Apriori算法在图书推荐服务中的应用与研究

林郎碟，王灿辉
（福州大学 软件学院，福建 福州 350108）

摘 要：数据挖掘是近年来数据库领域研究的热点问题之一。当今数字图书馆个性化服务已成为图书馆服务模式的主流，图书推荐服务是其重点之一。关联规则Apriori算法是数据挖掘的关键技术之一，主要是找出数据库中的所有频繁项集，然后由频繁项集产生关联规则。针对传统的Apriori算法存在的缺陷，利用“分割-整合”的思想改进了Apriori算法。将改进后的Apriori算法应用到图书推荐服务应用模型当中，并进行数据挖掘测试，通过与传统Apriori算法进行对比，改进后的Apriori算法的实际运行效果有明显的改进。

关键词：数据挖掘；关联规则；Apriori算法；图书推荐
中图分类号：TP311 **文献标识码**：A **文章编号**：1673-629X(2011)05-0022-03

Application and Research of Apriori Algorithm in Library's Book Recommendation Service

LIN Lang-die, WANG Can-hui
(Software College, Fuzhou University, Fuzhou 350108, China)

Abstract: Data mining is one of the research hotspots in database area. At present, digital personalized service has become the mainstream service model of library, and book recommendation is one of important services. Association rules Apriori algorithm is one of the key technologies in data mining, and its idea is find of frequent item sets in database, and produces association rules from the sets. Address shortcomings of Apriori algorithm, and introduce improved Apriori algorithm based on idea of "Division-Integrate". Then it applies the improved Apriori algorithm to books recommended service application model, and tests using data mining for service. By comparing with the traditional Apriori algorithm, the improved algorithm achieves obvious improvement in actual running.

Key words: data mining; association rule; Apriori algorithm; book recommendation

0 引 言

数据挖掘是从大量的数据中挖掘哪些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识[1]。关联规则是数据挖掘的典型方法，它是描述在数据库中数据项之间同时出现的规律的知识模式。关联规则的分析方法用于隐藏在大型数据集中令人感兴趣的联系，所发现的联系可以用关联规则或频繁项集的形式表示[2]。关联规则挖掘问题首先是由R. Agrawal等人于1993年提出的，而后又进一步提出了著名的Apriori算法，该算法的主要思想是首先寻找给定数据集中的频繁项集，然后通过频繁项集生成强关联规则[3]。

收稿日期：2010-09-19；**修回日期**：2010-12-25
基金项目：福建省自然科学基金(2009J05142)；福州大学校人才基金(2007-2012)
作者简介：林郎碟(1964-)，女，福建泉州人，硕士，研究方向为数据库与数据挖掘；王灿辉，副教授，硕士，研究方向为软件工程，数据库、分布式系统。

1 Apriori算法

Apriori算法是由Rakesh Agrawal和Rnamakrishnan Srikant在1994年提出的关联规则的经典算法，它是所有关联规则挖掘算法的核心。Apriori算法将关联规则挖掘划分为两个子问题：1)在事务集D中寻找满足所有最小支持度阈值min_sup的频繁项集。2)利用频繁项集来生成所有满足最小置信度阈值min_conf的关联规则。其中的子问题1是Apriori算法所要解决的核心问题。Apriori算法主要通过迭代的方法来求出事务集D中所有的频繁项集。

Apriori算法利用连接和剪枝两个步骤寻找出事务之间的强关联规则从而在商品零售业、网站开发、医学领域、金融投资业、图书管理系统等大型数据库中得到广泛的应用。

Apriori算法的计算复杂度受以下几个因素的限制[4]：1)最小支持度阈值和最小置信度阈值；2)项数(维度)；3)事务数；4)事务的平均宽度。

在实际应用中，发现Apriori算法存在如下一些主

要的缺陷[5]:1)频繁的扫描数据库;2)不适用于稠密集的关联规则挖掘;3)可能生成的关联规则过于庞大。

近年来,不少学者针对 Apriori 算法的缺陷对算法提出不同的改进策略,概括起来主要有以下的几类:1)基于逆向运算的优化策略[6];2)基于哈希表的优化策略[7];3)基于划分的优化策略[8];4)基于事务压缩的优化策略[9];5)基于采样的优化策略[10];6)基于数据库结构变换的优化策略[11]。

2 Apriori 算法的改进及在图书推荐服务中的应用

当今数字图书馆正将逐渐往个性化和智能化的方向前进,图书推荐服务在数字化图书馆个性化服务中非常重要,一方面能提高用户的满意度,另一方面可以让信息得到更好的利用。但是面对庞大的读者借阅记录数据时,传统的 Apriori 算法由于需要大量的扫描借阅数据库,所以算法的运行效率低下,因此需要针对图书馆借阅记录的实际特点,对传统 Apriori 算法加以改进,提高算法在实际运用中的性能。

