数据挖掘在超市大数据中的应用

■梁婧婕　曹　婷　南京邮电大学

本文受江苏省大学生创新训练计划项目资助,项目编号 SZDG2015037

摘　要:大数据时代,各行各业汇集了庞大的数据,如何使这些数据得到充分的利用,数据挖掘是最关键也是最基础的工作。在本次研究中,将数据挖掘技术与购物篮思想理念相结合,运用 R 语言,对南京一家超市五个月内的销售数据进行研究分析。具体有 65536 条数据,有 31869 条销售记录,2242 种类商品,将这些商品分为 192 小类商品,基于食品分类规则将该超市食品分为 38 类。此次研究过程如下:首先,运用 R 语言,使用编写字典的方式,对所获取的数据进行清洗,生成结构化数据。然后,在三方面对数据挖掘。一,数据描述性统计挖掘。二关联规则挖掘。最后,用图表的形式展示此次研究的成果。此次烟酒店意义:利用初级数据挖掘的理论支持,帮助企业更好地分析、了解客户,最终赢得客户的竞争是该研究的重要的意义和实际应用价值。

关键词:大数据;数据挖掘;购物篮;超市销售;关联规则

DOI:10.14013/j.cnki.scxdh.2016.07.007

一、研究背景

1998 年的《哈佛商业评论》刊登过这样一个案例,20 世纪 90 年代美国沃尔玛超市中，沃尔玛超市管理人员分析销售数据时发现了一个令人难以理解的想象:在某些特定的情况下,啤酒与尿布这两件毫无关联的商品会经常出现在同一购物篮中。1993 年美国学者 Agrawal 提出关于通过分析购物篮中商品集合,从而找出关联关系的关联算法,并根据商品之间的关系,找出客户的购买行为.Agrawal 从数学计算机算法角度提出了商品关联关系的计算方法 --Apriori 算法。沃尔玛尝试将 Apriori 算法引入到数据分析中,并获得成功,为超市销售产生了开拓性的影响。于是产生了“啤酒与尿布”的故事。

近几年,数据挖掘技术在零售业,电信业,金融业等许多领域得到了广泛的应用。为了更加清楚地了解学习数据挖掘在大数据环境下的应用。此次,我们对数据挖掘中的部分分析功能在零售业(基于一小型超市)的应用做一些粗略的研究与学习,基于关联规则,购物篮,Apriori 算法等分析商品销售状况,探索出更多的类似于啤酒与尿布这样的规则等，辅助决策者了解销售全局,降低库存成本,进行市场分析等。

二、文献回顾

数据挖掘出现于 20 世纪 80 年代后期,90 年代有了突飞猛进的发展。2001 年,GartnerGroup 的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首,并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。美国麻省理工学院在 2001 年 1 月份的《科技评论》(TechnologyReview)提出将在未来 5 年对人类产生重大影响的 10 大新兴技术，其中第 3 项就是数据挖掘。

数据挖掘技术已被广泛的应用于各个领域。在零售业领域，很多大型的零售商都采用了数据挖掘工具进行决策分析，关联规则挖掘已经投入应用领域,交叉管理,库存控制好客户分析设计都是零售业数据挖掘的主要内容。以沃尔玛为例他就采用了 BO 的方案。LuisCavique 的购物篮分析的可扩展算法研究;AndreasMilda,ThomasReutterer 提出了一个改进合作过滤方法以及预测二进制购物篮数据的交叉目录购买情况;HorngJinhChangd 的基于聚类分析和关联规则分析的潜在客户购买行为的期望模型研究;FransCoenen,PaulLeng 的基于分类精确度的关联规则阈值影响等。

国内对数据挖掘的研究较晚,没有形成整体的力量。1993 年国家自然基金首次提出支持数据挖掘领域的研究项目。目前,国内的许多科研单位和高等院校竞相开展数据挖掘和知识发现的基础理论及应用研究。复旦大学一直从事这方面的研究,朱扬勇等把一个应用于特征规则基于差异化的兴趣度定义运用到关联给则中,重新设立了兴趣度;武汉科技大学的张新霞等提出基于统计相关性的兴趣度量；东南大学宋爱波等提出了一种解决规则组合爆炸问题的方法，建立了一个带约束规则挖掘算法的模型,对 Apriorii 算法进行优化。还有其他相关研究。

