

基于GRI模型和C5.0规则集的图书推荐模式研究*

□ 苏静 赵捷 / 中国科学技术信息研究所 北京 100038

摘要: 通过应用Clementine 12.0中关联规则的GRI模型和C5.0推理规则集,对问卷涉及的205位青少年用户基本信息和图书购买信息进行数据建模和分析。研究结果表明,单个用户购买的多类图书之间,具有相同购买特征(如性别、年龄、教育背景等)的用户群和某类图书之间均存在一定的关联规律。在此基础上,提出了基于GRI模型和C5.0规则集的图书推荐模式,即跨类型和相同用户兴趣的图书推荐,试图为图书馆和图书零售市场开展图书推荐服务提供参考。

关键词: 图书推荐, 关联规则, GRI模型, C5.0规则集

DOI: 10.3772/j.issn.1673—2286.2012.10.006

1 引言

文化产业的繁荣推动了图书出版事业和电子图书市场的蓬勃发展。2011年上半年,中国图书零售市场动销品种数同比增长率为5.74%,达到了92.6万种^[1]。面对日益增长的图书种类和数量,如何根据用户的个性化行为习惯,主动推荐信息资源,以满足用户需求和提升信息服务水平,是图书馆和图书零售市场所要关注并亟待解决的问题。

近年来,数据挖掘作为一种从大量数据中提取隐含和潜在信息的新兴技术,备受各领域学者的关注,不少论文也就数据挖掘在图书推荐方面的应用进行了深入的探讨。按照与用户的个性化需求相结合的程度不同,图书推荐研究大致分为三大类:基于图书资源整合的图书推荐、基于用户定制的图书推荐和基于用户行为与人口特征的图书推荐。其中,第一类是通过RSS^[2]、电子邮件^[3]等方式将新到图书和热点图书推荐给用户的服

务模式,推荐信息具有容量大、内容新等特点;第二类是按照用户的定制要求,将其预先选定的知识门类、学科专业、信息内容等方面的需求汇总分析后,有选择地定期推荐给特约用户的服务^[4];第三类则是结合用户的行为习惯和人口统计学特征,运用关联规则的Apriori算法及其改进算法^[5-7]、ID3决策树算法^[8]、本体论^[9]和语义网络^[10]等,抽取用户的相似兴趣点,从而分析出可以推荐的内容。相比而言,第三类与用户潜在需求的契合度更高、针对性更强。

Clementine数据挖掘软件(PASW Modeler)^[11]将数据挖掘的理论成果与实践研发相结合,拥有着强大的数据挖掘算法,提供了决策树、人工神经网络、回归分析、关联分析、聚类分析等丰富的数据挖掘模型^[12],支持与数据库之间的数据和模型交换,能够从海量数据中提取辅助管理和决策的有价值信息,具有操作简明、界面可视化等特点,已经普遍应用于电子商务、社会经济、教育、金融等领域。

目前,鉴于涉及关联规则的GRI算法运用较少,Clementine在图书情报领域的实践分析不多、输出结果的解释较为简略、图书推荐的可视化分析手段单一等现状,本文拟采用数据挖掘中的关联规则和C5.0模型,运用Clementine 12.0对图书推荐模式策略进行深入的实证研究与翔实分析,以挖掘不同种类图书之间、图书与用户之间的潜在关联,进而分析获得适合不同类型用户的图书推荐规则,希望可以对各类网上书店和图书零售实体市场的资源推荐服务有一定的借鉴意义,尤其是为如何提升图书馆的用户个性化服务提供参考。

