基于 Apriori 算法的关联规则超市购物推荐算法研究

董阳光
（渭南瑞泉中学，陕西 渭南 714000）

摘要：数据挖掘是指从大量的、有噪声的、不完全的、模糊的应用数据中，提取出隐含在其中的、前所未有的、有潜在应用价值的信息和知识的过程。它是一门融合了许多技术的一个多学科交叉领域，已成功地应用到了许多现实行业之中。本文通过利用数据挖掘中的关联规则算法 Apriori，从超市的购物记录中发现顾客放入他们购物篮中商品之间的关联关系，分析顾客的购物习惯，帮助零售商了解哪些商品频繁地被顾客同时购买，从而帮助他们制定更好的营销策略。利用这种关联规则，我们可以更具有针对性地指导超市工作人员进行商品的摆放，在方便顾客进行商品挑选的同时，增加营业收入，实现消费者与供应商的互利共赢。

关键词：数据挖掘；人工智能；关联规则；购物篮分析
DOI:10.19474/j.cnki.10-1156/f.002867

一、前言

随着互联网技术的普及，人类进入了信息化时代。计算机技术的应用大大提高了人们的生产效率及生活水平，各行各业逐步实现了信息化的发展道路。新技术的产生为人们的工作、学习带来了翻天覆地的变化，越来越成为了人们生活中不可分割的一部分。与此同时，医学，金融，教育，航天等各行各业的信息化系统积累了大量的数据信息。面对信息社会中数据和数据库的爆炸式增长，人们分析数据和从中提取有用信息的能力，远远不能满足实际需要。如何对这些海量的数据进行统计、分析、利用、决策已经成为当前迫切需要解决的问题。在此背景之下，数据挖掘（data mining）应运而生[1]。

数据挖掘是指从存放在数据库中的大量的、有噪声的、不完全的、模糊的应用数据中，利用科学方法提取出隐含在其中的、前所未有的、有潜在应用价值的信息和知识的过程。数据挖掘是信息论、人工智能、数据库、计算机信息管理系统、计算及决策支持系统、运筹学及统计学等诸多学科结合的产物。基于其在处理大数据方面的显著优势，数据挖掘技术已经在许多领域取得了令人瞩目的成就，成为新世纪发展最为迅猛的学科方向之一[2]。例如，利用数据挖掘中的分类算法，通过分析历史客户数据，可以帮助银行信息系统准确地识别出潜在可能的流失客户，从而促使相关部门尽早进行预警操作；将数据挖掘技术应用于医学领域，通过深入分析临床疾病表征与人体生理机制之间的内在联系，可以帮助医生进行早期诊断治疗，有利于为病人建立个性化的治疗方案，促进人类健康事业发展[3]；将数据挖掘技术应用于客户关系管理系统，对客户信息数据进行分析[4]，可以有效识别在市场竞争中最佳客户群、确定目标市场，有利于针对不同客户，实施不同策略[5]。个性化推荐是目前商业应用中发展比较快的一个领域，大型门户网站（例如天猫，京东，亚马逊，聚美等）都拥有较为专业的商品推荐功能，根据用户的历史购物行为或者网页浏览记录，后台进行复杂的算法计算，并最终在网页中针对具体用户推荐它们最有可能购买的商品，从而刺激消费者进行消费行为。此外，其他类型的应用软件例如网易云音乐，以其在个性化音乐推荐上的特色优势，在同类型软件中得到了广泛青睐，并迅速成长为最受欢迎的音乐应用。

利用数据挖掘技术进行商业数据分析，是一项极具现实意义的尝试，它能够加速理论知识到实际应用的转化，彰显科学技术在生产力发展中的重要作用。

二、研究内容

超市购物信息系统在大中小型超市中的收款、仓储、进货等方面的应用已经非常广泛，与此同时也产生了大量的商品消费数据，这些数据详细记录了与该超市消费群体密切相关的消费行为。如何更好地利用这些数据来为企业的经营活动提供有用的信息决策及指导成为了一个急需解决的问题，这也就是购物篮分析[6]。

