文章编号：1007－5399（2014）05－0017－03

数据挖掘技术在报刊专刊中的研究与应用

李晓巍，侯 杰

（黑龙江省邮政公司，黑龙江 哈尔滨 150006）

摘 要：文章介绍了数据挖掘的概念及意义，从数据模型、模型转换、数据加载和实施过程等方面探讨了数据挖掘及分析在邮政报刊专刊中的应用步骤，并根据报刊专刊分析成果的应用效果提出了优化报刊专刊业务的发展建议。

关键词：报刊专刊；流转额；期发份数；订期结构；日常收订；对比分析法

中图分类号：F61 **文献标识码**：A

邮政报刊专业已于 2006 年完成业务与技术变革，目前报刊专业采用的全国业务数据集中存储处理模式已适应专业的发展要求，为报刊专业可持续发展，打造邮政报刊品牌做出了突出贡献。但是目前报刊专业市场竞争日益激烈，邮政传统报刊业务增长缓慢，竞争对手日益强大，各地报刊社自办发行不断增加。在如此严峻的形势下，邮政急需发挥自身优势，在保持现有业务增长的同时，转变经营理念，推行精细化、个性化营销，加大对内外环境因素及竞争对手的分析，加强营销手段，从报刊专刊入手，采取逐一分析、各个击破的原则，为开拓报刊发行业务潜在市场，提供差异化服务与精细化营销打下坚实基础。

1 数据挖掘技术简介

数据挖掘是指从大量、不完全、有噪声、模糊、随机的数据中，通过设置一定规则和算法，提取人们事先未知而又潜在有用信息的过程。数据挖掘可应用于任何行业领域，只要采集信息合理，积累数据足够，数据挖掘技术就会在不经意间被运用，从而得出更多有价值的数据信息或资料，为企业发展提供更好的决策技持。数据挖掘环境示意图如图 1 所示。

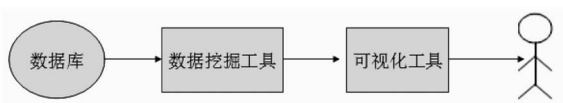

图 1 数据挖掘环境示意图

2 报刊专刊中的数据挖掘

报刊专刊分析的数据挖掘是从报刊生产数据库中把专刊信息按用户功能需求和结构要求选择挖掘算法、预测实体和业务数据，建立适当的关系数据挖掘模型或联机分析处理数据挖掘模型，最终发现并提取隐藏其中的信息或知识的过程，主要经过数据建模、数据模型转换、数据加载、数据实施过程四个阶段，如图 2 所示。

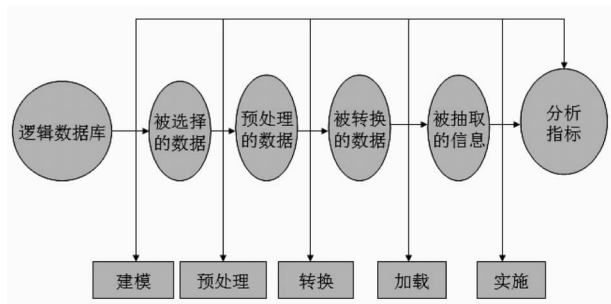

图 2 报刊专刊中的数据挖掘过程

2.1 数据模型

报刊专刊数据模型的设计采用面向主题自上而下的设计方法，通过客户、订期、渠道等多维概念获取所需数据，其中时间和地域是最基本和关键的维度，其设计方法经历了六个过程，分别是：业务理解、数据理解、数据准备、建立模型、模型评估和结果发布。