2 研究方法 with 数据来源

2.1 研究方法

关联规则(Association Rules)是由Agrawal等人于1993年提出的,用于挖掘商品销售中顾客行为的购买模式,常用于揭示大量

数据变量之间未知的依赖关系^[13]。GRI算法是关联规则的算法之一，侧重于关联规则的分析及应用，包括如何处理数值型变量、如何将单一概念层次的关联推广到多概念层次的关联等，进而描述事物的内在结构。它采用深度优先搜索策略实现算法，主要用于简单关联分析，一般表示形式是“ $X \Rightarrow Y$ (规则支持度S 规则置信度C)”，X称为规则的前项 (Antecedent)，Y称为规则的后项 (Consequent)^[14]。C5.0是决策树的经典算法之一，可以根据PRISM算法自动生成推理规则集，总是以期望类别的最大正确覆盖率为标准，用以实现数据集内在的规律探究和数据对象的分类与预测^[15]，一般表示形式为“如果<条件>则<结论>……”。

基于GRI模型和C5.0规则集，本研究运用Clementine 12.0软件的运算处理功能和可视化展现功能，对问卷信息进行数据建模和分析，挖掘用户购买的图书类型之间、具有相同购买特征 (如性别、年龄、教育背景等) 的用户群和某类图书之间的关联规律。研究路线如图1所示。

2.2 数据来源

本研究以网络问卷的形式对年龄段为13-23岁的青少年进行了问卷调查。调查问卷共有16个数据集 (见表1)，包括5个人口统计学变量和11个标识图书类型的字段。调查的主要目的是测量各变量之间的关联特征，被调查者同时购买两类或者多类图书视为一条关联规则，若只选择单一类型图书的问卷，视为无效问卷。经小样本的试调查，问卷的效度和信度符合研究要求。本次调查共发出问卷205份，回收问

卷196份，剔除无效问卷，最终得到有效问卷166份，回收率为80.1%。

3 新型图书推荐模式研究

3.1 应用GRI模型的图书关联分析

应用GRI模型，结合支持度 (Support)、置信度 (Confidence) 和作用度 (Lift) 等指标，分析用户购买的图书类型之间的关联强度并判断关联规则的有效性和实用性。其中，支持度度量规则的频率，即测度关联规则在数据集中的普遍性，表明出现前项X的概率；置信度

度量规则的强度，即衡量关联规则的准确度，表明出现前项X时也出现后项Y的概率；作用度反映出现前项X对出现后项Y的影响程度，大于1代表有促进作用^[16]。

按支持度从高到低排序，部分关联结果如图2所示。分析可得，财经类图书、管理类图书和成功类图书之间存在多种关联 ($\text{success} \Rightarrow \text{financial}$ S=38.79% C=79.69%; $\text{success} \Rightarrow \text{management}$ S=38.79% C=84.38%)，小说类图书和美容类图书之间出现双向关联规则 ($\text{cosmetic} \Rightarrow \text{novel}$ S=38.79% C=75%; $\text{novel} \Rightarrow \text{cosmetic}$ S=36.36% C=80%)。以第一条关联规则为

