用日期：2022年10月25日；发布日期：2022年11月1日

摘　要

烟草工商企业经过多年的协同发展，共同面向消费者营销模式在不断深化和变化，因为工业企业和商业企业在各自掌握的生产与消费信息不全问题，商业企业掌握了大量消费者的行为数据，工业企业掌握生产与消费者基本信息，没有实现整体信息的有效整合，导致在协同营销消费者环节深度不足，无法实现精准营销。本文主要分析了生产流程产生的一些问题与解决方法，探讨了如何利用新的数据流程框架和实际的技术对烟草生产到消费数据进行采集与集成。可以促使烟草流程信息一体化，为精准营销提供数据支持。

关键词

工商企业，流程框架采集与整合，信息一体化

Under the New Retail Model Research on the Data Collection and Integration Method of the Whole Industry Chain of the Tobacco Industry

Zhonghua Li, Yalong Liu, Yv Du, Jiaxian Zhou*, Liang Guo, Xuesong Pu

Marketing Center of Yunnan China Tobacco Co., Ltd., Kunming Yunnan

Received: Sep. 25th, 2022; accepted: Oct. 25th, 2022; published: Nov. 1st, 2022

*通讯作者。

文章引用: 李中华, 刘亚龙, 杜预, 周家贤, 郭梁, 蒲雪松. 新零售模式下烟草行业全产业链数据采集与集成方法研究[J]. 现代市场营销, 2022, 12(4): 81-88. DOI: 10.12677/mom.2022.124010

Abstract

After years of coordinated development of tobacco industrial and commercial enterprises, the common consumer marketing model is constantly deepening and changing. Because of the insufficiency of production and consumption information of industrial enterprises and commercial enterprises, commercial enterprises have mastered a large number of consumer behavior data, and industrial enterprises have mastered the basic information of production and consumers, and have not realized the effective integration of the overall information. As a result, the depth of collaborative marketing for consumers is insufficient, unable to achieve precision marketing. This paper mainly analyzes some problems and solutions of the production process, and discusses how to use the data collection and integration of the new data process framework and practical technology for tobacco production to consumption data. It can promote the integration of tobacco process information and provide data support for precision marketing.

Keywords

Industrial and Commercial Enterprises, Process Framework Collection and Integration, Information Integration

1. 引言

烟草行业对中国财政来说，一直都是支柱型产业。其一般都采用专营专卖，统一领导的体制方式。然而，随着近几年互联网的发展，特别是“互联网+”“新零售”等概念的提出，越来越多的行业与互联网接轨。新零售(New Retailing)，即企业以互联网为依托，通过运用人工智能、大数据等先进技术手段，对商品的生产、流通与销售过程进行升级改造，进而塑造全新的业态结构与生态圈，并对线上服务、线下体验以及现代物流进行深度结合的零售新模式。因此互联网成了各个行业提高效率、促进发展的一个不可或缺的工具。同时，随着互联网的加入，烟草行业信息愈发庞大，统筹方面愈发困难。因此，传统的烟草营销方式和信息整合方法已经不能满足当前的需要。

工商企业之间数据并没有进行打通，未实现数据共享，此外，消费者消费数据、行为数据、偏好数据较难获取，新零售模式下如何实现工商消费者数据共建共享，通过什么架构和技术来实现成为一项重要课题。

鉴于当前的行业背景，本文主要分析了生产流程产生的一些问题与解决方法，探讨了如何利用新的数据流程框架和实际的技术对烟草生产到消费数据进行采集与集成。可以促使烟草流程信息一体化，为精准营销提供数据支持。

2. 烟草行业数据采集与集成面临的问题

在构建新零售模式下生产者与消费者流程数据一体化工作中，很多研究人员提出了各阶段信息整合，却没有说明如何将信息源分类和具体整合技术方法。鉴于行业分工和行业地位，我们将烟草生产者到消费者流程数据分为三个阶段。每个流程阶段都会产生海量的数据。而本文将提出一种架构用于解决各个阶段产生的大量数据收集、整合问题，及三大阶段数据统合问题。第一个阶段是原料产地(烟农)到工业企