在大量的研究读者的借阅数据之后,发现读者对图书的借阅存在着一定的关联。不同的学科之间存在着关联,不同类型的读者对图书的借阅也存在着一定的关联模式[12]。大部分读者在一次特定的借阅中,往往只会借阅某一类别的图书,而通过传统 Apriori 算法对借阅数据进行挖掘后,也可以发现,关联规则的左部和右部均为同种图书的关联规则占总挖掘关联规则的 90% 以上。

根据上述借阅数据的特点,对传统 Apriori 算法加以改进,将改进的算法称为 L_Apriori 算法。根据《中图法》对图书的分类,可以通过图书分类号信息将整个图书馆分为若干个子图书馆,每个图书馆都只包含自己所属种类的图书。同样的,也可以将借阅数据库根据图书的分类号分隔成若干个子借阅数据库。L_Apriori 算法利用了"分割--整合"的思想,先在各个子数据库进行关联规则的挖掘,再将所挖掘到的关联规则进行整合,从而达到了对整个数据库进行关联规则挖掘的目的。假设借阅数据库有 M 条借阅数据记录,N 条书目记录,现将借阅记录平均分成 n 个子借阅数据库和书目记录库,这样每个子数据库包含 M/n 条借阅记录,N/n 条书目记录,在搜寻频繁 1 项集的时候,使用传统 Apriori 算法所需的时空开销为 $M \times N = MN$,而 L_Apriori 算法只需要 $n \times M/n \times N/n = MN/n$ 的时空开销,由此可见,在不使用其他改进策略的情况下,L_Apriori 算法已经比传统 Apriori 算法效率高。

其次 L_Apriori 算法在挖掘频繁项集的时候采用"剪枝"策略,在每一轮寻找候选项集的时候,会根据判断不断的"剪去"非频繁项集,从而逐渐减小候选项集的大小,进而减少扫描借阅数据库的次数,达到算法效率的进一步提升。

基于以上的考虑,将 L_Apriori 算法描述如下:

1)首先将读者借阅数据库经过数据预处理压缩成读者借阅事务数据库。

2)按照《中图法》的图书分类标准,将系统中的书目数据和借阅事务数据分成各个子数据库,每个子数据库只包含一类的图书和包含所有这类图书的借阅数据。

3)首先扫描第一个子数据库,对子数据库的借阅事务数据进行扫描,统计每一本书的借阅次数,并通过进行最小支持度的比较,搜索出频繁 1 项集。

4)把所生成的频繁 1 项集相互进行连接生成候选频繁 2 项集,此时 $K = 2$。随后再次扫描借阅事务数据库,统计每一候选 2 项集在事务数据库当中的出现次数,通过最小支持度阈值筛选出真正的频繁 2 项集。

5)根据所生成的频繁 K 项集相互进行连接生成频繁 $K + 1$ 项集,随后扫描借阅事务数据库,统计每一候选 $K + 1$ 项集在事务数据库当中的出现次数,通过最小支持度阈值筛选出频繁 $K + 1$ 项集。将此过程进行反复循环,直到再也生成不了新的频繁项集为止。在生成候选频繁 K 项集的时候,还进行"剪枝"过程。例如,若 $(a,b),(a,c),(a,d),(b,c),(b,d)$ 是算法所生成的所有频繁2项集,则在生成候选频繁3项集时只保留 (a,b,c) 这个项集,因为只有这个项集的所有子集均为频繁项集,而 $(a,b,d),(a,c,d),(b,c,d)$ 这三个项集的部分子集并不是频繁项集,所以在候选频繁项集生成时被"剪枝"。

6)将所生成的所有频繁项集进行整合,并计算每个频繁项集所生成的关联规则的置信度,通过与最小置信度阈值进行比较筛选出所有的关联规则。

7)重复 2)~6)步,对每一个子数据库进行关联规则的挖掘,直到所有的子数据库的关联规则都被挖掘出来。

8)将所有子数据库所挖掘出的关联规则进行合并,并将合并后的关联规则更新至系统关联规则库当中。

3 L_Apriori 算法改进的结果分析

3.1 测试环境与测试数据

针对 L_Apriori 算法在图书推荐服务中的应用所建立的测试系统使用 SQL Server 2005 数据库管理系统、Visual C#语言进行开发。所有实验基于双核 CPU 1.86GHz 以及 2G 内存,并运行在 Windows XP 操作系

