

基于 ETL-KETTLE 的贵州卷烟营销大数据分析及可视化^①

郭 丹, 樊 红

(武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430079)

摘 要: 为了实现对贵州卷烟销售大数据的分析及可视化, 基于开源数据分析工具 KETTLE 构建了 ETL 过程模型, 对现有的卷烟销售基础大数据进行抽取、转换、加载, 形成分析型数据, 快速高效地实现了数据集成. 经过 ETL 处理的集成数据, 为卷烟销售大数据可视化系统及进一步的数据挖掘、决策支持提供了数据基础. 同时, 采用 GIS 平台构建和可视化技术, 设计并实现了卷烟销售大数据可视化分析系统, 为企业的数据分析和决策支持提供了数据展示及分析平台.

关键词: ETL; KETTLE; 数据集成; GIS; 可视化

Analysis and Visualization of Cigarette Sales Data Based on ETL-KETTLE

GUO Dan, FAN Hong

(State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China)

Abstract: In this paper, based on the KETTLE which is an open source data analysis tool, an ETL process model is constructed to extract, transform and load the existing basic big data of the cigarette sales. Through the ETL model, the basic data is converted to analytical data, and the data integration can be implemented quickly and efficiently. In this paper, based on the integrated data, combining with the platform construction and visualization of GIS, the visualization and analysis system of the cigarette sales data is designed and implemented to provide data display and analysis platform for the enterprise data analysis and policy support.

Key words: ETL; KETTLE; data integration; GIS; visualization

卷烟作为一种特殊消费品, 国家实施严格的计划管理. 但是随着市场化变革的深入, 烟草行业要更加注意发挥市场需求对卷烟营销的拉动作用, 需要解决市场需求持续变化与计划供应之间的矛盾, 逐步实现, 在总量控制的前提下, “市场”要什么“计划”就给什么, “市场”要多少“计划”就给多少, “市场”什么时候要“计划”就什么时候给, “计划”与“市场”要紧密衔接、高度融合.

烟草产业是贵州“五张名片”和“八大支柱产业”之一, 在经济社会发展中占有重要的地位, 是贵州省财税增长、农民增收的重要来源, 是广大贫困地区脱贫致富的重要依托产业. 近年来, 贵州烟草企业信息化

建设取得了显著成绩, 伴随现代营销体系的逐步优化升级, 尤其是最近几年网上订货、网上营销和现代零售终端建设, 烟草营销逐步进入大数据时代. 虽然分布广泛的营销网络每天都会产生大量的数据, 但目前烟草商业企业并没有形成一套系统的、科学的数据处理方式和数据决策模式^[1-3]. 为此, 贵州烟草要不断提升信息化能力积极响应“互联网+”、大数据、云计算、可视化等现代技术和管理手段, 全面谋划贵州省烟草商业信息化各项工作, 推进烟草产业与信息化深度融合. 这样的信息化将有助于解决烟草行业面临的“市场”与“计划”之间的突出矛盾.

数据抽取(extraction)、转换(transformation)、

① 基金项目: 国家自然科学基金(41471323); 中国烟草公司贵州省公司科学研究与技术开发项目(合同号 201407)

收稿时间: 2016-04-12; 收到修改稿时间: 2016-06-21 [doi:10.15888/j.cnki.csa.005592]

加载(loading)(ETL)是多源异构数据集成的有效方法。目前,越来越多的厂商致力于 ETL 工具的研发,如 IBM 的 Datastage、甲骨文的 OWB(Oracle Warehouse Builder)和 ODI(Oracle Data Integrator)、KETTLE 等。其中, KETTLE 使用较为广泛^[4]。本文利用数据分析工具 KETTLE 构建了高效多源异构数据 ETL 解决方案,对贵州现有的卷烟营销基础大数据进行抽取、转换、加载,形成分析型数据,快速高效地实现了数据集成,为卷烟销售大数据可视化系统及数据挖掘提供了优化的数据源。同时,采用 GIS 平台可视化技术,设计并实现了卷烟销售大数据可视化分析系统,优化后的数据进行可视化、分析和挖掘,为企业决策提供支持。