但是,当前国内零售业数据挖掘工作还处于探索阶段。据了解,许多零售业企业使用收账结账设备获取的相关销售数据,都没有得到充分利用，这些数据本来都可以帮助零售企业实施交叉销售,控制库存,降低库存风险等创造更大的商业价值,却被忽略。所以,我们以南京市一家苏果超市为主体,使用购物篮的思想,从数据的获取,到数据清洗,再到关联规则分析等一系列系统的方法,研究与运用数据挖掘技术。

三、研究对象及方法

本研究所用的超市销售数据来自于南京市某一家苏果便利店的一个月内的月销量数据。数据大约有六万多条。包括商品的单号,商品销售时间,商品名称,销售单价,销售数量,销售金额。其中,部分是一个单号包含一个商品,其余为是一个单号包含多

个商品。所以,本次研究不仅对购买了一个商品的购物篮进行描述分析,同时也对购买多个商品的购物篮进行关联规则分析。

采用购物篮分析方法。购物篮分析就是通过购物篮所显示的交易信息来研究顾客的购买行为，其直观意义就是顾客在购买一种商品的同时有多大的意愿购买另一种商品。研究商品之间的关联规则。这一规则中包含两个参数:支持度(support)和置信度(confidence)。支持度(Support)的公式是:Support(A->B)=P(AUB)。支持度揭示了A与B同时出现的概率。置信度(Confidence)的公式是:Confidence(A->B)=P(A|B)。置信度揭示了A出现时,B是否也会出现或有多大概率出现。

四、数据清洗

随着信息技术的不断发展，各行各业都建立了很多的计算机信息系统,所以也就产生了大量的数据。当需要对数据进行分析的时候,直接获取的数据并不能够直接进行数据分析。主要表现在:数据冗余、数据重复、脏数据等问题。为了使得数据能够有效地支持相关的运作与分析,必须对数据进行清洗与处理,使之成为结构化数据。所以数据清洗也就是各种数据分析如OLAP(关联分析)、数据挖掘的前提与基础。在R软件中,通过建立字典的方式进行数据的清洗。

我们在对超市数据进行数据清洗的方法是构建字典，具体步骤如下：

1.建立链接:用read直接读取数据所在的文件,建立链接。

2.编写字典:根据商品的货号,提取出每一种商品的关键字,定为搜索的字符(searchword)赋予它替换的名称(replacenames),把类似的商品给予相同的名称,如:洗衣护理剂,柔顺剂,都给它附名柔顺剂。其中,忽略商品的生产厂商。(注:因为主要研究方向与生产厂商无关)这样的话,可明确商品类型,依据连锁超市商品分类明细表指标,对所有的商品进行分类(category),如:家居用品,饮料,调料制品,粮食类等总共38种。

3.名称替换:使用for循环语句,按照字典里的关键字对原始数据里所有的商品进行对比,测试,找到相同的赋与替换名称与分类。结果如下(部分)。如果没有搜索到对应的关键字,则用other_names代替。这样,打开清除后的数据文件,查看清洗后的结果,对没有与之相对应的关键字的商品再进行字典的补充,知绝大部分的商品都搜索到与之相匹配的关键字。这样,就完成了字典的编写,与得到清洗后的结构化数据。

4.数据的重组:对于相同单号的数据合并在一起,则为一个顾客购买的商品。加载reshape程序包，把整体的数据打碎(melt),让其回到一个一个数据点的状态,根据观测的id名称和变量名称定为,再根据id名称和变量名称进行重新的组合,将同一个顾客买的所有商品都排列到一行。这里,假定购买最多的一个客户买了20种商品。在每一行显示该客户所买商品名称,买的不足20种的则用"@"表示。得到的数据就是完全清理好的数据,保存到新的文件夹.

