数据时代数据挖掘方法及改进应用于零售业中的实践研究

罗嗣扬
上海英夫泰尔克软件开发有限公司　上海　200233

摘　要：最近几年，发生了以物联网，云计算，大数据等为主要特征的数据革命。在消费者，企业，各经济领域都在持续发掘数据潜力的情况下，这一充分触及，分析并利用高价值机遇的前瞻性途径通过多种途径创造了价值。文章主要针对数据时代零售业在利用大数据，数据挖掘技术方面存在的诸多问题进行分析，并且针对这些问题的应用展开了相关的研究。

关键词：数据挖掘；零售业；数据时代

Research on the practice of data mining method and its improved application in retail industry in data age

Siyang Luo
Shanghai Infortech Software Development Co., Ltd，Shanghai，200233，China

Abstract: In recent years, there has been a data revolution characterized by the Internet of Things, cloud computing, and big data. This forward-thinking approach to reaching, analyzing, and exploiting high-value opportunities creates value in multiple ways as consumers, businesses, and all sectors of the economy continue to tap the potential of data. This paper mainly analyzes the problems existing in the use of big data and data mining technology in the retail industry in the data era and carries out relevant research on the application of these problems.

Keywords: Data mining; Retail industry; The Age of Data

引言

在信息剧增的今天，信息过量差不多成了每个人都要面临的一个难题。如何及时地从其中发现有用知识而不是淹没在信息中，需要一种全新的办法对这些大量数据进行处理。因此人们将统计学中的数据库，机器学习和其他技术相结合提出了数据挖掘的方法以解决这一困难。数据挖掘与知识发现正是解决这类问题的一种有效途径。在理论和应用研究日益深入的情况下，它正日益显示出旺盛的生命力。数据挖掘是一种从数量大，不全面，带噪声，模糊，对随机数据，从中抽取出隐含的，人们预先不知道的，却是潜在的有用信息与知识。数据挖掘历史不长，从 20 世纪 90 年代开始就得到了迅猛发展。数据挖掘作为一门交叉学科聚集了来自各个领域的研究人员，特别是数据库，人工智能数理统计，可视化和并行计算的研究人员以及工程技术人员，属于多技术融合的产物[1]。

一. 零售业的特点

零售，就是将货物或服务直接卖给最终消费者。所以零售业介于生产者和消费者或者批发企业和消费者中间的一个环节，它所面向的目标就是最为广泛的消费者。这一界定确定零售业的主要特征是：

（一）从顾客方面看，个人消费者占绝大多数，该群体人数多、消费水平不均衡、多集中在商场进行消费、和零售商之间关系中断。

（二）在货物方面，除专门特卖店外，一般零售商包括货物种类庞大；零售商以经销、代销和联销为主要销售形式。

（三）在供应商方面，所涉货物来源、掌握的资料也非常多。

二. 零售业应用数据挖掘的背景

结账扫描仪建立的初衷纯为操作之便利，因其具有与人工相比较集中设定货物价格，提高付账速度，精确定价及控制库存所无法比拟的优点。而结账扫描仪所带来的另外一项重要效应在最初并未被重视，即生成了用 GB 乃至 TB 来衡量的海量数据。当今，随着超市信息化建设与应用进程不断加快，条码，电子收款机， POS 系统等一大批信息技术已经遍布各大超市。信息系统 8 益巨大的积累着海量数据，怎样用这些海址数据分析什么东西好销，什么东西难销，什么顾客适合什么东西，以

及货物间怎样匹配等都让超市 CIO 头痛[2]。

自 1992 年中国开放零售业后，外资注入国内商业领域的份额逐年上升。并随着莉 WTO 设定 3 年保护期届满。外国资本进入中国会更顺畅，中国零售业也面临空前压力与挑战。仅仅 3 年时间，像家乐福，沃尔玛，麦德龙这样一些国际知名的大财团就已经在中国主要城市连锁店百花齐放。在贴心服务、多样化选择面前，客户也日趋理性、需求个性化。完善客户关系、保留老客户、赢得新客户成了各个零售企业追求的目标[3]。

三. 数据挖掘的主要技术

（一）分类

分类是指对数据类别的判别。首先对数据中已划分类别的训练集进行筛选，并在此训练集中利用数据挖掘分类的相关技术建立分类模型对未划分类别的数据实施分类。使用时，不仅可利用该模型对现有资料进行分析，还可用于对未来资料进行预报。数据挖掘算法的操作方法就是通过对已知分类信息历史数据进行分析，归纳出一种预测模型。在此用来构建模型所使用的数据叫做训练“集”，它一般指已被人们所拥有的历史数据。

（二）估值

分类刻画了离散型变量产出，估值则对连续值产出进行加工，与此同时，分类类别为决定个数估值数量不定。

（三）预测

预测就是构建并利用模型对无标号样本类进行评价，也可以对给定样本中可能存在的属性值或取值范围进行评价。分类与回归为两大预测问题类型，分类为预测离散或者标称值问题，回归则主要应用于连续或者有序值问题[4]。