2.1.1 业务理解

理解项目背景，即报刊专刊的远景和发展目标，理解每个分析指标的内涵和外延。

2.1.2 数据理解

着手对源数据进行收集，从报刊生产数据库中发现专刊隐藏的信息或探测臆想的数据子集。

2.1.3 数据准备

在源数据的基础上运用建模工具建立最终的数据集。数据准备可能会重复多次，主要目的是使用建模工具传输和清洗数据，包括表、记录和属性等。

2.1.4 建立模型

在数据准备基础上，建立报刊专刊数据模型，是一个多次校对、多次审核的过程，通过选择和应用多种建模技术，力求达到最理想的参数值。

2.1.5 模型评估

基于数据分析的观点建立一个或多个高质量模型。在配置这些模型前，最重要的是对已经建立的模型进行彻底评

估，并回顾建造模型的每个步骤，确保业务目标完成。

2.1.6　结果发布

根据用户需要可能只是简单地创建一个报表，或是一个重复、复杂的数据挖掘过程。大多数情况下，模型应该由用户而不是数据分析师来配置，重要的是让用户预先理解所要执行的配置动作，从而让用户使用创建的模型。

2.2　数据模型转换过程

为了更好地完成报刊专刊的决策分析任务，需要把普通面向交易的数据库转换为面向分析的数据库。面向分析的数据库侧重查询和分析，由于二者的应用不同，所依赖的数据模型也有所不同，分析数据库主要采用多维模型，而业务数据库采用关系模型，所以数据抽取过程是一次模型的转换。为了区分二者，将分析数据库专门进行报刊决策分析的系统称为报刊专刊分析数据库，它是从报刊管理系统中按关联规则抽取的生产业务数据，专门完成分析决策的功能。报刊专刊分析数据库主要采用 OLAP 技术及思想，分为服务器和前端产品，OLAP 服务器提供多维数据的存储，为前端报表或控件提供易于分析的直观多维数据，是数据与信息之间的桥梁。

2.3　数据加载

数据加载即数据抽取、转换和装载的过程，它是构建数据分析的重要环节。目前邮政报刊系统采用全国数据集中存储的模式存储生产业务数据，而报刊专刊分析也将采用全国数据集中存储的方式进行存储，后者的数据是从生产业务数据库中经过数据加载过程而生成，因此这个过程是决定报刊专题分析成败的关键因素。

2.3.1　数据抽取

数据抽取需要在调研阶段做大量工作，从业务需求角度出发，根据报刊生产业务数据库的数据源，如卡片、要数、结算、分发等数据，按照时间戳的方式和一定规则把生产业务数据抽取到数据分析库中。

2.3.2　数据转换

从业务系统到操作型数据存储做清洗，将脏数据和不完整数据进行过滤，在从操作型数据存储到分析库的过程中转换，进行业务规则的统计、计算和聚合。

2.3.3　数据装载

以增量并发方式把生产库中的数据按分析要求建立多维度存储，提高统计分析效率。

2.4　实施过程

2.4.1　服务定制过程

报刊专刊分析数据模型的建立主要针对不同分析结果采用不同转换模式，对于相对固定的统计数据则采用在数据迁移时进行汇总运算的方式，从而在存储数据时节约开销。另外，为了增强交互性，采取服务定制方式，用户可以通过友好的服务定制界面，输入统计的起止时间、产品类别、区域范围和环节数据等信息，把服务定义信息存储于分析数据库的服务定制表中，后台程序根据用户编制的服务信息，从相关的业务表中建模、转换模型、加载数据、统计生成分析报告，供用户次日查看。用户对报刊专刊进行分析，只需在服务定制界面中增加要分析的专刊代号、年度信息，即可完成服务订制过程。

2.4.2　分析报告的形成过程

通过定量计算与分析，以报告形式展示分析结果，从而为业务提供一定的参考依据。决策分析人员只需在相应的功能上点击查询按钮就可以把系统生成的、针对专刊的全方位分析报告展示出来。

