基于 Android 平台的图书阅读推荐系统

丁 勇 朱长水
(南京理工大学泰州科技学院 泰州 225300)

摘 要 随着移动互联网技术的发展,通过手机进行阅读已经成为人们的一种生活习惯。为了帮助读者在成千上万的"书海"中找到自己喜欢的图书,提出将经典的频繁项集挖掘算法 FP-Growth 应用到图书推荐系统中。算法根据读者的历史阅读记录,挖掘频繁出现的图书阅读组合,提取满足最小支持度和最小置信度阈值的强关联性规则,并根据关联规则进行图书智能推荐。实例证明该系统能够为读者提供快速、准确的智能推荐服务。

关键词 图书推荐,Android,关联规则,FP-Growth
中图法分类号 TP311 **文献标识码** A

Books Recommended Reading System Based on Android

DING Yong ZHU Chang-shui
(Taizhou Institute of Science and Technology, Nanjing University of Science and Technology, Taizhou 225300, China)

Abstract With the development of mobile internet technology, the mobile reading has become a kind of life habits. In order to help the reader find the favorite books in tens of thousands of "Sea of books", this paper applied the classic FP-Growth frequent itemsets mining algorithm into book recommendation system. Based on the reader's reading history records, the algorithm mines frequent book combination, extracts association rules satisfying minimum support and minimum confidence threshold. According to the rules, intelligent books can be recommended. Experiments show that, this system can offer rapid and accurate books recommended service for the readers.

Keywords Books recommendation, Android, Association rules, FP-Growth

1 引言

个性化推荐[1]是指将用户的历史行为数据通过数学建模,来挖掘用户的兴趣和偏好,以此将用户可能感兴趣的信息筛选出来。近年来,随着移动互联网技术的发展,通过手机进行阅读已经成为人们的一种生活习惯,手机阅读也逐渐成为移动运营商的主流增值业务。如此巨大的电子图书信息已经远远超过了人们的鉴别能力,使用传统的搜索引擎已经无法快速获取所需的信息,因此实现一个基于移动平台的图书推荐系统,为读者提供快速、准确的图书推荐服务,帮助读者在成千上万的"书海"中找到自己喜欢的图书,是十分必要的。

随着数据挖掘技术的广泛应用,当前有不少学者提出将数据挖掘相关技术应用到"个性化图书推荐"中。陈亮等[2]提出构建基于数据挖掘技术的图书智能推荐系统,并将经典的 Apriori 算法应用到图书推荐中,通过对读者的借阅记录进行深层次挖掘,得到置信度高的关联规则,并以此作为图书智能推荐的依据。黄洋等[3]结合分类、聚类和协同过滤等技术,设计实现了一种图书个性化推荐算法,算法对数据库中保留的大量用户的图书借阅记录进行挖掘,发现其中的规律,并向读者提供个性化图书推荐服务。高成等[4]提出一种基于 LDA 模型的主题句抽取方法,并将其应用于图书个性化推荐中,通过追踪用户阅读兴趣的变化,对用户的历史阅读行为日志进行分析,从而建立具有代表性的用户-主题模型,最后基于该模型实现图书智能推荐服务。荆月敏等[5]提出应用关联规则、决策树等数据挖掘方法对读者的借阅记录进行挖掘,分别从读者和图书分类角度对数据进行处理,得到读者群感兴趣的图书,最后通过分析读者是否符合该读者群的特征,来决策是否向该读者推荐这类图书。

2 基于 FP-Growth 的图书推荐算法

由于 Apriori 算法在挖掘频繁模式时会产生大量的候选项,使得算法时间和空间复杂度较高,因此本系统采用 FP-Growth 算法作为图书推荐算法。FP-growth 算法也是一种频繁模式挖掘算法,该算法最大的优势在于挖掘频繁项集时不产生大量的候选集,提高了算法执行的效率。FP-Growth 算法的思想是尽可能减少扫描事务数据库的次数,从而提高挖掘的效率。算法首先定义一种称为频繁模式树(Frequent Pattern Tree)的数据结构,FP-tree 是一种特殊的前缀树,包括项头表和项前缀树,前缀树存储候选项集,树的节点用来存储后缀项,分支用来标识项名,路径用来表示项集。然后,对频繁项进行排序,将支持度高的项排在前面,这使得 FP-tree 中出现频繁的项被共享的可能性增大,从而有效地节省存储