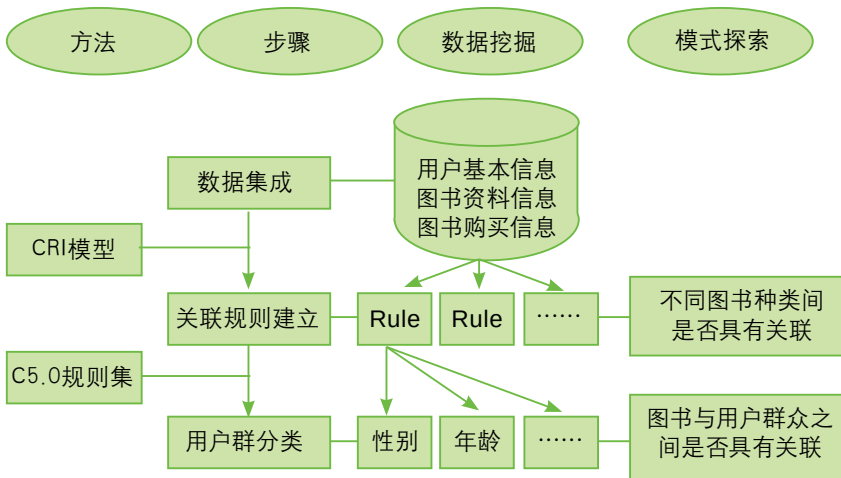


图1 新型图书推荐模式研究路线

表1 调查问卷数据集内容

序号	字段名称	字段意义	序号	字段名称	字段意义
1	id	用户唯一标识符	9	popular science	购买科普类图书
2	gender	性别	10	teaching resources	购买教辅类图书
3	age	年龄	11	travel	购买旅游类图书
4	education	教育背景	12	health	购买保健类图书
5	expenditure	月购书花销	13	management	购买管理类图书
6	novel	购买小说类图书	14	financial	购买财经类图书
7	art	购买艺术类图书	15	success	购买成功类图书
8	cartoon	购买卡通类图书	16	cosmetic	购买美容类图书

例,当用户购买了成功类的图书,系统可以将财经类和管理类图书推荐给用户。

值得注意的是,虽然旅游类图书和成功类图书之间具有关联关系,但置信度较低($C=59.38\%$)。这说明,规则的可信程度差,旅游类图书和成功类图书之间的关系是间接联系,由其他类图书(财经类或管理类)作为二者的纽带。因此,根据关联规则结果,当用户购买旅游类图书时,不能直接向其推荐成功类图书。

按置信度从高到低排序(见图3),设置最小置信度为90%,最小支持度为8%,以确保规则的可信性和普遍性。由此,得到三条规则:科普类图书和教辅类图书之间($\text{popular science} \Rightarrow \text{teaching resources}$ $S=9.09\%$ $C=100\%$),教辅类图书、旅游类图书和财经类图书之间($\text{teaching resources and travel} \Rightarrow \text{financial}$ $S=8.48\%$ $C=100\%$),教辅类图书、成功类图书和管理类图书之间($\text{teaching resources and success} \Rightarrow \text{management}$ $S=8.48\%$ $C=100\%$)均具有一定关联性。另一方面,这几条关联规则具有这样的特征:置信度高,支持度较低。这说明规则的可信度高,但普遍性差,也就是说购买a类图书的同时购买b类图书的概率极高,但是a类图书本身在所有购买活动中出现的概率较低,造成规则的适用范围有限、应用机会较少。因此,由于上述规则不具有一般性,仅作为推荐方案之一,而不是重点或是热门推荐活动,例如可以根据第1条规则将科普类图书和教辅类图书进行捆绑式销售。此外,以上关联规则的作用度均大于1,说明关联结果具有实用性和实际意义。

	Consequent	Antecedent	Support %	Confidence %
1	novel	cosmetic	38.79	75.0
2	travel	success	38.79	59.38
3	financial	success	38.79	79.69
4	management	success	38.79	84.38
5	cosmetic	novel	36.36	80.0
6	management	novel	36.36	70.0
7	travel	management and success	32.73	62.96
8	travel	financial and success	30.91	62.75
9	success	travel	27.27	84.44
10	financial	travel	27.27	82.22

图2 GRI节点执行后的关联规则(以支持度排序)

	Consequent	Antecedent	Support %	Confidence %
1	teaching resources	popular science	9.09	100.0
2	financial	teaching resources and travel	8.48	100.0
3	management	teaching resources and success	8.48	100.0
4	management	art and success	7.27	100.0
5	teaching resources	cartoon and popular science	1.82	100.0
6	novel	art and cosmetic	1.82	100.0
7	novel	health	1.21	100.0
8	novel	travel and cosmetic	1.21	100.0
9	management	novel and success	2.42	100.0
10	novel	art and health	0.61	100.0

图3 GRI节点执行后的关联规则(以置信度排序)

综合而言,由于关联规则本身并没有关于预测精度和误差的评价指标,因此通常不直接用于预测建模。但可以利用关联规则对样本隐含的规律进行归纳和总结,考察大部分用户的购买/借阅习惯和结构模式,从而开展相应的图书推荐服务。