1 数据抽取、转换、加载技术

1.1 ETL 过程模型

ETL 用来描述将数据从来源端经过抽取(extract)、转换(transform)、加载(load)至目的端的过程,使这些数据成为 BI(商业智能)系统需要的高质量的数据,是构建数据仓库、大数据分析、商业智能的重要一环^[5-7]。

ETL 包含了三方面:一是数据抽取,将数据从各种原始的基础数据中读取出来,这是所有工作的前提;二是数据转换,按照预先设计好的规则将抽取得数据进行转换,使本来异构的数据格式能统一起来;三是数据装载,将转换完的数据按计划增量或全部导入到数据仓库中^[9,10]。

1.2 KETTLE

KETTLE 是“Kettle E.T.T.L. Environment”只取首字母的缩写,是一款国外开源的 ETL 工具,纯 java 编写,可以在 Window、Linux、Unix 上运行,绿色无需安装,数据抽取高效稳定。KETTLE 支持文本文件、数据表、各种商业或免费的数据库引擎等多种输入输出格式, KETTLE 实现对数据的抽取、转换以及加载的核心是作业(job)以及转换(transformation)。ETL 活动是一个四元组 $A=(ID, I, O, S)$, ID 是活动标识符, I 是输入模式的集合, O 是输出模式的集合, S 是一个或多个扩展的关系代数表达式,表示每个输出模式的语义^[9-14]。图 1 是 KETTLE ETL 的概念模型。

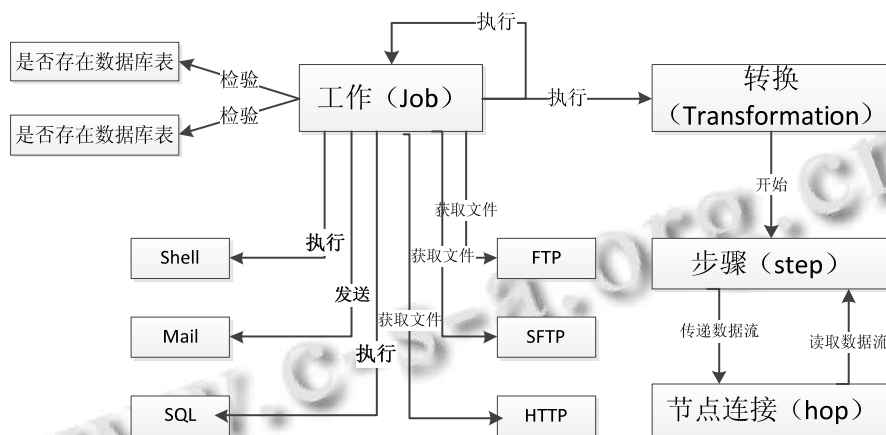


图 1 KETTLE ETL 概念模型图

基于 KETTLE 的 ETL 工作流程图如图 2 所示。KETTLE 主要包含 Spoon、Kitchen、Pan 三个主要组件。Spoon 是一个图形化的界面,是创建 Job 和 Transformation 的数据集成环境,制定整个 ETL 的解决方案;Kitchen 用于调用作业(Job)流程;Pan 调用转换(Transformation)流程。通过 KETTLE 中设计的作业和转换,有两种方式存储:资源库和文件,通过加载作业与转换流程文件,可以方便地实现对作业、转换流程的优化与重新定制,提升 ETL 处理效率,改善数据

质量^[9-14]。

2 数据 ETL 处理

2.1 基本数据的获取

贵州各地区销售的卷烟细分品牌共有 345 个,来自全国 20 个省、自治区、直辖市和国外,国外品牌共有 30 个,国内品牌 315 个。贵州各地区销售的国内卷烟品牌主要分布于长江流域,共 239 个卷烟品牌来自云、贵、川、鄂、湘、皖、苏、沪共 8 个长江流经的