报刊专刊分析主要从单日累计流转额、收订方式分析、日常收订与退订、订期结构、长短期、订户数量分布、客户群体分析等方面进行统计分析，并自动形成分析报告。分析报告分为单项分析结论和整体分析结论，对应的分析方法是对比分析方法和 k-means 分析方法。报告由三部分组成：第一部分标题区是专刊简介，简单介绍报刊名称、单价、产品分类等基本信息；中间部分是系统自动分析生成的九张报表，每项报表根据分析数据和图表自动生成单项分析结论；最后是对整个报表的总体分析结论。

2.4.2.1　对比分析方法

对比分析法通常是把两个相互联系的指标数据进行比较，根据分析结果并结合图表展示和说明研究对象的规模大小、水平高低、速度快慢以及各种关系是否协调。在对比分析中，选择合适的对比标准是关键步骤，选择合适才能做出客观评价，选择不合适可能得出错误的评价结论。在专刊分析报告中，主要采用对比分析方法对每个报表进行统计分析，并根据分析数据形成单项分析结论，其中分析基准主要依据业务沉淀的业务指标经验值或系统根据数据结果计算的均值，单项分析结论在分析基准的基础上给出每个分析样本在整体分析指标中所处的位置或百分比，以准确、精炼的文字做出总结性的单项结论。

单日累计流转额分析：展示当年和上两年的流转额绝对值，按日进行统计和展示，汇总每月流转额，采用分级绘图对不同年度进行对比，分析三年收订的高峰，形成一定的经验值，为报刊业务有针对性地开展年度收订提供依据。

收订方式分析：收订方式统计数据是当年与前两年不同收订方式的对比，分析专刊收订渠道在整体收订方式中的占比，体现收订渠道在各年所占的百分比及发展趋势，加强各类订阅渠道的宣传力度，以提升订阅量。

年度日常收订量与退订量分析：由本年度新增和退订两个报表组成，分区域展示区域及客户对专刊的贡献率，按区域和单位展示流转额排名。

订期结构分析——全年期发份数：展示当年和前两年的月汇总期发份数绝对值，利用体积柱状图形象地展示出三年的期发份数，整体分析专刊发展形式。

订期结构分析——订阅长短期分析：对当年和前些年专刊在各省或是本省各地市各订期的订阅份数进行汇总，分别针对各省每年不同订期的数值进行对比展示，为业务提供订期结构在总份数增长的情况下，短期订阅向长期订阅转移的情况，进而分析客户稳定程度。

订户数量分布：对三年各地区的订户数量进行汇总，对比展示各年各省订户数量的变化情况和各省之间同年订户数量的对比，比较专刊区域客户的贡献率，从而有针对性地开发潜在地区的订户数量。

排名前 100 位的订户：列出三年中大客户的相关信息，分年度对各省数据进行汇总，并展示饼状图，分析订阅此报刊的客户所在单位及地址，从而分析出客户喜好并开展针对性营销。

客户群体分析：由于客户群体分析需要对所有客户所属群体进行多级细分，所以在展示上也进行分层显示，主报表内只展示大类细分，通过链接方式进行二级和三级分类的细分。统计三年的分类流转额汇总，绘制体积柱形图展示，从整体上分析各行业为专刊的贡献率，从而针对行业进行精准营销。由于分析重点在于对比各省之间的客户群体差异，不再进行年度对比，因此分别对各年度数据进行展示，从省和分类角度进行体积柱形图展示。

结合报刊专刊分析形成的单项分析报告，制定有针对性的营销策略，周密安排多渠道建设，改变报刊收订人员坐等用户来窗口订阅的方式，改进对目标客户的宣传策略，增加各项促销活动。为营造专刊的收订氛围，在所有网点悬挂报刊收订宣传条幅、张贴活动海报，并在当地报纸、电视台刊播公告。各级投递部门组织人员深入社区、进校园，有针对性地开展现场收订营销活动，从而对专刊进行有效促销。