3.2 应用C5.0规则集的图书关联分析

在上述关联规则的基础上,本研究使用C5.0规则集创建了双向关联的Web图(见图4),用于发现目标群体的统一特征。图中节点形状代表不同种类的图书,节点间的连线粗细代表二者的关系强弱。为了有效区分不同购买意向的用户群、避免强链接过多,图中的strong links(强链接)阈值设为45。

结果显示(见图4),购买财经类图书、管理类图书和成功类图书的用户具有类似的标识特征(R1);购买管理类图书和美容

类图书的用户具有类似的标识特征(R2);购买小说类图书和美容类图书的用户具有类似的标识特征(R3),这三条关联规则所对应的用户群就是相应种类图书的未来市场发展的目标群体。

基于已经标识的三个用户群(R1、R2、R3),建立C5.0模型来辨别每个用户群的人口统计学特征,将输出类型设置为规则集,结果为真(T)的规则集如表2所示。经分析可知,性别、年龄和教育背景是最重要的三个变量。当性别为男性,年龄大于16岁时,R1成立;当性别为女性,教育背景为大学时,R2成立;当性别为女性时,R3成立。根据数据挖掘结果,我们可以以用户的人口统计学特征为基点开展图书推荐服务,即用户的人口统计学特征满足一定条件时,可推测其有可能性同时购买某类图书,则将相应的图书种类的畅销图书信息发送给用户。以R1为例,若注册用户的性别为男性并且年龄大于16岁

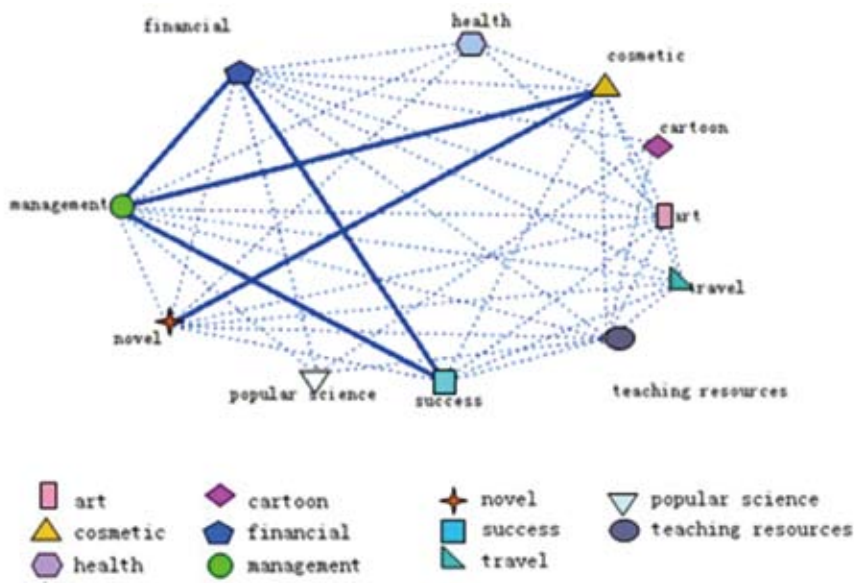


图4 图书种类关联Web图

时,则可以将财经类、管理类和成功类的热点图书信息推荐给用户。

3.3 基于GRI模型和C5.0规则集的图书推荐模式

上述结果表明,利用GRI模型

和C5.0规则集来分析用户数据,能够挖掘出不同图书种类之间、图书种类与用户群之间所隐含的关联性。由此,我们提出一种新的图书推荐模式:跨类型和相同用户兴趣的图书推荐。其中,跨类型是指根据不同图书种类之间的关联性推