省、市、自治区。烟草品牌来源复杂，区域销售情况参差不齐，实现市场引导和计划营销势在必行。

烟草企业各业务层面的信息系统在支撑企业运营管理的同时，积累了大量的基础数据。伴随现代营销体系的逐步优化升级，尤其是最近几年网上订货、网上营销和现代零售终端建设，烟草营销逐步进入大数据时代。在本文中，我们的实验数据来自贵州省某烟草销售公司的品牌数据、客户零售门店相关数据、订

单数据、消费者信息数据、以及各个零售户的库存及销售数据等。表 1 给出了经过整理以及去重表简化之后的原始数据表的字段。从表中我们可以看出，因为基础数据来自不同的系统，数据结构差异很大，字段中英文都有，数据冗余度大。烟草营销数据进入大数据时代，海量的数据，如果没有好的数据表结构，后为后期的数据管理、数据分析以及数据挖掘造成困扰。

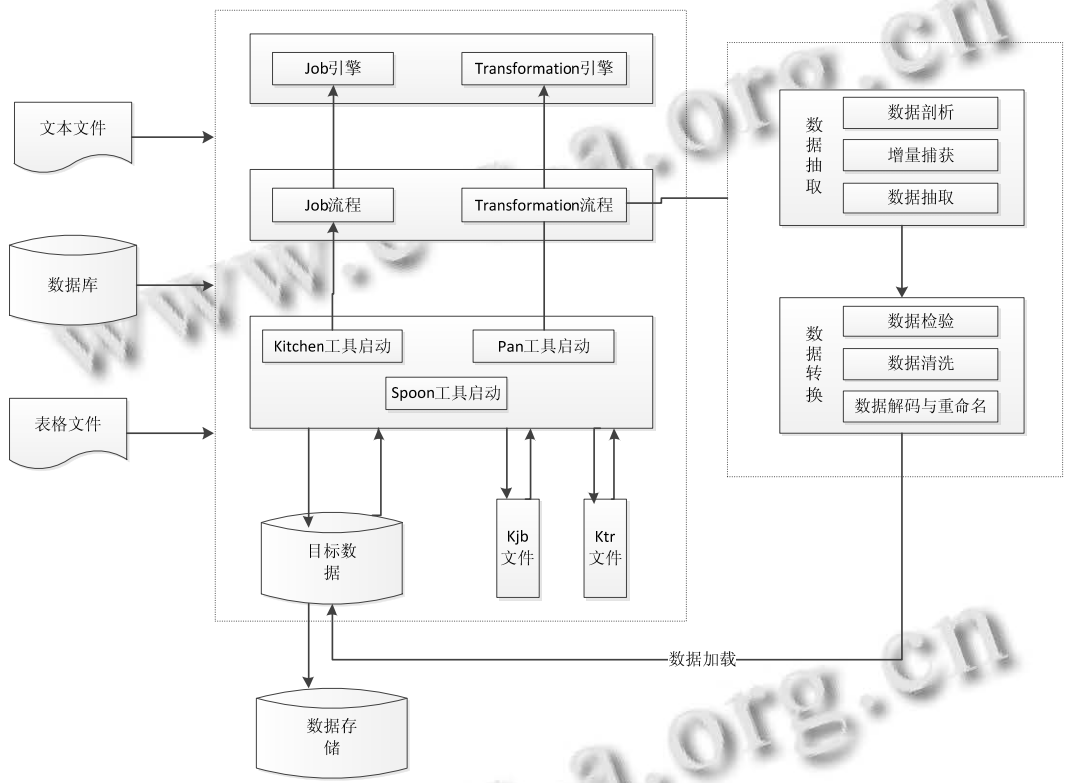


图 2 KETTLER 工作流程图

表 1 原始数据表字段

品牌信息	品牌系列	品牌编码、品牌名称、品牌拥有者编号、是否重点骨干品牌
	品牌拥有者	品牌拥有者编号、品牌拥有者名称、品牌拥有地址、简称、注释
	品牌编码数据字典表	商品编码、商品名称
	重点品牌目录表	时间区间、排名类别、排名、品牌统一码、是否当前信息