本文受 2015 江苏省高校自然科学研究面上项目(15KJB520016)资助。
丁 勇(1980—),男,硕士,副教授,主要研究方向为数据库理论与数据挖掘,E-mail:4383526@qq.com;朱长水(1981—),男,硕士,讲师,主要研究方向为虚拟现实、图像处理,E-mail:shui_zc@163.com。

空间。最后，不断迭代 FP-tree 的构造和投影过程，对每个频繁项，构造它的投影 FP-tree 和条件投影数据库，对每个新构建的 FP-tree 重复这个过程，直到构造的新 FP-tree 为空或仅包含一条路径。当构造的 FP-tree 为空时，其前缀即为频繁模式；当仅包含一条路径时，枚举所有可能组合，并与树的前缀进行连接操作，从而得到频繁模式。算法描述如下。

算法 1 procedure Insert_tree(P,N)

输入：阅读数据库 D；最小支持度阈值 min_sup

输出：频繁阅读模式集合

1. 扫描阅读数据库 D
2. 收集频繁项集 F，并计算它们的支持度
3. 对 F 中的频繁项按支持度降序排序，记为 L
4. 创建 FP－树的根结点
5. 对于 D 中每个阅读记录 trans，对 trans 中的频繁项，按 L 中的次序排序
6. 设排序后的频繁项表为[p|P]，其中 p 为第一个元素，P 是剩余元素的表
7. 调用 insert_tree([p|P],T)
8. if T 有结点 N 且满足 N.item－name＝p.item－name
9. N 的计数加 1
10. else
11. 创建一个新结点 N，计数为 1，链接到它的父结点 T，链接到具有相同 item-name 的结点
12. 如果 P 非空，递归调用 insert_tree(P,N)

算法 2 procedure FP_growth(Tree,α)

1. if Tree 包含单个路径 P then
2. for 路径 P 中结点的组合 β
3. 产生模式 β∪α，support＝β 中结点的最小支持度
4. else
5. for each ai 在 Tree 的头部 {
6. 产生一个模式 β＝ai ∪ α，support＝ai.support
7. 构造 β 的条件模式基，构造 β 的条件 FP－tree Treeβ
8. if Treeβ ≠ Ø then
9. 调用 FP_growth (Treeβ,β)；
10. }

3 图书阅读关联规则提取算法

对于频繁阅读的项集中的每个频繁项 $\alpha=\{\alpha_1,\alpha_2,\alpha_3,\cdots\}$，提取关联规则的方法是依次去掉其中的某项(如 α_1)，构成一个子项 $\beta=\{\alpha_2,\alpha_3,\cdots\}$，则形成一个规则 $\alpha_1\geqslant\beta$，α_1，β 分别为规则的前件和后件，计算其置信度：$Congfidence(\alpha_1\geqslant\beta)=support(\alpha)/support(\alpha_1)$，规则提取算法如下所示。

/＊关联规则提取算法＊/

Procedure ExactorRules()

输入：频繁项集集合 F；最小置信度 min_conf

输出：规则集合 R

1. For each α∈F
2. For each sub item $\alpha_1\in\alpha$
3. β＝Project(α,sub_α)
4. conf＝α. sup/α_1. sup
5. If conf＞＝min_conf
6. Rule rule＝new Rule(α_1＝＞β)
7. R. add(rule)
8. Return R

4 图书推荐算法的应用

下面以一个具体的实例来说明算法的应用。根据读者历史阅读记录提取读者库、图书库、阅读数据库，分别如表 1－表 3 所列。

表 1 读者库

读者编号(TID)	读者姓名	性别	年龄
T1	张兵	男	18
T2	高强	男	20
T3	刘军	男	25
T4	周丽	女	32
…	…	…	…

表 2 图书库

图书编号	图书名称	作者	出版社
a	计算机基础	亓常松	清华大学出版社
b	C 语言程序设计	谭浩强	清华大学出版社
c	数据结构	严蔚敏	清华大学出版社
d	操作系统	庞丽萍	人民邮电出版社
e	数据库系统概论	王珊	高等教育出版社
…	…	…	…