荐:若用户购买/借阅A类图书,图书馆或图书零售市场可以将具有一定关联关系的B类图书推荐给用户;相同用户兴趣则是针对具有相同个性化特征的用户进行关联性推荐:若用户符合一定条件下的人口统计学特征(如年龄、性别、教育背景、收入水平、购书花销、职业等),则依照现有规则,将对应的图书种类的热点图书推荐给用户。

跨类型和相同用户兴趣的图书推荐模式,对于高校图书馆、专业图书馆以及各类书店都有一定的实用价值。以高校图书馆为例,要实施这种推荐模式,首先需要构建相应数据集,包括用户基本信息、图书资料信息、图书借阅信息(见图5)。其中,图书借阅信息是由借阅行为本身发生而产生的数据信息。然后,运用关联规则GRI模型和C5.0规则集寻找数据对象的关联特征和逻辑特征,如根据图书借阅信息挖掘用户专业和借阅图书分类号的关联性、用户年龄和借阅图书分类号的关联性、借阅图书之间的关联性等。需要说明的是,这种推荐模式并非一成不变,而是通过客户端与Web服务器的不断交互,实时更新数据库信息。同时,借助Clementine Server的模型库评价、分析和预测用户需求,以生成和完善关联/推理规则集,从而形成了一个循环往复的运行流程,具体流程如图6所示。

表2 人口统计学规则集

规则	R1	R2	R3
前提	IF gender=male and age>16	IF gender=female and education =undergraduate	IF gender=female
结果	THEN T	THEN T	THEN T