消费者信息	消费者信息	消费者编号、消费者卡号、证件类型、地址类型、地址、性别、出生日期、开始吸烟日期、焦油含量、吸食结构、喜好类型、喜好厂家、日吸量、推荐零售户、邮政编码、月收入、学历、职业、当前积分、积分类型、消费者状态、当前等级、购买总量、购买金额、工作单位、口味变换、积分汇总日期、备注、更新日期、更新时间、喜好卷烟、人口类型、是否有意愿购买新品、喜好品牌、消费群体、购买场合、希望吸食品牌
订单信息	订单信息	订单编号、公司、客户编号、类型、状态、付款状态、生效日期、创建日期、创建时间、需求数

		量合计、数量合计、应收金额合计、记账日期、是否已经发货、是否已提交配送、营销中心、营销部、客户经理、业务库编号
零售户信息	零售户	创建日期、组织编码、组织名称、卷烟编码、卷烟名称、卷烟批发价格、客户编码、客户名称、经营地址、经营业态、零售户状态、支付类型、客流量、市场类型需求数量、审核数量、订购数量、价类、TAR_CONT、销售额、需减订、审减订
	客户信息	LICENSE_CODE、CUST_NAME、BUSI_ADDR、MANAGER
库存信息	零售客户扫码库存	LICENSE_CODE、ITEM_CODE、QTY、DATE1、TIME1、COM_NAME
销售信息	客户扫码销售	SALE_ID、NO、LICENSE_CODE、PUH_TIME、ITEM_CODE、QTY_ORD、PRICE、AMT、COM_NAME