五、结果分析

1.数据描述性统计分析

(1)数据的基本信息

在65536条销售数据中，分类汇总产生结构化数据后共有31869条消费记录，其中购买一件商品的顾客购物篮有19778个,购买一件以上商品的购物篮有12091个,分别占总体销售数据的62.06%和37.94%,购买一件商品的比例稍高;在包含一件以上商品的12091个购物篮中,顾客大多购买2-4件商品,占总体的88%左右。

通过分析销量最多的10种商品发现，该超市销售38类商品中,销售量最多的是饮料(19.1%),其次是熟食速食(14.0%),第三是休闲食品(11.5%)。销量最多的10种商品的销量占总销量的81.0%。销量最少的10种商品的销售比例只占0.6%,其中最少的三种商品是服装服饰、鞋帽类、土产干货,其销量的比例都不到0.02%。

通过分析销售金额最多的10种商品，该超市销售38类商品中,销售金额最多的是烟草(19.8%),其次是饮料(13.9%),第三是蛋奶类(9.4%)。销售额最多的10种商品的占总销量销售额比例为82.4%。

2.关联规则的发现

(1)总分类关联规则

①由图1可知,在对饮料进行关联规则挖掘时发现,顾客在购买了饮料时,有可能同时购买蔬果(1.2%),熟食速食(1.3%),营养保健品(7.9%),塑料制品(21.6%)与休闲食品(34.3%)。顾客在购买饮料时一般有34.3%的可能性会同时购买休闲食品,这里的塑料制品被认为是塑料袋,这是一般规则。同时,顾客还有可能购买营养保健品,由此推断,顾客买饮料可能是看望长辈或家庭宴会,所以有7.9%的可能性购买营养保健品。

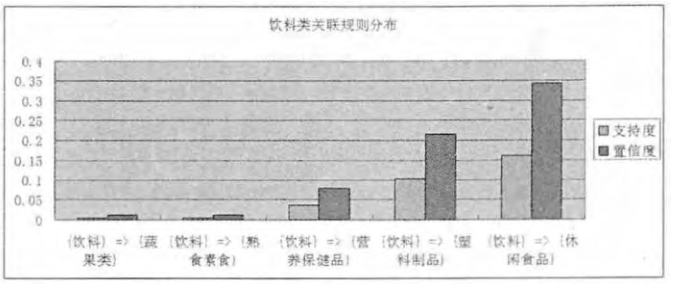

图1

②由图2可知,顾客在购买休闲食品时,有1.3%的可能会同时会购买饼干糕点或糖果类商品，这是我们生活经验的一般规则。同时发现,顾客在购买休闲食品时也极有1.33%的可能购买蔬果,3.67%的可能性购买酱菜。据推测,这可能是主妇在为孩子购买零食时,会购买生活必需品。

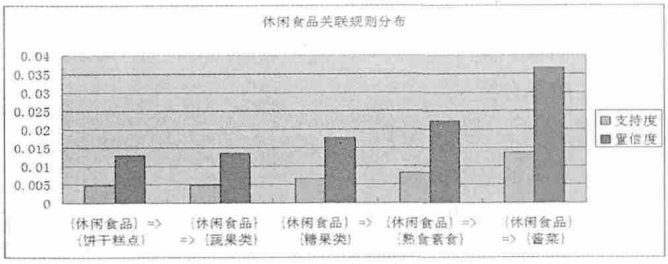

图 2

③根据图 3，顾客在购买调味制品时，会对粮食和酱菜有需要，购买可能性分别为 1.5%和 4.6%，这是主妇在购物时的一般规则；同时发现，顾客在购买调味制品时，也可能购买家用清洁（2.3%）和个人洁护（3.5%）等日用品，这可能是主妇在为家庭内添置一些食品，日用品等生活必需品。