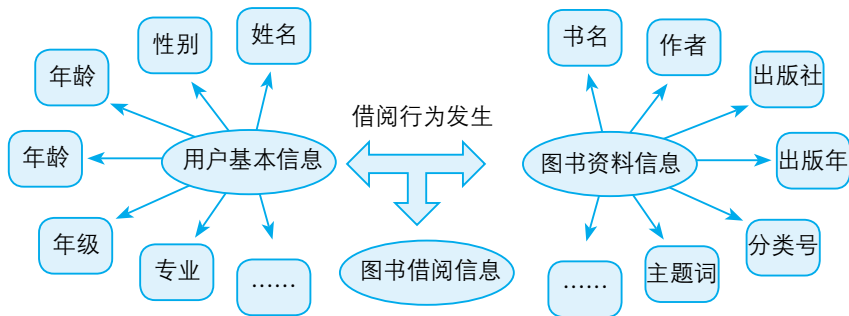


图5 数据集及其关联性

4 结语

图书馆馆藏的借阅流通数据中蕴含着大量用户的图书兴趣关联,同样地,各类书店的销售记录中也隐藏着丰富的图书关联。这些信息记录了用户的真实信息需求,体现

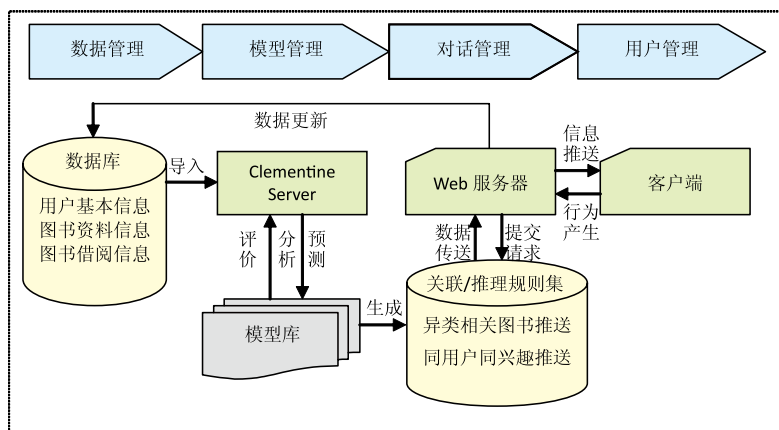


图6 图书推荐模式流程

着用户需求的内在联系和规律，分析和挖掘这些信息，对于提升个性化服务的质量能够起到不小的促进

作用。同时，在信息爆炸时代，数据挖掘技术是探索性分析和归纳的重要研究手段，以满足用户个性化信

息服务为出发点，以发现用户信息的需求模式为目标，本身具有较强的针对性和实用性。基于数据挖掘的图书推荐模式，适合快速定位用户的信息兴趣点，进而对实现个性化和主动式的图书推荐服务提供决策支持。虽然本研究采用的数据具有一定的时效性和局限性，也受到市场波动、环境变化、名人效应等诸多因素的影响，但所提出的新型图书推荐模式有利于高效地获取图书关联信息，更能够人性化地为推荐相关图书，在高校图书馆、专业图书馆和各类书店中都有一定的应用价值。

参考文献

- [1] 开卷发布“2011上半年中国图书零售市场报告”[EB/OL]. [2011-11-12]. http://news.ifeng.com/gundong/detail_2011_09/05/8939421_0.shtml.
- [2] 胡新平,张天俊,董建成.基于数字图书馆的RSS推送服务研究[J].情报杂志,2008,27(1):55-56.
- [3] 袁利民,周东升.新书书目邮件推送服务的实现[J].现代情报,2004,24(6):157-158,160.
- [4] 杨九龙.基于The Shifted Users的图书馆个性化信息推送服务模式研究[J].情报杂志,2008,27(12):139-141.
- [5] 王燕,温有奎.基于关联规则的推荐系统在数字图书馆中的应用[J].情报科学,2007,25(6):877-880.
- [6] 丁雪.基于数据挖掘的图书智能推荐系统研究[J].情报理论与实践,2010,33(5):107-110.
- [7] 熊拥军,陈春颖.基于关联挖掘技术的数字图书馆个性化推送服务[J].图书情报工作,2010,4(1):125-129.
- [8] 王殿佑.一种基于数据挖掘技术的馆藏资源个性化推荐服务[J].情报学报,2008,27(6):813-818.
- [9] 郑美玉.基于本体论的高校图书馆个性化书目推送系统[J].图书情报工作,2010,54(11):108-111,135.
- [10] 赵杨.基于语义网络的数字图书馆知识推送服务系统研究[J].情报科学,2007,25(12):1869-1873,1882.
- [11] HENSCHEN D. IBM Software Makes Sense Of Smileys [J]. Information week, 2010(17):18.
- [12] STALLMAN A T, BRADFORD E. Predict the Future with SPSS Clementine Data Mining Software [J]. iSeries news, 2009(6):35-36.
- [13] 张炜.读者借阅行为的关联知识发现实证分析[J].图书馆工作与研究,2010(12):38-41.
- [14] 薛薇,陈欢歌.Clementine数据挖掘方法及应用[M].北京:电子工业出版社,2010:245-249,256-260.
- [15] ROIGER R J, GEATZ M W.数据挖掘基础教程[M].北京:清华大学出版社,2003:78-83.
- [16] DUNHAM M H.数据挖掘教程[M].北京:清华大学出版社,2005:141-144.

作者简介

苏静(1988-),女,中国科学技术信息研究所硕士研究生,研究方向:引文分析、知识链接。E-mail: owensujing@163.com
赵捷(1959-),男,高级工程师,中国科学技术信息研究所信息资源中心副主任,研究方向:信息构建、数据库建设。

A Study of Recommend Modes in Books Service Based on the GRI Algorithm and C5.0-Based Rule Set

Su Jing, Zhao Jie / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: This paper obtains the data of the personal information and books purchased information through the questionnaire survey (n=205). Moreover, based on the GRI algorithm and C5.0-based rule set in Clementine 12.0, it discovers a thorough understanding about some significant links among all categories of books purchased, but also between one category of books and simultaneous user group flags, such as gender, age, education background and so on. Finally, it intends to put forward a new recommendation model and its application in library services.

Keywords: Recommendation models in books service, Association rules, GRI model, C5.0-based rule set

(收稿日期: 2012-06-25)