业阶段。第二个阶段是工业企业到商业企业阶段。第三个阶段是商业企业到消费者阶段。本文又从烟草生产到销售的结构体系和具体技术两个部分来解决多源数据异构问题。即三个阶段，两大部分。基本结构如图 1 所示。

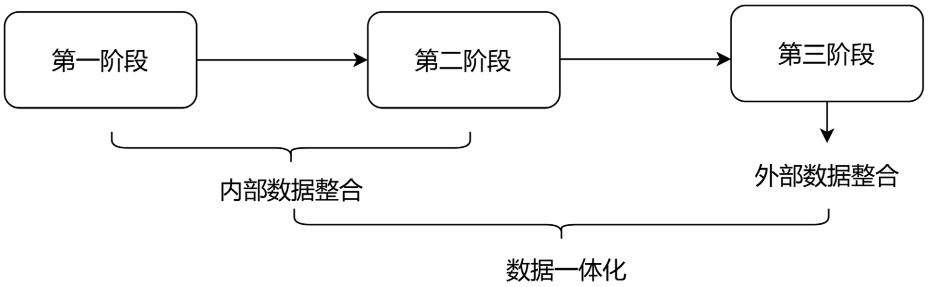

Figure 1. Schematic diagram of the basic structure
图 1. 基本结构示意图

第一阶段和第二阶段的数据是烟草生产和销售流程的过程中产生，这部分数据结构是可控的，所以可以从统一的管理系统和数据库结构可以解决。而第三阶段中的公司外部数据是不可控的，所以需要从技术角度解决。即图 1 中内部数据整合部分和外部数据整合部分。将所有数据整合后才能实现最终的数据一体化。

3. 数据来源

工业企业有自己的生产数据库和营销系统。主要包括烟叶采购，烟叶运输，烟叶存储，卷烟生产，卷烟运输，卷烟分配，卷烟销售，客户信息等数据。因此在生产销售过程中必然产生大量数据。同时，烟草商业企业为卷烟零售终端客户提供了多种信息化系统，在新零售模式下，商业企业为零售客户提供基于线上线下为一体的数字化转型工具，积累了大量的消费者基本信息和行为数据，包含消费者订单数据、消费者支付数据、消费者基本信息、消费者喜好等信息。

工商企业之间数据并没有进行打通，消费者数据是完全独立，未实现数据共享，如何通过技术手段实现新零售模式下如何实现工商消费者数据共建共享，通过什么架构和技术来实现成为一项重要课题。

3.1. 内部数据源

内部数据源整合需要设立一个新的烟草生产到销售的数据结构体系。首先，内部数据产生于三大阶段的第一和第二阶段。这部分数据结构是可控的，所以可以从统一的管理系统和数据库结构可以解决。

1) 第一个阶段是原料产地(烟农)到工业企业阶段。

首先烟农会将种植好的烟草原料卖给烟草工作站(点)，会产生最原始的烟草收购数据和烟农个人信息数据。从烟叶工作站(点)到中心仓库或周转仓库，会产生烟草存储和转移数据。从中心仓库或周转仓库到烟草工业企业的烟叶复烤公司，会产生烟草总体的收购数据和总体原料数据。注，在此过程中卷烟厂(工业企业)不能直接收购烟叶，而是需要烟草专卖局(商业企业)收购烟叶然后才能将烟叶交付卷烟厂，因此这个转交的过程又会产生大量的数据。第一阶段结构如图 2 所示。

第一阶段产生的问题：a) 烟叶收购评级分拣问题。当前，烟叶评级分拣工作还是采用人工的方式。这种方式对于当前信息化，高效化的社会来说太过原始。b) 烟叶存储运输问题。当前阶段，存储运输的投资成本完全取决于管理层的个人判断。其投资决策、预算需求、实际需求必然存在很大误差。实际投资一般都大于实际需求。造成了仓库存储资源冗余的情况[1]。