表 3 阅读数据库

读者	图书项集	预处理过的图书项集
T1	d,b,c,a	a,b,c,d
T2	b,f,a,e	a,b,e
T3	b,c,a	a,b,c
T4	d,b,c,f,a,e	a,b,c,d,e
T5	d,c,e	c,d,e
T6	b,c,e	b,c,e
T7	d,c,a	a,c,d
T8	d,b,c,g,a	a,b,c,d
T9	d,b,g,a	a,b,d

实例中阅读数据库包含 9 个读者，每个读者对应一个项集，其中包含该读者的图书阅读记录。每个项集中没有重复的数据项，若读者多次阅读同一本书，则该书在项集中只记录一次。表 1 中用 TID 表示读者的唯一标识，Items 表示项集。实例挖掘表 3 阅读数据库中的频繁项集，设最小支持度为 3，具体步骤如下。

首先，对阅读数据库进行预处理，包括对读者的所有项进行支持度统计，去掉每一个阅读记录中的非频繁项，同时将每个阅读记录中剩余的频繁项按其支持度递减的顺序进行排列。例如，通过对表 3 阅读数据库中每一项的支持度进行统计，得到以下频繁项：$L=\langle a:7,b:7,c:7,d:6,e:4\rangle$，其中数字代表该项的支持度，支持度相同的项按照其字母顺序进行排列。

然后，对预处理过的阅读数据库构建 FP-Tree。首先建一个根节点(root)，扫描第一条阅读记录 $T1=\{a,b,c,d\}$。为其中的每一项建立相应的节点，并记录该项当前的出现次数。扫描第 2 条阅读记录 $T2=\{a,b,e\}$。由于该阅读记录的前 2 项与第 1 条阅读记录相同，因此不再为此建立新的节点，只需把这 2 项所在节点的支持度加 1。第 3 项为 e，由于图 1 中 b 节点后面没有 e 节点，因此需要在 b 节点后新建一个 e 节点来存储 e 项并记录其支持度。同理，依次扫描第 3－9 条阅读记录，构建最终的 FP-Tree，如图 1 所示。

最后，在构建的 FP-Tree 上采取“分而治之”的策略挖掘频繁项集，如表 4 所列，并根据关联规则提取算法提取强关联规则作为系统推荐的依据，如表 5 所列。

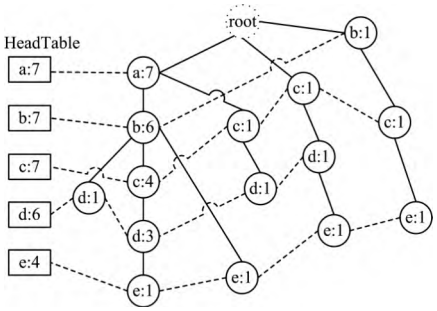

图 1　最终构建的 FP-Tree

表 4　频繁阅读项集

频繁项集	支持度
计算机基础	7
C 语言程序设计	7
数据结构	7
操作系统	6
数据库系统概论	6
计算机基础，C 语言程序设计	6
计算机基础，数据结构	5
计算机基础，操作系统	5
C 语言程序设计，数据结构	5
C 语言程序设计，操作系统	4
C 语言程序设计，数据库系统概论	3
数据结构，操作系统	5
数据结构，数据库系统概论	3
计算机基础，C 语言程序设计，数据结构	4
计算机基础，C 语言程序设计，操作系统	4
计算机基础，数据结构，操作系统	4
C 语言程序设计，数据结构，操作系统	3
计算机基础，C 语言程序设计，数据结构，操作系统	3