分析烟草营销数据的特点，按照“准确定位，有机对接，突出品牌，全面提升”的工商协同营销总体要求和“精确信息，精确投放，精细管理”的精确营销工作

要求，在原有的数据表结构的基础上，本系统设计数据库表结构如图 3 所示。

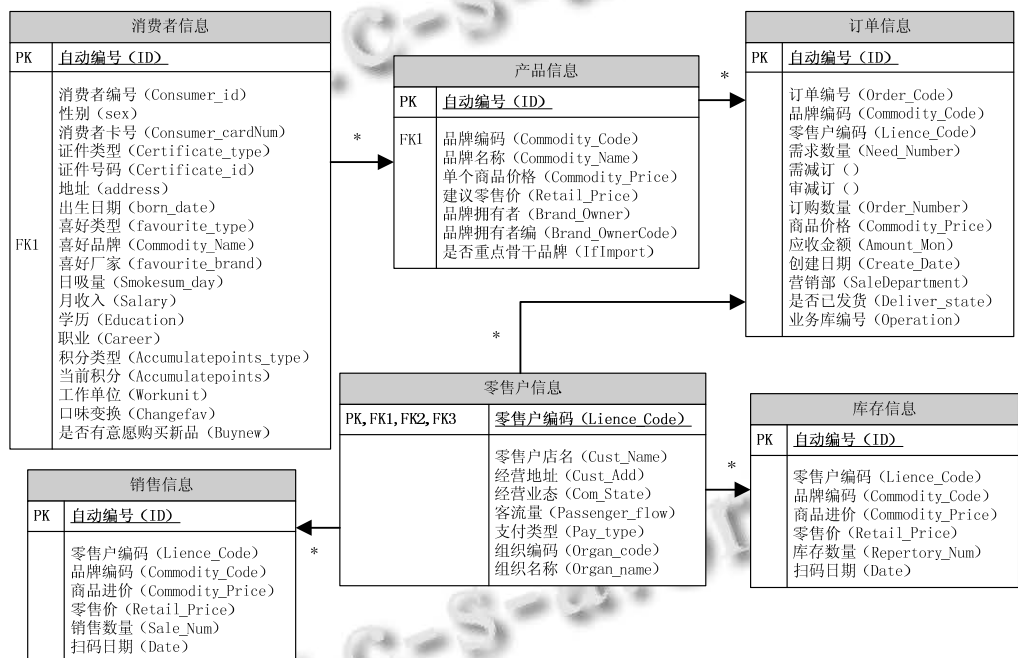


图 3 系统数据库表结构

经过对基础数据的分析，初步设计了产品信息表、订单信息表、零售户信息表、零售户库存信息表、零售户销售信息表以及消费者信息表等存储经过 ETL 转换后的数据。其中产品信息表记录卷烟的品牌商品信息，每个卷烟品牌是一条记录，有独立的编码，是卷烟销售数据库的基础表；订单信息表用来记录每个零售店的订单信息，单个零售户的每个卷烟品牌的订单形成一条记录，可以和产品信息表和零售户信息表关联，获取品牌和零售户的信息；零售户信息表用来记录零售门店的信息，每个零售户是一条记录，有独

立的编码，是卷烟销售数据库的基础表；零售户库存信息表用来记录各个零售用户的卷烟库存信息，每个品牌某个时间节点的扫码库存形成一条记录；零售户销售信息表用来记录各个零售门户的卷烟销售信息，每个品牌某个时间扫码的销售形成一个记录；消费者信息表用来记录登记的消费者的基本信息，一个消费者形成一条记录，是基础表，可以和品牌信息表关联，得到消费者的个人偏好。

2.2 基础数据的 ETL-KETTLE 的实现

针对贵州卷烟营销大数据抽取分析的应用需求，

需要将积累的基础数据按照新的数据设计导入新库中,此过程可能涉及到表结构不一致、大数据量(千万级,甚至上亿)、数据冗余等情况,处理过程包对异构数据的抽取、清洗等工作.本文中利用开源的 ETL 工具 KETTLE 实现了数据的 ETL 过程,部分复杂工作通过 Java、JavaScript 扩展 KETTLE 工具实现.

(1) 作业流程

根据设计的表结构,本文确定了产品维度、零售门店维度、订单维度、销售事实、库存维度、消费者维度六个 ETL 处理流程,实现对整个卷烟销售数据集成的控制.数据处理 ETL 处理流程图如图 4 所示.按照流程,经过产品维度作业流程处理后形成产品信息表,经过零售门店维度作业流程处理后形成零售门店信息表,经过订单维度作业流程处理后形成订单信息表,经过销售事实作业流程处理后形成销售信息表,经过库存维度处理后形成库存信息表,经过消费者维度处理后形成消费者信息表.

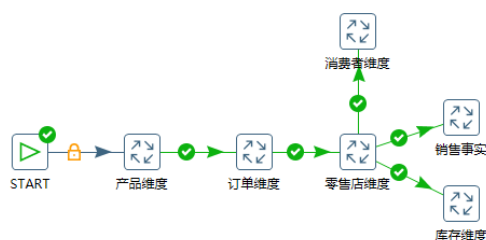


图 4 数据处理 ETL 作业流程图

各个维度与事实作业的二级作业流程图如图 5 所示.由于该 ETL 流程会被重复使用,数据流经每个二级作业流程时,首先检查数据库服务器中相应流程的表是否存在,如果存在,执行更新原表数据的转换流程,如果不存在,需要在服务器端的数据库中创建新的表格,执行插入数据的转换流程.