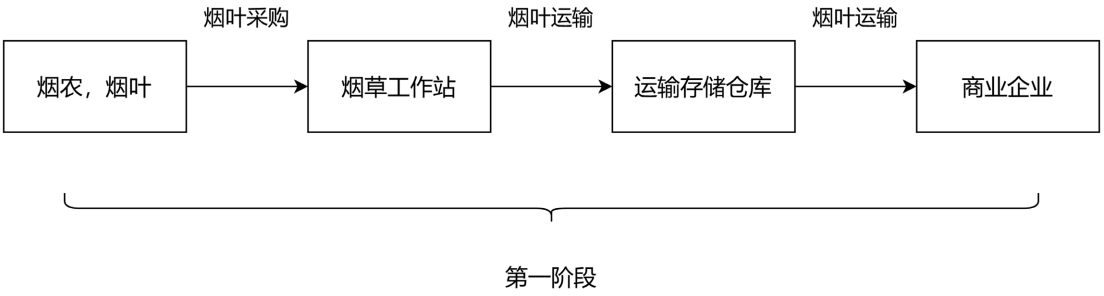

Figure 2. Schematic diagram of the first stage structure
图 2. 第一阶段结构示意图

2) 第二个阶段是工业企业到商业企业阶段。

这个过程中工业企业生产的大量卷烟会交付给商业企业，这个过程又会产生大量数据。再由市级商业企业获得后交付给下级商业企业，同时还要结合之前时间段下级单位的销售情况，考虑各个下级单位商品分配问题，这个过程会产生更多的卷烟数据。最后再将分配的卷烟交给零售终端去售卖，这里同样要考虑卷烟分配问题，因此同样会产生更多的数据。第二阶段结构如图 3 所示。

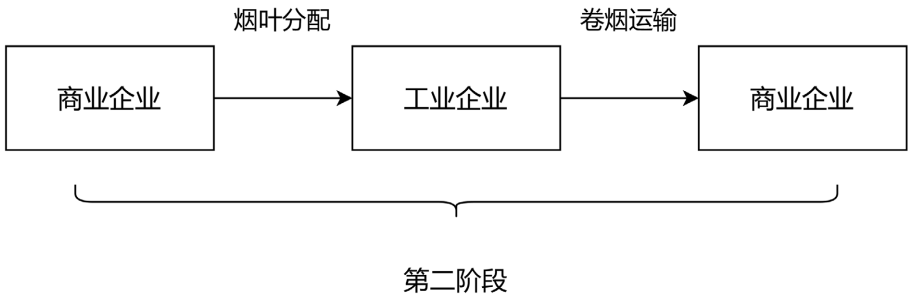

Figure 3. Schematic diagram of the second stage structure
图 3. 第二阶段结构示意图

第二阶段产生的问题：工业企业到商业企业数据整合一体化问题。当前情况下，工业企业有自己的数据，商业企业同样有自己数据。两者的数据不能互通，这对烟草行业的数据一体化产生了巨大障碍。数据的传输和处理，分析，都将难以进行。更不用说整合、预测。整个行业的数据都如一盘散沙，各个部门都是信息孤岛，数据冗余和数据重用必然发生。解决这个问题刻不容缓。也是本文要解决的问题核心。

3.2. 外部数据源

由于烟草公司卷烟需要交由零售商发售，而零售商售卖方式各种各样，特别是新零售模式下，线上线下方式结合，产生的数据种类更多。主要的信息来源是各种外部平台，与内部数据完全不同。因此将这部分的信息数据称为外部数据，产生过程称为第三阶段。

第三个阶段是商业企业到消费者阶段。在这个阶段中，商业企业的终端零售店将卷烟卖给给大消费者。会产生更加海量的数据，例如：消费者订单数据、消费者支付数据、消费者基本信息、消费者行为信息、消费者喜好等信息。

第三阶段的数据产生的问题：

1) 各大软件平台烟草购买数据的多源异构问题。在烟草行业，在数据量极为庞大的基础情况下，各

大数据平台还存在数据管理及存储模式多样性。数据库模型一般存在两种：关系型数据库与非关系型数据库。这两种数据库存储的方式不同。而就算数据存储方式相同，各个数据库表的格式也可能不同。以上各种原因都会最终导致数据整合困难。