（四）关联规则

关联分析就是在数据库中查找值之间的关联。2 种常见技术分别为关联规则与序列模式。关联规则就是要找出同一事件发生时不同项目之间的关联，比如单次购买所购买物品之间的关联等。序列模式也和这种模式相似，所寻求的也就是事件间的时间相关性，例如分析商品的价格上升或下降。关联分为完全关联与部分关联两种。例如进行关联规则分析时，甲代表事件“买电脑”，乙代表事件“买鼠标”。甲乙，就说乙与甲是完全相关的，也就是说购买计算机者必然要购买鼠标，如果是局部相关的话，根据粗集理论甲与乙局部相关程度是可利用的，那么购买计算机者就有可能购买鼠标[5]。

（五）聚类

聚类分析算法中输入集为未校准记录集合，即这时输入记录尚未被任何划分。它旨在对记录集合按某种规则进行合理分割，用显式或者隐式来描述各种分类。并通过聚类分析工具，界定了赖以存在的上述规律。聚集前不知将数据划分为若干组或如何划分（根据哪些变量）。所以，集合后应该由熟悉生意的人来讲解这种分群是什么意思。更多时候，单次聚集获得的分群对于自己的生意可能不是很好，此时需要通过删除或者添加变量来影响分群，多次重复后最终获得理想效果。神经元网络及 K —均值等聚集算法较为普遍[6]。

（六）描述和可视化

也就是把资料归约，概化，或者用图形描述等等。

四. 数据挖掘在零售业中的应用

数据挖掘产生于商业上的直接需要，在很多领域具有广泛使用价值，零售领域就是数据挖掘应用的一个主要方面。数据挖掘技术主要用于零售业：

（一）商品分组布局、购买推荐及商品参照分析。通过对销售记录进行相关信息挖掘，可找出购买某一物品的客户同时购买其他物品的可能性。这些信息可以用来形成定地购买建议或者获得物品的最优分组布局以帮助顾客挑选物品并节约其购买时间以实现激发其购买欲望和提高销售量[7]。

（二）对促销活动效果进行分析。零售业往往采取广告、优惠券、各种折扣、让利等促销手段来达到宣传商品，招徕顾客。只有对顾客有了全面的了解，才能够对促销对象有一个精准的定位，增强针对性和减少活动成本。运用数据挖掘技术能够分析企业应在何时何地，用什么方法以及向哪些人群促销，尽可能避免企业资源浪费。同时数据挖掘还能利用以往有关促销数据找到在将来投资时收益最高的用户。

（三）客户忠诚分析。零售企业往往采取办会员卡和建立客户会员制度等手段对客户消费行为进行追踪。通过数据挖掘客户会员卡信息，能够记录客户购买顺序，并对同一位客户不同时间段所购买物品进行分组，利用顺序模式挖掘能够分析出客户购买趋势或者忠诚度变化情况，并依次调整更新价格及物品花样，从而保留老顾客并吸引新顾客。

（四）顾客细分。顾客细分就是把庞大的消费群体分成若干个小细分群体，同属细分群体的顾客消费特征类似。客户细分能让商户在不同细分群内区别对待顾客。比如经济学里二八法则是怎么分辨出顾客和顾客？只有对数据进行深层次挖掘，才有助于商家在大量顾客中归类找到属于自己的顾客[8]。

（五）交叉销售。零售业与顾客的关系具有持续性与发展性，零售业一般采用如下 3 种方式维系与强化这一关系：1. 尽可能地延长维系这一关系的期限。2 尽可能频繁地和顾客进行成交。3 努力使每一笔交易都能获得最大盈利。所谓交叉销售，就是指向老客户推销新产品或者提供新服务。交叉销售以买卖双方互惠互利为原则，顾客由于获得了更多、更好地满足其需要的服务而受益，商家则由于销售的增加而受益。交叉销售具有商家能更方便地获取老客户更丰富信息等优点，对数据挖掘精度要求高海量数据。一般我们指交叉销售，和初次销售是不一样的。企业拥有的顾客信息，尤其是先前的

购买行为信息可能正在蕴含着决定该顾客下次购买的主要信息。此时，数据挖掘的功能表现在能够帮助商家找到那些对客户购买行为有影响的数据与因素[9]。

五. 数据挖掘的流程及案例分析

数据挖掘过程一般由如下 3 个基本环节组成：首先是业务分析。其次是数据理解的准备和建模。三是评价和开发。以下就英国 Safeway 公司的案例，分析它们是怎样成功地执行数据挖掘项目[10-12]。