表 5　推荐规则

关联规则	置信度
C 语言程序设计，操作系统＝＞ 计算机基础	1
C 语言程序设计，数据结构，操作系统＝＞ 计算机基础	1
C 语言程序设计＝＞ 计算机基础	0.86
计算机基础＝＞ C 语言程序设计	0.86
操作系统＝＞ 计算机基础	0.83
操作系统＝＞ 数据结构	0.83
C 语言程序设计，数据结构＝＞ 计算机基础	0.8
计算机基础，数据结构＝＞ C 语言程序设计	0.8
计算机基础，操作系统＝＞ C 语言程序设计	0.8
数据结构，操作系统＝＞ 计算机基础	0.8
计算机基础，操作系统＝＞ 数据结构	0.8
计算机基础，数据结构＝＞ 操作系统	0.8
数据库系统概论＝＞ C 语言程序设计	0.75
数据库系统概论＝＞ 数据结构	0.75
C 语言程序设计，操作系统＝＞ 数据结构	0.75
计算机基础，数据结构，操作系统＝＞ C 语言程序设计	0.75
计算机基础，C 语言程序设计，操作系统＝＞ 数据结构	0.75
计算机基础，C 语言程序设计，数据结构＝＞ 操作系统	0.75
C 语言程序设计，操作系统＝＞ 计算机基础，数据结构	0.75
数据结构＝＞ 计算机基础	0.71
计算机基础＝＞ 数据结构	0.71
计算机基础＝＞ 操作系统	0.71
数据结构＝＞ C 语言程序设计	0.71
C 语言程序设计＝＞ 数据结构	0.71
数据结构＝＞ 操作系统	0.71
操作系统＝＞ C 语言程序设计	0.67
计算机基础，C 语言程序设计＝＞ 数据结构	0.67
计算机基础，C 语言程序设计＝＞ 操作系统	0.67
操作系统＝＞ 计算机基础，C 语言程序设计	0.67
操作系统＝＞ 计算机基础，数据结构	0.67

5　系统界面

图书阅读推荐系统基于 Android 移动平台，利用 Intellij IDEA 开发工具，使用 Java 语言实现，系统提供了图书浏览、图书查询、图书阅读、图书推荐等功能。系统登录界面、图书推荐界面分别如图 2、图 3 所示。

图 2　系统登录界面

图 3　图书推荐界面

结束语　本文在研究国内外个性化推荐相关算法的基础上，针对目前移动阅读市场的现状和实际需求设计开发了一款基于 Android 移动平台的电子图书阅读推荐系统。为了挖掘图书阅读的关联关系，提出将频繁项集挖掘算法 FP-Growth 应用到图书推荐系统中，根据读者的历史阅读记录，挖掘他们之间的关联关系，从中提取满足支持度和置信度阈值的强关联性规则，并根据规则进行图书智能推荐。实验证明，基于 Android 移动平台的图书阅读推荐系统能够为读者提供快速、准确的图书推荐服务。

（下转第 554 页）

本设计中 PC 机位于采油现场的监控处，因此其数据接收是实时并与上位机采集的数据是同步的。设计中上位机是 1 分钟采集一组压力和一组温度值，每 30 秒切换一次，即 PC 机是每一分钟接收一组数据。而为了降低功耗和准确性，传送到云端服务器的数据采用每 5 分钟接收一次，且这个数据是经过上位机软件 C++查表法得到的一个最准确值。实验于 2014 年 10 月份在胜利油田东营宪河采油厂所属油井进行，其深度大约为 2700m，接收一段时间的温度和压力数据，选择了 3 个时间段的数据，如表 1 所列。

表 1　PC 机和 Android 客户端接收的数据

时间/min	1	2	3	4	5
PC T/P	86.4/7.2	86.5/7.2	86.7/7.2	86.4/7.1	86.5/7.2
Android T/P			86.6/7.2		
PC T/P	81.2/8.1	82.5/8.0	82.6/8.1	82.5/8.3	81.1/8.6
Android T/P			82.2/8.3		
PC T/P	87.3/7.0	87.5/7.1	87.6/7.1	87.6/7.1	87.8/7.3
Android T/P			87.5/7.1		

结束语　1)实验进行将近一个月，设计的整个系统性能比较稳定，没有出现什么问题，对比较偏远地区采油井的采油温度和压力的监测很准确，对工作人员有很大帮助，且效果良好。

2)用手机可以在具有 3G 网络的地方对潜油电泵进行控制，实现了对油井远程控制与监控的功能，能够对油井开采中的采油泵起到保护作用，从而降低了运营成本。

STM32F072 处理器应进一步优化，改进安装工艺来提高系统可靠性与实用性，对潜油电泵实现更多参数的监测；同时可以应用更多的 ZigBee 终端节点，实现同时对多口采油井进行监测。