2) 对供应链过程优缺点逆向分析。在数据整合完成后，数据的分析，判断非常重要。一旦出现产品问题、客户投诉、客户购买量出现正增长或者负增长等情况，我们需要知道问题根源在哪。而根源的最初源头可能就在供应链的源头。因此我们需要通过数据分析把握住整个供应链的情况。

4. 数据采集与集成的方法

为了解决上文中出现的两大数据源各自产生的问题和需求。这一部分将烟草公司的数据整合方法对应的分为两大方面：

4.1. 烟草公司内部信息整合

在烟草生产过程中，从烟叶到将卷烟分配给零售商这段步骤可以看出，所有的信息的结构都可以由卷烟厂自己控制。即，可以设计一套卷烟厂自己的数据结构将所有信息记录下来。而记录信息必须要设计烟草公司自己的数据库，同时还要考虑到个部门信息交换的便利性，所以必须要设计一个网上烟草信息管理平台，从而从源头上对烟草数据进行管理是必不可少的。

1) 对第一阶段信息的整合。

- 人工智能处理烟叶收购评级分拣问题。

当前，烟叶评级分拣工作还是采用人工的方式。针对这个问题，我们可以利用人工智能的方法，先利用图片对人工智能网络进行训练，训练完成后就可以采用准确率高的权值，然后将跑好的智能程序安装在各个识别烟叶终端的电脑上，再利用摄像头对烟叶进行拍照，上传照片，然后根基比对情况，电脑会自动给出最终结果。其实当前在物品评级分拣方面，很多地方都使用了人工智能技术。何永康就设计一套不同于机械臂的流水线拨杆式可回收物智能分拣系统[2]。在处理过程中产生的数据就可以上传到总数据库。这解决了人工评级分拣的三个问题。在这一阶段需要保存原料产地(烟农)的基本信息，烟草收购信息等。同时为接下来的阶段提供数据基础。

- 处理烟叶存储运输问题。

根据上个阶段数据可以知道各个站点收购情况，可以得出总体收购情况，从而计算出各个区域所需仓库实际大小，再也不需要对建造仓库的大小进行主观猜测。即，通过上个阶段收购烟叶数据的分析可以直接解决烟叶存储运输问题，为烟草公司节省大量冗余开支。同时这阶段产生的各种数据同样需要上传总数据库，为之后的阶段提供基础数据。

2) 对第二阶段信息的整合。

要工业企业到商业企业数据整合一体化问题，必须要有一套线上烟草信息管理系统。系统必然涉及web 开发，Java，数据库技术，服务器等。可参考王金凤的数字化下烟草行业流程管理评价模型构建研究[3]与程雲霄的烟草物流信息管理系统[4]。在系统中可以根据各部门等级，分工，地域的不同设定不同的管理权限，最高权限由总公司掌握。这样全公司各个部门都可以拥有一套统一的数据库存储结构，只有统一的数据库存储结构才能避免信息孤岛的存在，从根本上解决数据冗余问题。才能更好的对公司数据进行统一管理和分析。三大阶段信息将全部统一存储于烟草信息管理系统下面的总数据库。

总之，有统一的管理系统和数据库，还有从烟叶采购源头阶段就定好的数据采集和数据构成方法。将使公司内部信息实现一体化，从源头解决了公司内部信息的多源异构问题。

4.2. 各大软件平台烟草购买的多源异构外部数据整合

1) 这一个方面主要是对第三阶段信息的整合。在这一阶段，烟草公司主组要的信息来源是各种外部平台，烟草需要交由零售商发售，而零售商售卖方式各种各样，特别是新零售模式下，线上线下方式结合，产生的数据种类更多。因此多源异构数据无法避免，这就需要用技术手段来解决问题。