（一）商业分析：Safeway 公司年销售量突破 100 亿美金、拥有近 7 万员工、在英国排名第三、提供 34 个服务类型。公司信息部拥有 2 台 System/390 服务器并行完成 DB2 工作，最大一周需要管理 800 万交易和大约 4TB 磁盘储存容量。但是在运营期间，公司管理层发现了运营危机。他们觉得英国市场处于饱和状态，要在其中使用诸如低价，店面较多或者产品种类繁多之类的传统竞争技术已变得越来越难，应该说多数竞争对手在价格和产品范围上旗鼓相当，再加上受土地和扩充成本限制，任何一家都不可能在这一领域拥有比对手更多的资源。所以他们认识到，一定要把重点由产品和店面角度转到顾客角度上来，把顾客当作向导。这意味着企业要更加理解客户个人，要理解他们 600 万名客户进行的每一笔交易和交易之间的联系，要构建面向客户的市场。

（二）数据的理解；准备和建模：Safeway 针对以上问题开始向客户发放会员卡并使用卡片结账，客户可获得各种折扣，这类卡片成了公司 500 个店面收集 600 万条客户信息的“网络”，收集到的信息由公司每周从主数据库中抽取并存入数据仓库（一周约 500GB）。因为这个资料来源简单、标准统一，所以企业对于资料的质量还是比较放心的，而且企业只为了大概知道自己的客户是谁，资料清晰并没有太大的麻烦。所以当数据存入数据仓库后，企业直接把顾客划分为 150 个类别，利用关联规则技术对数据集进行比对，然后列出产品吸引力列表，也就是挖掘出来的关联规则。例如，买烤肉炭的顾客，多买打火机燃料。" 由于企业采用自动化的分析工具，所以这些挖掘工作并不是一次性完成的，而变成企业持续开展的日常事务中的一环。

（三）评价和开发：Safeway 公司经过持续的发掘工作，已经找到许多对于公司决策颇有帮助的资料。例如他们发现某种奶酪产品尽管销售额排在 209 位，但在消费额最大的顾客中，有一些人经常购买这类奶酪。这类顾客是英国 Safeway 公司最不愿意冒犯的顾客。若采用常规分析方法，这类产品也许会迅速撤下货柜但真相。上好这类产品颇为重要。他们又发现 28 个牌子橘子汁里有 8 个格外受青睐。于是公司可以重新布置货架上的陈设，从而使橘子汁销售达到最高。除此之外，当企业知道顾客每消费一次就买什么商品时，也可借助数据挖掘的序列发现功能来检测是否存在长时间频繁购买的情况。然后结合主数据库人口统计数据，Safeway 行销部门便可依据各家各户之特性，如什么季节买什么商品之倾向等等来发信。据资料显示，该公司某年共发送完全客制化邮件 1200 万件，对销量的增长起了决定性作用。

六、结语

当前市面上存在着各种适用于各种商业模式求解的普适数据挖掘系统，但是在现实中它们都不太好用，一般用户很难将它们运用到商业中去。问题是怎样把数据挖掘技术和已有技术进行良好的融合，没有特殊领域商业逻辑和数据仓库技术的融合，数据挖掘分析效果及收益就无法到达顶峰。对系统进行定制、软件供应商与企业之间进行相互沟通、对系统功能进行持续改进与扩展等都能实现某种程度上解决这一难题。

参考文献：

[1] 祝万晨 . 数据时代数据挖掘方法及改进在零售业中的应用与研究 [D]. 安徽理工大学 , 2014.

[2] 汪祖云 . 数据挖掘技术研究及其在零售业中的应用 [D]. 北京工业大学 , 2003.

[3] 刘芳 , 王璐鑫 . 数据挖掘技术及其在零售业中应用的初步研究 [J]. 福建电脑 , 2009(8):2.

[4] 彭虎 , 田俊峰 , 余玛俐 . 数据挖掘技术在零售企业成本分析中的应用研究 [J]. 科技信息 (科学教研), 2007, 000(032):85-86.

[5] 张杰 . 面向零售行业的数据挖掘技术的研究及应用 [D]. 江西理工大学 , 2011.

[6] 陶俊 . 数据挖掘在零售业销售管理中的应用 [D]. 武汉理工大学 , 2011.

[7] 沈富泉 . 数据挖掘技术在商业银行零售业务精准营销中的应用研究 [D]. 厦门大学 , 2011.

[8] 顾美芳 . 数据挖掘技术在零售业客户关系分析中的应用研究 [D]. 苏州大学 , 2006.

[9] 冯瑶 . 基于零售业的数据挖掘技术和关联规则算法的改进研究 [D]. 河北工业大学 .

[10] 任际范 . 数据仓库与数据挖掘技术的应用研究——以家电零售企业为例 [D]. 山东大学 .

[11] 陈竞 . 基于数据挖掘技术的零售业精确营销应用研究 [J]. 中国市场 , 2010(14):3.

[12] 陈竞 . 基于数据挖掘技术的零售业精确营销应用研究 [J]. 中国市场 , 2010.