零售业市场决策层越来越重视从已有的销售数据中寻找产品与产品之间存在的潜在的某种关系，这就需要从海量数据中快速准确地挖掘出有价值、能够描述数据项之间相互联系的有关知识，这是数据挖掘研究的一个重要研究领域。基于此理念得到的信息，可以指导销售人员在超市的商品摆放过程中，把可能被顾客同时购买的商品尽可能放在临近的位置，这样，即可以方便顾客，同时又能够增加营业额，是商业追求的互利共赢模式。关联规则挖掘技术在超市购物篮分析中的应用是当前研究的热点问题之一，在相关领域已经有了一定的应用和发展。本论文的目的是通过对某超市一段时间的购物记录数据进行分析，探索该超市服务地区顾客购物行为习惯的一般规律，并以此规律为一般结论帮助零售商做选择性销售和安排货架空间，进行超市商品摆放指导。

三、研究方法

本实验的主要目的是利用数据挖掘方法对数据进行处理，从超市购物篮的购物行为中发现顾客放入他们购物篮中商品之间的关联关系，分析顾客的购物习惯，帮助零售商了解哪些商品频繁地被顾客同时购买，从而辅助他们制定更好的营销策略。作为数据挖掘领域的一个重要的应用方面，购物篮分析的主要目标是在顾客的购买交易中分析出能够同时购买一类产品或一组产品的可能性，这是数据挖掘中关联规则所关注的问题，因此可以通过当前发展较为成熟的关联规则挖掘算法进行问题的分析研究。

目前，最常用的关联规则挖掘算法是 Apriori[7]，该方法于 1994 年提出，为布尔关联规则挖掘频繁项集的经典算法并得到了广泛应用。Apriori 算法是用一种称为逐层搜索的迭代方法。在进行算法实现之前，需要事先定义最小支持度与最小置信度。最小支持度是人为设定的表示频繁项集需要满足的最小商品出现概率；而最小置信度用于衡量所挖掘出的规则至少满足多少才是可信的：

支持度 support（A->B）=

置信度 confidence（A->B）=

具体的实施步骤如下：

（1）扫描数据库，也就是所有的购物清单记录，计算每

个商品（项）被购买的次数，产生候选频繁 1 项集。然后根据设定的最小支持度大小，保留满足最小支持度的项，产生频繁 1 项集 L1。

（2）在产生的频繁 1 项集的基础之上，找出所有购买的商品中任何两个商品一起购买的组合，并计数其被同时购买的次数，产生候选的频繁 2 项集。根据最小支持度大小，去掉频次小于最小支持度的项集，产生频繁 2 项集 L2。

（3）重复以上步骤（2），直到不能找到频繁项集。

在以上频繁项集的产生过程中，为了提高频繁项集的产生效率，Apriori 算法充分利用了两条基本原理：

（a）频繁项集的子集必为频繁项集。例如频繁项集 I1 包含商品 [a1,a2,a3]，说明 I1 的支持度大于最小支持度，那么所有 I 的非空子集 [a1,a2],[a1,a3], [a2,a3],[a1,a2,a3] 也都满足最小支持度，因此它们也都是频繁项集；

（b）非频繁项集的超集一定不是频繁项集。例如项集 I2 包含商品 [a1,a2]，他不满　足最小支持度，那么任何包含 I2 的项集 [a1,a2,ai] 都不是频繁项集。

因此当产生了频繁项集 Lk 的时候，我们可以仅仅在 Lk 的基础上寻找 Lk+1 频繁项集，从而大大地减少了计算复杂度，这就是 Apriori 算法的优势所在。

四、实验分析

我们的实验数据来自某超市近几个月的消费记录（图 1），共包含 46200 条消费记录，其中每一条消费记录由 20 种不同的商品构成。数字 0 表示该条消费记录没有购买该商品，数字 1 表示购买了该商品。在这里我们设定最小支持度为 5%，最小置信度为 50%。

```
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0
```

图 1：实验数据，每一行代表一次消费行为，每一列代表是否购买该商品

在实验的具体实施中，根据 Apriori 算法，我们分别计算频繁 1 项集，并按照最小支持度准则去掉不满足条件的结果，然后在频繁 1 项集的基础上产生频繁 2 项集，重复该过程，知道没有更高的频繁项集产生。由于该实验涉及的数据量巨大（46200 条消费记录，20 种不同的消费商品），因此需要借助专业的数据挖掘分析软件 Clementine[8] 来进行数据的分析。Clementine 最早由听过 ISL 公司开发，适用于大型工业和商业特实践，具有可视化的操作界面，是数据分析领域最常使用的工具之一。