统之上。本次实验所使用的测试数据是某高校图书馆的读者借阅数据。

3.2 改进算法性能测试

Apriori 算法和 L_Apriori 算法在基于统一的最小支持度上的时间性能测试，实验测试事务数据量全部为 12000 条，实验结果如图 1 所示。当最小支持度的值比较高的时候，传统 Apriori 算法和 L_Apriori 算法时间开销相差并不是很大，这是因为较高的支持度会导致候选项集的减少，从而减少算法在扫描数据库上的时间开销。当最小支持度设置的较低时，传统算法与改进算法的时间开销就开始逐步拉大，这是因为传统算法会比改进算法产生更多的候选项集，进而加大算法扫描数据库的时间开销。

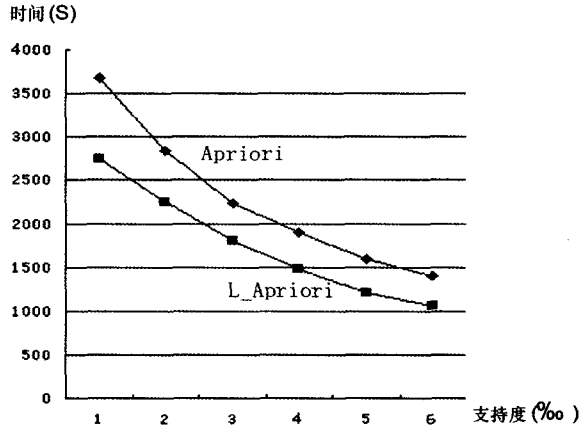

图 1 基于同一最小支持度的时间性能测试结果

Apriori 算法和 L_Apriori 算法在基于统一的事务数下的时间性能测试，作为实验测试条件的事务数据量从 2500 条到 15000 条之间变化，实验测试的最小支持度数值全部为 4‰，实验结果如图 2 所示。

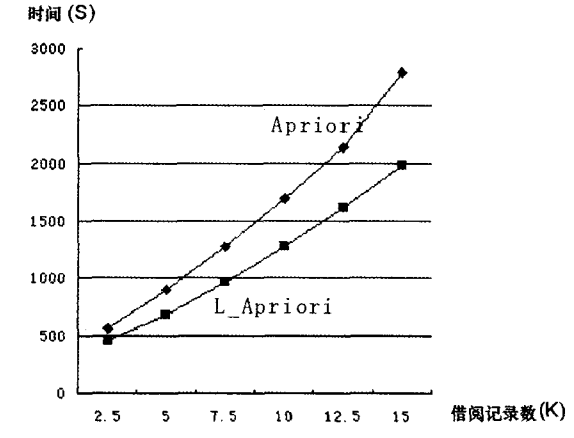

图 2 基于统一事务数据量的时间性能测试结果

当事务数据量比较少的时候，传统算法和改进算法的时间开销相差很小，这是因为事务数据量小的时候，传统算法和改进算法不断扫描事务数据库所花的时间开销就会相对较小，所以两种算法的总时间开销不会相差太大。当事务数据量比较大的时候，扫描一遍事务数据库的时间开销就会加大，由于改进算法扫描事务数据库的次数比传统算法的次数少，所以改进算法的时间开销比传统算法的时间开销小。

3.3 挖掘性能测试结果

比较传统 Apriori 算法和 L_Apriori 算法在基于统一的最小支持度上的挖掘性能测试，作为实验测试条件的最小支持度在 1‰到 6‰范围之内变化，实验测试事务数据量全部为 12000 条，最小置信度为 60%，对算法所挖掘出来的关联规则进行图书关键字匹配度计算来判断此条关联规则是否对读者有意义，结果如表 1。

表 1 基于统一最小支持度的挖掘性能测试结果

支持度	算法	总挖掘数量	无意义的挖掘数量	无意义占比例
1‰	传统	83987	14661	17.4%
	改进	70321	3254	4.6%
2‰	传统	53378	9173	17.1%
	改进	44268	2001	4.5%
3‰	传统	31039	5275	16.9%
	改进	24679	1373	5.5%
4‰	传统	16775	2988	17.8%
	改进	13432	866	6.4%
5‰	传统	9543	1683	17.6%
	改进	6755	331	4.9%
6‰	传统	4956	775	15.6%
	改进	4190	174	4.2%

从结果上来看，传统算法挖掘无意义的关联规则数量占总数量在 17% 左右，而改进算法挖掘无意义的数量占总数量在 5% 左右，由此可见改进算法比传统算法在挖掘无意义的关联规则数量上要少，但是，在无意义的关联规则占总量少的条件下，改进算法也比传统算法少挖掘出一些有意义的关联规则，所以在挖掘性能测试的测试当中，改进算法比传统算法的精确度要略微下降。