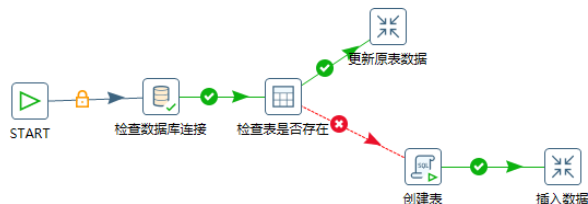


图 5 各个维度的二级作业流程图

(2) 转换流程

作业流程是对整个 ETL 处理的总体规划,协调执

行过程和相关依赖性的 ETL 活动.作业流程中对基础数据和目标数据的映射关系的实施是通过转换流程实现的,经过排序、去重复数据、字段选择、平滑去噪等转换操作,最后实现加载到相应的维度或事实信息表中.以订单维度作业流程为例,相应的插入数据转换流程如图 6 所示.经过订单维度作业流程需要去除无用重复的冗余字段,选择有效和对后期数据分析有意义的字段,对其中格式不统一的字段进行转换(这个维度中主要是对日期字段转换为年/月/日的格式),同时与基础数据中的零售户信息进行零售户编码的检校,与产品维度表中的品牌编码及品牌名称进行检校,校验通过形成有效数据插入新的数据库订单表中,供后期分析处理使用.

3 可视化系统构建

本文结合 GIS 空间数据可视化管理及分析技术,开发了贵州省某烟草公司卷烟销售数据的 GIS 分析平台,可以对前面经由数据分析工具 KETTLE 进行 ETL 处理后的销售、采购和库存等为多种维度数据进行初步的分析挖掘和分析成果可视化展示,为提高卷烟销售水平提供了决策支持依据.

3.1 可视化系统构建需求分析

卷烟营销信息具有数据量大、结构复杂、种类繁多、内容丰富等诸多特点,利用 KETTLE 对贵州省某烟草销售公司积累下来的大量基础数据进行 ETL 处理后,形成了新的集成的可用于分析的集成的数据集.同时,地理空间区域对烟草的销售有着不可忽视的影响.本文根据各省、市、县卷烟销售公司和烟厂的历史销售数据,结合 GIS 空间数据可视化管理及分析技术,建立了烟草销售分析 GIS 系统,实现卷烟购销存分析、品牌分析、市场分析、贡献度分析和趋势分析等数据分析工作.

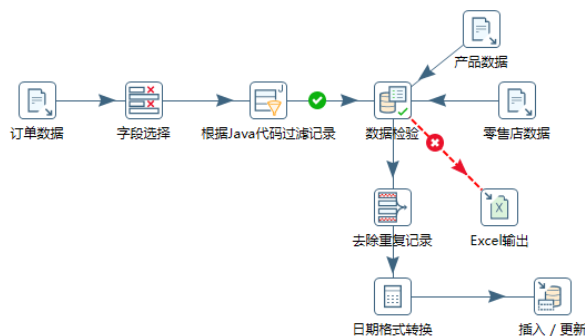


图 6 订单维度插入转换流程图

3.2 平台设计

根据需求,平台共分为数据管理、报表分析、R 数据分析以及 GIS 数据可视化四个功能模块。其中数据管理模块对卷烟营销数据进行增、删、改、查等基本操作;报表分析模块,针对不同的数据,形成报表分析,包括有品牌占比图、各季度品牌重要性对比图、销售走势图等,提供决策支持;R 数据分析,基于 R 统计计算和制图的功能,提供查询与预测分析;GIS 可视化模块,结合了 GIS 可视化分析的功能,

地图与卷烟销售数据的结合,实现了数据上图可视化显示,分地区数据分析,生成卷烟销售数据专题图。

3.3 平台实现

本文采用了基于 Java 语言的 J2EE 框架开发,采用了 Struts + Hibernate + Spring 的轻量级框架开发了 B/S 系统。图 7 给出了报表分析中的分季度各品牌烟草重要性对比图,图 8 是 GIS 可视化中分区域统计烟草销售分布饼状图。



图 7 分季度各品牌烟草重要性对比图

4 结语

本文针对贵州省某烟草公司卷烟销售营销数据分析和挖掘研究中数据处理和平台开发的需求,基于 ETL-KETTLE 的数据集成的过程及其关键技术,制定了针对其现场业务数据优化处理的 ETL 模型,对数据进行了抽取、转换、加载,形成了分析型数据,完成了数据的初步抽取和集成实验。ETL-KETTLE 能够高效地完成批量数据的迁移工作,为后续的数据集成分