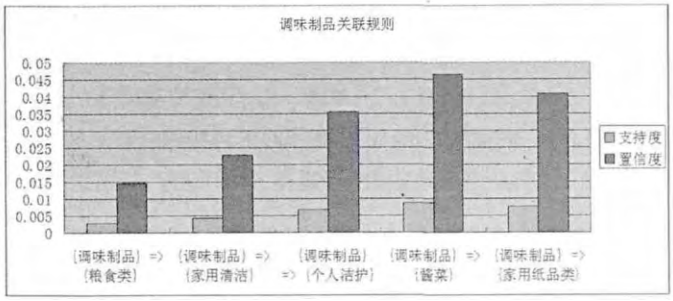

图 3

⑵ 明细分类关联规则

明细分类商 3 中销量前 5 的商品分别为水，香烟，肠，茶和购物袋。下面对香烟，香肠和茶作明细商品关联规则分析。

①对香烟的关联规则

由图 4 可知，香烟与打火机这两类商品的置信度较高，为 3.2%，说明此次关联规则挖掘贴近顾客的日常需求，这是对一般关联规则的有效验证。同时还发现，顾客在购买香烟的同时，有 1.3%的可能性会购买香皂，有 1.5%的可能性会购买鞋刷。据推测，这可能是由于顾客在购买香烟时会帮妻子购买一些日用品。除此之外，发现顾客在购买香烟时，对粥和可乐的购买分别为 2.9%和 3.5%，有较高的关联度。

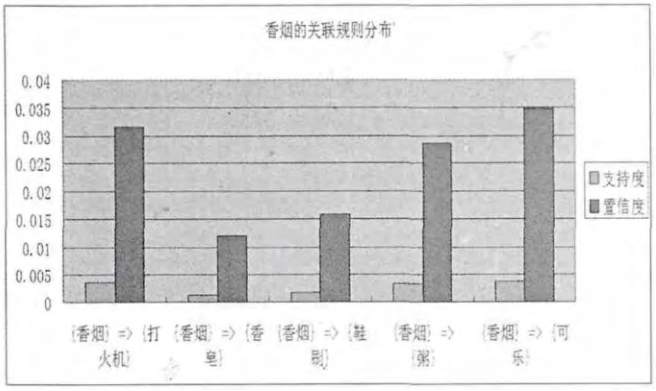

图 4

②对香肠的规则

在对香肠进行关联规则挖掘时，由图 5 可知，顾客也会同时购买其他零食，其中对丸子和凤爪的购买可能性分别为 1.9%和 1.1%。同时发现，顾客在购买香肠的同时有 1.03%的可能性会购买杯子。据推测，顾客可能是由于要宴请客人所以会同时购买香肠和水杯等餐桌必需品。

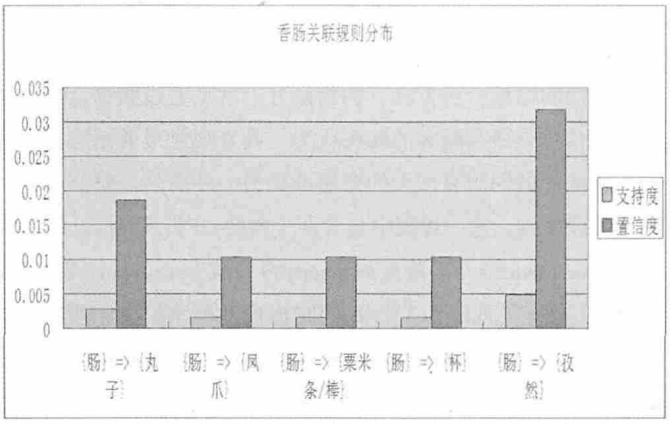

图 5

③对茶的关联规则

对茶进行关联规则挖掘，由图 6 可知，顾客在购买茶时，购买咖啡，可乐等替代饮品的可能性分别为 5.0%和 3.6%，拥有较强的可信度，说明对超市商品的分类摆放其实是有助于商品销售的。同时发现，顾客在买茶的同时，有 1.3%的可能性会购买鞋刷，据推测这可能是由于妻子在为丈夫购茶饮时，会同时买家用清洁用品。