Sort by: **Support**

Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2
9468	20.500	51.500	1.117	milk	biscuits	
7045	15.200	52.200	1.131	milk	yoghurt	
5363	11.600	53.200	1.520	pasta	tomato souce	
5363	11.600	51.300	1.112	milk	tomato souce	
4417	9.600	56.000	1.214	milk	water	pasta
3818	8.300	53.300	1.155	milk	juices	
3590	7.800	57.900	1.254	milk	biscuits	pasta
3007	6.500	51.800	1.479	pasta	rice	
2855	6.200	55.100	1.195	milk	tomato souce	pasta
2772	6.000	56.900	1.234	milk	coffee	pasta
2734	5.900	60.200	1.306	milk	biscuits	water
2751	5.900	57.400	1.243	milk	brioches	pasta
2750	5.900	57.200	1.634	pasta	tomato souce	milk
2441	5.300	58.900	1.277	milk	yoghurt	pasta

图 2 按照支持度（support）大小顺序显示出的关联规则

图 2 和图 3 显示了 Apriori 算法对数据进行关联规则挖掘出的结果，为了直观显示，我们分别按照支持度，置信度由高到低的顺序分别进行了显示。由图 2 可知共有 14 条关联规则满足最小支持度和最小置信度。从所有的关联规则里可以看到，和 milk 一起购买的商品较多，它和 biscuits、yoghurt、tomato source 等在内的许多商品在购物篮中都存在关联关系。其中具有最大支持度的前五条规则是：

{ biscuits} => milk　　{ yoghurt} => milk　　{tomato source} => pasta

{ tomato source } => milk　　{water, pasta} => milk

以上关联规则表示消费者购买了其中一个或两个的同时也购买了关联规则另一边的商品。当按照 confidence 的大小顺序显示的时候，前 5 条规则是：

{ biscuits, water } => milk　　{ yoghurt, pasta } => milk

{pasta，biscuits } => milk　　{pasta，brioches} => milk

{ tomato source，milk} => pasta

虽然关联规则本身没有变，但是像{biscuits, water} => milk，{yoghurt, pasta} => milk 这些规则虽然没有较大的support，但是置信度较高，说明该关联规则的确定性更高。

Sort by: Confidence

Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2
2734	5.900	60.200	1.306	milk	biscuits	water
2441	5.300	58.900	1.277	milk	yoghurt	pasta
3590	7.800	57.900	1.254	milk	biscuits	pasta
2751	5.900	57.400	1.243	milk	brioches	pasta
2750	5.900	57.200	1.634	pasta	tomato souce	milk
2772	6.000	56.900	1.234	milk	coffee	pasta
4417	9.600	56.000	1.214	milk	water	pasta
2855	6.200	55.100	1.195	milk	tomato souce	pasta
3818	8.300	53.300	1.155	milk	juices	
5363	11.600	53.200	1.520	pasta	tomato souce	
7045	15.200	52.200	1.131	milk	yoghurt	
3007	6.500	51.800	1.479	pasta	rice	
9468	20.500	51.500	1.117	milk	biscuits	
5363	11.600	51.300	1.112	milk	tomato souce	

图 3：按照置信度（confidence）大小顺序显示出的关联规则

五、总结

数据挖掘作为一种蓬勃发展的计算机技术，能够针对大数据进行处理分析，帮助我们发现隐藏在数据中的有价值信息[9]。本论文主要利用数据挖掘中关联规则挖掘算法对超市的购物消费记录进行关联规则挖掘，通过利用经典的 Apriori 算法，我们发现了消费者购物行为中一些不为平时所注意的关联关系。利用这种关联规则，可以更具有针对性地指导超市工作人员进行商品的摆放，在方便顾客进行商品挑选的同时，增加营业收入，实现消费者与供应商的双赢[10]。此外，我们发现，Apriori 算法每进行一次频繁项集的产生，都需要扫描一次数据库，因此会使计算效率变低。因此，在接下来的工作中，需要研究其他方法解决这个问题。