参 考 文 献

[1] 崔健，段振刚，刘志男. 基于物联网云平台的壁挂炉远程控制系统[J]. 计算机系统应用，2015，24(9)：56-60

[2] 周承民，张宗元. 油田工况数字化与无线视频监控[J]. 石油仪器，2003，17(5)：1-4

[3] 高守玮，吴灿阳. ZigBee 技术实践教程[M]. 北京：北京航空航天大学出版社，2006

[4] 唐雄燕. 宽带无线接入技术及应用—WiMAX 与 WiFi[M]. 北京：电子工业出版社，2006

[5] 王运红，何灵娜. 基于 Android 平台智能家居客户端的设计与实现[J]. 机电工程，2014，31(8)：1086-1089

[6] 覃征，王志敏，王利荣. 基于 Internet 的在线压缩传输模型[J]. 小型微型计算机系统，2002，23(2)：156-158

[7] TEXAS INSTRUMENTS. A True System-on-Chip Solution for 2.4-GHz IEEE 802.15.4 and ZigBee Applications[K]

[8] 江发全，杜坚. 基于 CC2530 的工业无线传感网节点设计[J]. 仪表技术，2013(3)：23-26

[9] 衣翠平，柏逢明. 基于 ZigBee 技术的 CC2530 粮仓温湿度检测系统研究[J]. 长春理工大学学报(自然科学版)，2011，34(4)：53-57

[10] 王风. 基于 CC2530 的 ZigBee 无线传感器网络的设计与实现[D]. 西安：西安电子科技大学，2012

[11] 梅思杰，邵永实，刘军，等. 潜油电泵技术(下册)[M]. 北京：石油工业出版社，2004

[12] Wen De-sheng. College of Mechanical Engineering, Yanshan Univer-sity, Qinhuangdao 066004, China. Theoretical analysis of output speed of multi-pump and multi-motor driving system[J]. Science China(Technological Sciences), 2011, 54(4): 992-997

[13] 章伟聪，俞新武，李忠成. 基于 CC2530 及 ZigBee 协议栈设计无线网络传感器节点平[J]. 计算机系统应用，2011，20(7)：184-187

[14] Wireless Medium Access Control (MAC) and physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Network (LR-WPANs): IEEE 802.15.4-2003Std

[15] 朱琎，杨占勇. 基于 CC2530 的无线振动监测传感器节点设计[J]. 仪表技术与传感器，2012(8)：56-59

[16] Wolters K M, Engelbrecht K P, Godde F, et al. Making it easier for older people to talk to smart homes: the effect eraly help prornpts[J]. Universal Access in the Information Society, 2010, 9(4): 311-325

(上接第 525 页)

参 考 文 献

[1] 王国霞，刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用，2012，48(7)：66-76

[2] 陈亮. 图书智能检索系统中的数据挖掘技术研究与应用[D]. 哈尔滨：哈尔滨工程大学，2012

[3] 黄洋. 基于聚类和项目类别偏好的协同过滤推荐算法研究[D]. 杭州：浙江理工大学，2013

[4] 高成. 基于标签主题建模的图书推荐系统研究[D]. 杭州：浙江大学，2014

[5] 荆月敏. 基于数据挖掘的图书馆书目推荐服务的研究[D]. 太原：中北大学，2014

[6] 邓娟，陈西曲. 基于用户兴趣变化的协同过滤推荐算法[J]. 武汉工业学院学报，2013(4)：48-51

[7] 吉红蕾. 基于高效用模式挖掘的推荐方法研究[D]. 北京：北方工业大学，2014

[8] 秦健. 基于信息可视化与数据挖掘的高校图书馆推荐系统的设计与实现[D]. 北京：北京交通大学，2014

[9] 黄平运. 关联规则算法在图书馆智能 OPAC 系统设计中的应用研究[J]. 电子技术与软件工程，2014(6)：26

[10] 胡文江. 基于关联规则与标签的好友推荐算法[J]. 计算机工程与科学，2013，35(2)：109-113

[11] 章志刚. 一种基于 FP-Growth 的频繁项目集并行挖掘算法[J]. 计算机工程与应用，2014，50(2)：102-106

[12] Achar A, Laxman S, Sastry P S. A unified view of the Apriori-based algorithms for frequent episode discovery[J]. Knowledge and Information Systems, 2012, 31(2): 1-28