数据集成：是把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。在企业数据集成领域，国内外已经有了很多成熟的框架和方法可以利用。美国 CSC 公司开发的 Multibase 是一种集成异构式分布系统[5]。该系统为用户提供了统一的全局模式和单一的高级查询语言，同时本地数据库保留了更新的自主权[6]。北京大学开发的 CoXMLV10，能够实现数据的采集、管理和共享。该系统主要以关系模型数据库为基础，开发了协同查询应答框架，实现了与其他数据库和数据源之间的查询应答机制[6]。但是该系统在 XML 数据 DTD 描述方面还存在不足[7]。北京理工大学开发的 UUHDB 是跨平台异构型数据库，支持 GSOL，但是全局查询的功能没有经过系统地优化，目前的效率较低[6]。以上各种异构数据整合方法都有这样或那样的缺点，因此我将在后文中单独介绍更加符合当前烟草公司情况和数据异构类型的框架和方法。

JTidy 是 HTML Tidy (一个 HTML 语法检查器和优雅的打印编排工具)的 Java 移植，本身具有的清除 HTML 错误内容或文件难看的功能外，同时还提供了一个 DOM 接口，我们可以将 JTidy 当作一个处理 HTML 文件的 DOM 解析器来使用。

首先我们需要对数据进行采集，如果是网络上的数据，最简单的方法就是用 python 爬虫在网上爬取，如果是零售终端的数据，可以向零售商索要香烟交易数据。

采集过后的 HTML 等数据利用 JTidy 转换为 XML，然后利用配置文件 XSLT 转换成需要的数据。采集过来的 HTML、XML、JSON 源数据，因为格式是多样的，不利干进行格式化，所以我们统一利用 JTidy 转换为 XML，由于源数据可能会很多，而我们需要的关键数据只有部分，所以设计在配置文件中进行匹配采集，采集规则设计如下两种模式：

- 标签规则：标签模式规则采集数据是指在 htm1 元数据中指定一个起始位置和结束位置，将这两个标记中间的数据采集出来，当存在多次匹配时，会让自动累加数据到 XML，比如常见的 HTML 标签 divform、body 等，标签规则又分成 3 种模式[8]。采集模式，过滤模式，替换模式。
- 正则规则：正则模式是指使用标准的正则表达式去匹配要采集的数据，并将匹配到的数据保存到元数据中，当存在多次匹配时，自动累加数据到 XML。正则格式则是标准的正则格式，正则规则有 4 种模式。采集模式，过滤模式，替换模式，格式化输出模式。

通过 JTidy 和规则将源数据转化为 XML，这个 XML 就是我们想要得到的可以进行格式化的数据，我们利用 XSLT 格式化 XML，输出自己想要的数据格式。

最终得到的数据就是我们需要的整合好的异构数据源的数据。我们只需要将其保存到总数据库中即可。该方法简单直接，适用性强，同时适用于结构性数据和非结构性数据，比其它方法更适用于烟草行业。

2) 利用数据链数据对供应链过程优缺点逆向分析。

从三个阶段的数据整合结果来看，我们会得到从烟叶到最终客户购买数据，从这个角度来说，必然会形成一整条数据链。即，我们可以从客户购买的哪包烟逆向追踪到哪一个卷烟厂，哪些存储仓库，哪一个生产原料基地，哪一个收购烟叶的站点。由此就可以分析出商品问题事情的根源在哪。再结合大数据和人工智能就可以对未来做出合理预测和规划。例如，从卷烟的购买量，客户的购买喜好，我们可以知道哪些烟草更受喜欢，哪些烟叶种类更受客户青睐。从而合理的调整生产工艺和调整烟叶生产基地的规模。