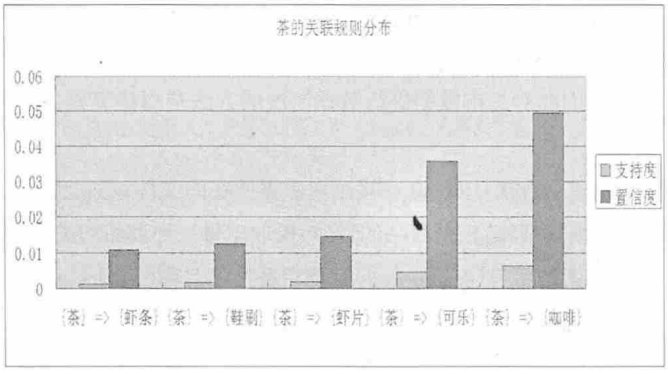

图 6

3.时间序列挖掘

根据折线图 7，我们推断出以下结论：图中七天每天营业时间 7：00am~21:00pm 中，商品销量和销售额均先走势平缓，之后达到峰点，随后下降，即在 18：00pm~20:00pm 达到峰点，说明此超市在此时间段人流量最大，推测可能是 18：00pm 以后，人们下班回家，会顺便带生活必需品或休闲食品等需要的商品回家。

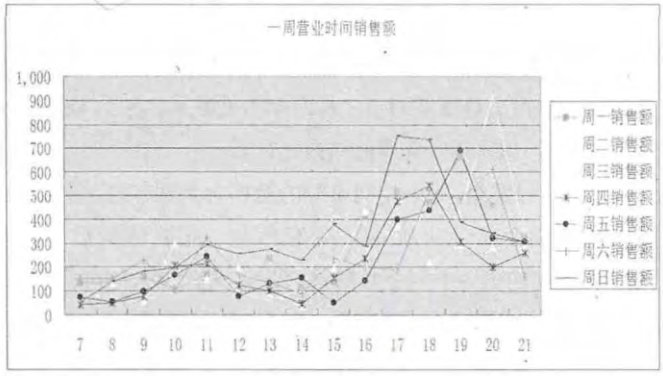

图 7

由图 8 发现，发现周销量和销售额的变化趋势相同，均是由平缓到峰值再下降，且峰值出现点均在 18 ∶00pm～20 ∶00pm 间，销售额 = 销量 × 单价则说明在高峰期单价销售差别不大，即此超市在 18 ∶00pm～20 ∶00pm 每日销售商品类似。

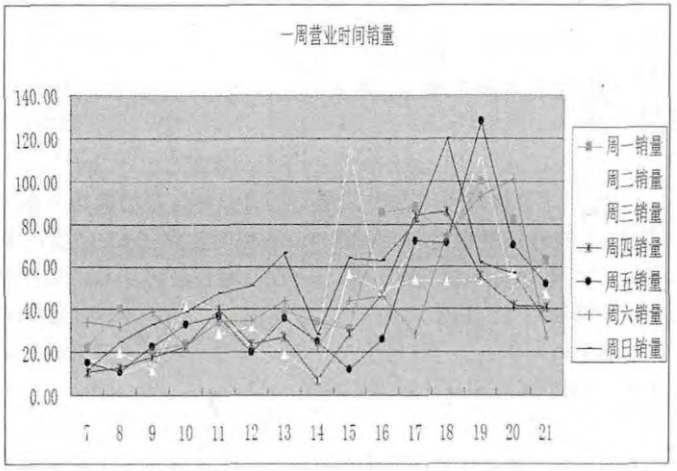

图 8

六、结论与建议

通过对六万条超市数据进行挖掘，获取描述统计信息，为管理者进货安排给出合理化建议。关联规则挖掘，在本次研究中发现，买饮料的人极有可能会买营养保健品 购买休闲食品的时候会买酱菜；买香烟的人极有可能会买鞋刷；买香肠的人会买水杯。根据得出的这些隐含的规则，可帮助管理者进行更好地货架拜访，从而提高超市销售量，同时增强顾客的购物体验。这一挖掘内容延于沃尔玛的啤酒与尿布的故事，也是本次研究的重点。其中挖掘到的新的关联规则是本次研究的创新点。时间序列分析可得到超市的购买高峰期，从而帮助管理者合理的进行人员的安排。