2.4.2.2　K-means 分析方法

聚类分析也称群分析、点群分析，是研究分类的一种多元统计方法，是将样本个体或指标变量按各自特性进行分类的一种统计分析方法。聚类分析的任务有两个：一是寻找合理的度量事物相似性的统计量，二是寻找合理的分类方法。报刊专刊分析主要利用 K-means 方法对单项分析结论进行总体分析、分类，从而形成本次分析的最终结论。

具体做法是：将全国 31 个省的分析样本数据随机分为高、中、低三个等级，通过平均值和标准差计算得到每一等级的中心，再按欧氏距离计算方法得到每个样本重新分配到距离它最近等级的质心，然后计算分配到每个等级样本的均值向量，最终经过多次递归直到三个等级的质心点不再发生变化或准则函数收敛时即完成本次聚类处理。最终将各方面较类似的省份划分在同一分类中，并根据它们与分组平均值间的大小差别判断分组特点，给予各分类相应的发展建议。

3　报刊专刊分析成果的应用效果及业务管理建议

3.1　报刊专刊分析成果的应用效果

目前本统计分析报告已经在黑龙江省邮政报刊发行局上线使用，有针对性地对十多种报刊专刊的发行情况进行统计分析，取得了一定效果，主要表现在以下方面。

自动生成报告，节省了人员和时间成本。按原处理方式，需要多人采用手工处理方式进行计算，分析九张报表，现在只需在服务定制功能中增加要分析的报刊，系统就会自动生成报表及分析报告，节省了时间和人员的工作量。

联邦快递在西班牙新建处理站

为拓展网络覆盖范围，联邦快递近日在西班牙新建处理站点，其中一个位于马德里的赫塔费，另一个位于巴塞罗那，此举为当地创造了 58 个新的就业岗位。

2013 年，公司已经在西班牙开设了 11 个处理站点。目前，联邦快递在西班牙共有 14 个处理站，为当地提供了 140 个就业岗位。在过去两年半的时间里，联邦快递在欧洲新建了 100 余个处理站。

在马德里和巴塞罗那新建处理站后，公司可以在 1～2 个工作日内为 90％以上的当地企业提供投递服务。

上述两家处理站提供一系列服务，如目的地为欧洲的次日递服务，投递时限为两天的洲际运输服务等。此外，公司还在欧洲、美国及亚洲等地提供资费较低的经济型服务。

在近日召开的新闻发布会上，联邦快递南欧区副总裁以及西班牙地区运营经理纷纷对公司在西班牙和欧洲地区的发展形势进行了展望。

（兰翔英　译）

报刊专刊分析为刊社提供了更优质的服务，通过系统自动生成统计分析报告的形式能够为刊社提供更加准确、快速的服务，也可以为刊社开通服务定制功能，使刊社亲自使用专刊分析功能订制服务，查看系统自动生成的分析报告，促使邮政与刊社的合作更加融洽。

3.2　专刊管理建议

根据以上分析，本文对报刊专刊分析结论提出以下优化建议。一是精细化发行。通过分析订阅数据，查找目标读者群，刊社与邮政共同宣传，合力开发大客户，开展集订分送等业务，拓展报刊订阅市场。二是开拓发展渠道。采取全国通用报刊专用单据、文化礼品的方式，以礼品营销实现立体化报刊销售渠道拓展。

参　考　文　献

1　Jiawei Han，Micheline Kamber. 数据挖掘概念与技术. 北京：机械工业出版社，2001

2　陈封能，斯坦巴赫，库玛尔. 数据挖掘导论. 北京：人民邮电出版社，2010

3　罗永辉，陈明亮. 商业智能的“操作性和提升性”转换－商业智能（BI）的三维框架. 工业技术，2010，6

4　梅长林，范金城. 数据分析方法. 北京：高等教育出版社，2006

收稿日期：2014－03－19

作者简介：李晓巍（1977～），男，黑龙江七台河人，高级工程师，主要从事邮政企业信息化建设研究；侯杰（1978～），女，黑龙江哈尔滨人，高级工程师，主要从事数据分析研究。