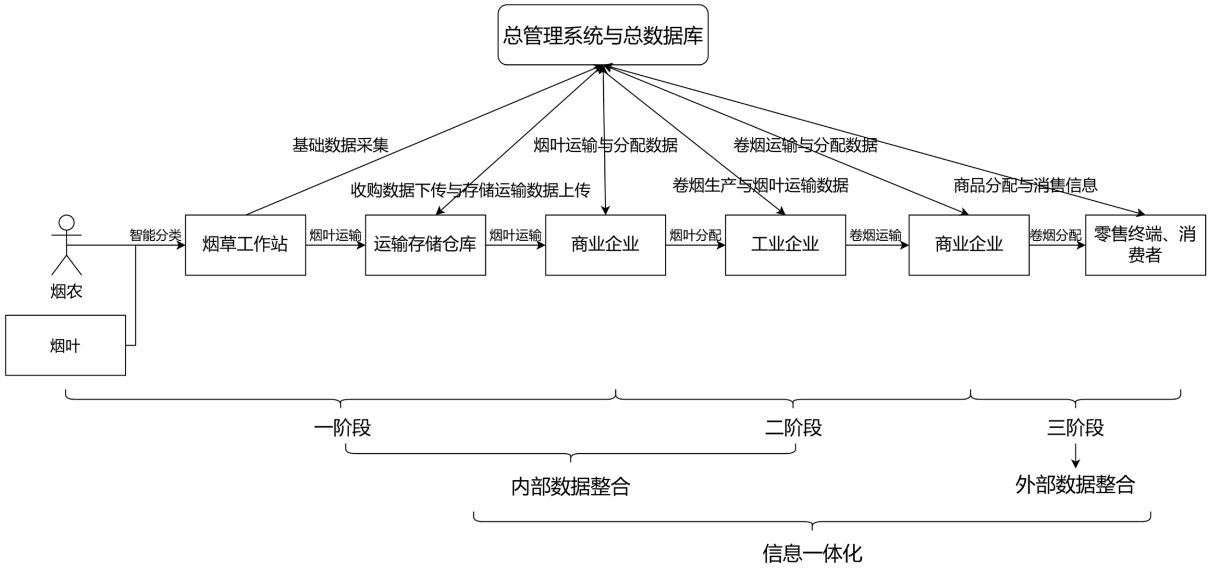

Figure 4. Schematic diagram of the overall structure
图 4. 总体结构示意图

从此一个正向的数据链体系将会建立起来，一个逆向的反馈机制同样也会建立起来。再结合数据处理。烟草公司的数据一体化将会实现，如图 4。

5. 结语

本文从烟草生产到销售的结构体系和具体技术两个方面来解决多源数据异构问题。新的体系解决了从烟叶开始的公司内部数据的多源异构问题，这部分数据结构是可控的，所以可以从统一的管理系统和数据库结构解决。而公司外部数据是不可控的，所以需要从技术角度解决。同时内部大量数据的整合也简化了外部数据的整合，不然内部数据与外部数据同时只用技术手段解决的话，将会使得数据处理难度增加很多量级，也不利于数据链的建立和数据一体化。所以，本文创新性的将烟草数据分成两块，用两种方式分开整合。

从科技发展的角度来说，烟草公司的数据一体化是必然发生的。当前很多技术都早已经在其它领域使用，基本已经成熟，因此本文中提到的方法都是可行性非常高。

基金项目

云南中烟工业有限责任公司科技计划项目(任务书编号 2022XX01)。

参考文献

[1] 周荣翼, 赵翰声, 徐莉萍. 烟农合作社下烟叶生产与销售流程再造研究[J]. 商业会计, 2016(16): 22-25.

[2] 何永康, 宋连庆, 颉清云, 郭瑞鸿. 可回收物智能分拣系统的设计与实现[J]. 设计, 2022, 43(6): 1582-1591. https://doi.org/10.16208/j.issn1000-7024.2022.06.011

[3] 王金凤, 郑成德, 张洪利, 潘滨. 数字化下烟草行业流程管理评价模型构建研究——以山东烟台烟草有限公司为例[J]. 财务管理研究, 2022(4): 9-20.

[4] 程雲霄. 烟草物流信息管理系统的设计与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.

[5] 李群, 张传顺, 吴忠明. 分布式数据库管理系统之现状[J]. 计算机工程, 1986(1): 36-47.

[6] 王航. 多源异构数据整合系统的设计与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2020. https://doi.org/10.27005/d.cnki.gdzku.2020.001173

[7] 刘云峰, 杨冬青, 唐世渭, 王腾蛟, 高军, 李缨. 基于XML数据集成与交换中的完整性约束研究[J]. 计算机工程 2005(9): 39-40+224.

[8] Chi, Y.T., Guo, P., Gao, H.J., *et al.* (2014) Automatic Batch Extraction of Specific Content of HTML Based on Tag Locations. *Applied Mechanics and Materials*, **602**, 3826-3830.