研究后发现，我们可进行如下改进，在数据的获取方面，在条件允许的情况下获取更加丰富的数据资源，此作为深入研究的必备条件。再者，进一步的学习关联规则算法，为研究提供理论支持。最后，多方面的进场分析，从客户以及决策中两方面的角度考虑分析，是研究结果更加全面。

参考文献：

[1]郑继刚.数据挖掘研究的现状与发展趋势[J].红河学院学报,2010.

[2]袁剑秋.基于关联规则算法在数据挖掘中的研究与应用[D].成都理工大学硕士论文,2009.

[3]T.Mitchell.MachineLearning.McGraw-Hill,Boston,MA,1997.

[4]林凡.数据挖掘在零售业交叉销售的作用[D].黑龙江硕士学位论文,2009.

[5]R.C.Holte.Very Simple Classification Rules Perform Well on Most Commonly Used Data sets.Machine Learning,11:63-91,1993 .

山东省居民消费与经济增长关系的实证分析

■崔铃敏　国　凡　青岛大学经济学院

摘　要：本文采用 1984 年 -2013 年时间序列数据，运用 VAR 模型对山东省城乡居民消费对经济增长的影响进行了实证分析。研究结果显示，城乡居民消费与经济增长存在格兰杰因果关系。山东省城乡居民消费均会拉动经济增长，其中城镇居民在拉动经济增长方面具有更大的作用。

关键词：城乡居民消费；经济增长；VAR 模型

DOI:10.14013/j.cnki.scxdh.2016.07.008

一、引言

从 90 年代中期以来，中国经济开始发生改变。消费、投资和出口拉动经济增长的“三驾马车”出现了不同的发展趋势。我国的投资率从 1978 年的 38.2%上升至 2013 年的 47.8%，我国货物和服务净出口占 GDP 比重也从 1978 年的 -0.3%上升至 2013 年的 2.4%，而消费率则是呈现下降趋势，由 1978 年的 62.1%下降 2013 年的 49.8%，通过这些数据可以看出消费增长不足，限制了我国经济持续的增长。因此，如何形成主要依靠消费需求拉动经济增长的格局成为社会关注的焦点之一。山东省 2013 年人均 GDP 为 56184 元，但近几年的居民消费对经济增长的贡献率却在 40%以下，因此山东省促进居民消费，以此实现实现消费与生产的相互促进租用，对于促进经济增长有重要作用。

二、基于 VAR 模型的实证分析

1.变量的选取

本文采用山东省的地区生产总值指数来衡量经济增长，城乡居民的消费分别用城镇居民和农村居民的人均生活消费衡量。以上变量通过对各年《山东省统计年鉴》整理所得，样本区间为 1984-2013。为剔除了价格因素采用价格指数进行平减。用 GDPZS 表示经济增长，以 CZ 和 NC 来表示城乡居民的消费。

2.变量平稳性检验

为了检验数据的平稳性，本文采用 ADF 方法检验 GDPZS、NC、CZ 是否存在单位根。根据 ADF 检验结果显示 NC、CZ、GDPZS 的一阶差分序列 ADF 检验值都小于临界值 (5%) 的值，因此这些变量都是一阶单整的。检验结果如下表。

变量	ADF 检验值	临界值（5%）	结论
GDPZS	0.529608	-2.971853	不平稳
DGDPZS	-3.259079	-2.971853	平稳
CZ	0.49077	-2.967767	不平稳
DCZ	-5.213855	-2.976263	平稳
NC	2.564339	-2.976263	不平稳
DNC	-3.492715	-2.976263	平稳

3.滞后阶数确定和模型稳定性检验

本文通过统计软件利用 LR(似然比)检验以及 AIC 信息准则