在现实生活中，我们可以充分利用数据挖掘技术进行各种各样的实际问题的分析，尤其是在商业中的应用，可以促进理论到实践的转化，是一项极具现实意义的探索[11]。

参考文献：

[1] 吕成哲，赵晓明，王起伟．浅谈数据挖掘理论．中国西部科技（学术），2007, 39-42

[2] 朱建平，张润楚．数据挖掘的发展及其特点．统计与决策，2002, 71-72

[3] 李明江，唐颖，周力军．数据挖掘技术及应用．中国新通信，2012, 66-67+74

[4] 罗鹏．数据挖掘在客户关系管理（crm）中的应用．昆明理工大学，2013

[5] 黄林军，张勇，郭冰榕．机器学习技术在数据挖掘中的商业应用．中山大学学报论丛，2005, 145-148

[6] 刘锡铃．关联规则挖掘算法及其在购物篮分析中的应用研究．苏州大学，2009

[7] 赵洪英，蔡乐才，李先杰．关联规则挖掘的 apriori 算法综述．四川理工学院学报（自然科学版），2011, 66-70

[8] 张帆．基于 clementine 的广告客户数据挖掘模型设计与实现．北京邮电大学，2010

[9] 赵倩倩，程国建，冀乾宇．大数据崛起与数据挖掘刍议．电脑知识与技术，2014, 7831-7833

[10]崔贯勋，李梁，王柯柯．关联规则挖掘中 apriori 算法的研究与改进．计算机应用，2010, 2952-2955

[11]张春华，王阳．数据挖掘技术、应用及发展趋势．现代情报，2003, 47-48+50

作者简介：

董阳光，男，陕西渭南人，渭南瑞泉中学。

上接第039页

争夺市场，获得发展主动权，经常会采用“免费骑行”等营销策略，这就从侧面反映了，共享单车在实际的应用发展中，企业早期的盈利目标是难以达成的。企业为了获得更加具有实效性的发展，一般会选择向社会融资。从城市共享单车的发展趋势来看，我国未来的共享经济要获得更长远的发展，就需要结合实际的市场情况，充分应用网络平台，实现多元化的融资模式，积极拓展融资渠道，为共享经济的发展注入足够的资本。由当前的这种共享经济模式，我们还可以联想到共享房产，这是一种在线的短租业务，共享型的房屋在不同的网络平台上线后，能够为短租客户，包括游客、白领等提供住房便利，一定程度上缓解了人们的购房压力，解决人们的住房问题，同时降低了空房率，提供房屋资源的利用率。

（四）增强服务性，实现长效性的经济发展

共享单车之所以能够发展起来，主要还是市场需求，人们自己购买单车，需要一定费用，且不能随时随地骑行或停放，还要担心失窃问题，共享经济下的单车共用直接就解决了这些问题，单车使用权在时段性的转移中，人人都可以使用，且应用成本低，不需要考虑失窃问题。随着我国城市化建设的不断发展，城市人口增加，人们对于这种时段性的租赁服务经济需求将进一步增加，共享内容需要更加丰富化，共享服务业需要更加人性化。仍旧是以共享单车为例，企业要获得更大的市场竞争力，在产品内容和性能、质量以及服务上要不断提升，如智能车锁、GPS 定位系统、车筐、升降座椅等均能够为用户提供更好的体验。共享经济的发展进一步形成了一种以租代买的商业业务模式，例如大学生的空调、桌椅、板凳等用品，学生可以以租赁的形式获得使用权，在大学毕业后，学生可以再以转租的形式脱手，实现了资源的循环应用。

三、结语

共享单车本质上就是具有公共属性的城市共用类型的单车，共享单车的使用与发展较快，从中我们也可以发现一些经济发展中存在的问题。例如现代化城市经济的迅速发展，也促使城市交通的发展，小汽车、公共汽车数量大大增加，城市地面交通的荷载压力也加大，交通拥堵问题是城市居民出行的常遇问题，而共享单车的出现，却为人们的出行带来较大的便利，既有利于人们锻炼身体，又能够节能减配，起到绿色出行的作用。由共享单车的使用发展，可以分析出未来中国在经济大建设与发展中，共享经济作为一种租赁服务经济将进一步发展，且随着城市人口的不断增加和互联网信息技术的不断应用，共享资源市场将不断扩大，但是共享经济在商业模式上和竞争管理体系上仍旧需要在实践中不断完善。

参考文献：

[1]. 共享单车与城市可持续发展——中国城市交通发展论坛 2017 年第一次专题研讨会 [J]. 城市交通 ,2017,03：1-8.

[2] 廉伟 . 城市共享资源的发展之路在何方———以共享单车为例 [J]. 现代交际 .

[3] 王楠 . 从共享单车上树现象谈如何有效管理推动共享经济健康发展 [J]. 中国战略新兴产业 ,2017,12：23-24.

[4] 李琨浩 . 基于共享经济视角下城市共享单车发展对策研究 [J]. 城市 ,2017,03：66-69.

作者简介：

赵汶轩，洛阳市第一高级中学。