析、数据挖掘提供了可靠的数据支撑。

在 ETL-KETTLE 处理后的集成数据的基础上,结合 GIS 空间数据可视化分析的技术,本文设计并实现了卷烟销售大数据可视化及分析平台,对卷烟品牌、销售、采购和库存等多种维度数据分空间区域进行了初步的统计分析挖掘和分析成果可视化展示,从而为制定更优化的企业卷烟营销购销存策略,提高卷烟营销的整体水平提供决策支持依据。

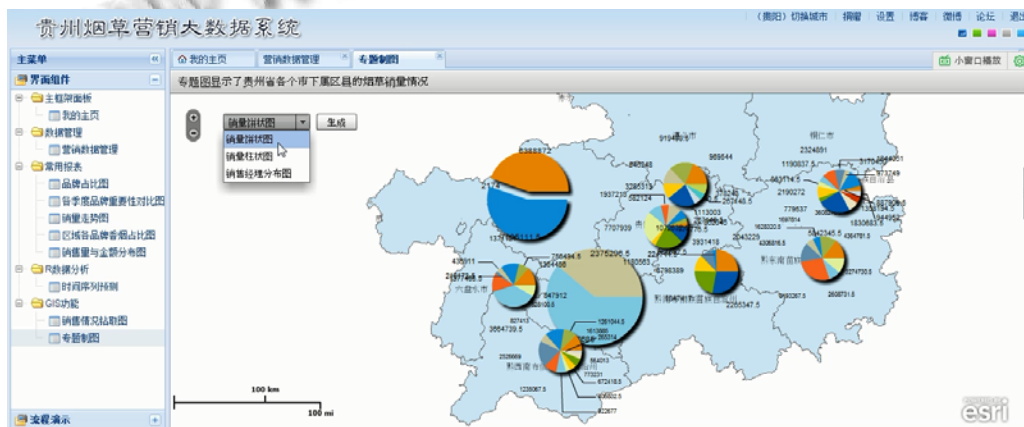


图 8 基于 GIS 的烟草销售饼状图

参考文献

- 1 姚丹丹.基于数据挖掘的红塔集团数据库营销系统的研究与实现[硕士学位论文].杭州:浙江理工大学,2013.
- 2 高山等.基于大数据思维的卷烟零售店微商圈分析研究.电子技术与软件工程, 2015,2:209-212.
- 3 聂佳,等.巴蜀中医药古籍医案数据挖掘系统构建及应用.中国中医药图书情报杂志,2015,4:13-15.
- 4 武剑.数据集成平台中 ETL 的研究与设计[学位论文].北京:华北电力大学,2007.
- 5 陈荣鑫,付永钢,陈维斌.基于 Pentaho 的商业智能系统.计算机工程与设计,2008,9:2407-2409.
- 6 余海钊.基于 Pentaho 的水泥企业商业智能信息系统研究与开发[硕士学位论文].长沙:中南大学,2010.
- 7 韩冰.基于商业智能的高校辅导员决策支持系统研究与应用[硕士学位论文].长春:东北师范大学,2013.
- 8 卢达.警用信息智能分析系统的研发与应用[硕士学位论文].北京:华北电力大学,2013.
- 9 尹晓楠,邹晓涛,张冬.基于 kettle 的北京市水务普查数据的提取与转换.中国水利,2013,21:57-59,42.
- 10 崔有文,周金海.基于 KETTLE 的数据集成研究.计算机技术与发展,2015,4:153-157.
- 11 Van DJ, Bouman R. Pentaho Solutions: Business Intelligence and DataWarehousing with Pentaho and MySQL. John Wiley & Sons Inc., 2009.
- 12 钟华,冯文澜,谭红星,黄涛.面向数据集成的 ETL 系统设计与实现.计算机科学,2004,31:87-90.
- 13 刘充.基于 KETTLE 的高校多源异构数据集成研究及实践.电子设计工程,2015,23:23-26.
- 14 崔友洋,崔有文.基于 ETL-Kettle 的中药饮片企业商业智能研究.产业与科技论坛,2014,22:47-50.