3.4 L_Apriori 算法性能分析

经过时间性能分析测试和挖掘性能分析测试，可以得出改进后的 L_Apriori 算法比传统 Apriori 算法性能更高效，尤其是在数据量十分庞大的数字图书馆数据库当中进行挖掘时，L_Apriori 算法能节省更多的时间，虽然比传统算法少挖掘出一些关联规则，但也同样避免了大量的无意义的关联规则被挖掘出来，从而节省了一部分的空间开销，提升了挖掘出的关联规则的整体质量，进一步证明了改进的 L_Apriori 算法在效率上的优越性。当然，L_Apriori 同样存在一些缺陷，它是使用分而治之的思想提出的改进策略，所以在挖掘的过程中不同类别的图书之间存在的关联规则就无法

（下转第 28 页）

3 结束语

软件回归测试的关键是测试用例的优化选择[12]。文中借用数据挖掘中的规则分类技术给出一种等价类划分方法，主要介绍了在回归测试中利用基于决策树规则的分类技术来划分等价类的方法模型。可借由此方法编程实现测试用例的自动生成工具。该方法模型采用基于规则的排序策略对决策树规则进行排序，按分类规则将测试用例划分为若干个等价类，然后在每个类中选择少数有代表性的测试用例进行测试，测试成本，实现最小回归测试集的生成。尽管数据挖掘技术在其他方面得到了广泛的应用，但是在工程实践中，某些方法的效率和有效性还需要提高及检验；随着计算机学科技术的发展和研究的深入，将数据挖掘的相关技术越来越多的应用到软件测试中会成为软件测试有效方法之一。

参考文献：

[1] Rothermel G, Harrold M J. Analyzing regression test selection techniques[J]. IEEE Transactions on Software Engineering, 1996,22(8):529-551.

[2] Last K,Friedman M. The Data Mining Approach to Automated Software Testing[C]//Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]:[s. n.],2003:388-396.

[3] 许 静,陈宏刚,王庆人. 软件测试方法简述与展望[J]. 计算机工程与应用,2003,39(13):75-78.

[4] 李克文,杨志霞. 基于回归测试模型的用例集的优化方法研究[J]. 微计算机应用,2008,29(10):7-11.

[5] 谭俊璐,武建华. 基于决策树规则的分类算法研究[J]. 计算机工程与设计,2010,31(5):1017-1019.

[6] Chen Leida. Data mining methods, applications, tools[J]. Information System Management,2000,17(1):65-70.

[7] 马 菁,顾景文. 决策树在软件测试用例生成中的应用[J]. 计算机技术与发展,2008,18(2):66-69.

[8] Han Jiawei,Kamber M. 数据挖掘：概念与技术[M]. 北京：机械工业出版社,2008.

[9] Tan Pangning,Steinbach M,Kumar V. 数据挖掘导论[M]. 北京：人民邮电出版社,2006:127-136.

[10] 王小丽,段永颢. 软件回归测试用例选取方法研究[J]. 空间控制技术与应用,2010,36(3):47-49.

[11] Wu Jianhua,Song Qinbao,Shen Junyi. A novel association rule mining based missing nominal data imputation method[C]// Proceedings of the Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. [s. l.]:[s. n.],2007:44-249.

[12] 曾 强,洪 玫,杨昊苏,等. 软件回归测试中的自动化测试生成方法[J]. 计算机应用研究,2009,26(6):2349-2351.

(上接第 24 页)

被改进算法挖掘出来，虽然这些关联规则占总关联规则的比例不大。

4 结束语

文中对 Apriori 算法的复杂性及存在的缺陷进行分析，提出了 L_Apriori 算法，并在图书推荐服务进行了应用。测试发现以改进后的 L_Apriori 算法为核心的数据挖掘引擎比传统 Apriori 数据挖掘引擎在图书推荐服务模型中能发挥出更大的作用。通过对核心算法的改进，L_Apriori 算法能更好地应用于图书推荐服务模型中，为读者提供更为方便快捷的个性化服务。

参考文献：

[1] 康敏旸,张 安. 改进的 Apriori 数据挖掘算法的应用[J]. 火力与指挥控制,2009,34(10):111-114.

[2] 黄 鹤. 关联规则算法综述[J]. 软件导刊,2009,8(3):56-58.

[3] 陈则芝,李冬梅. 数据挖掘关联规则 Apriori 算法的优化[J]. 山西大同大学学报(自然科学版),2008,24(4),35-37.

[4] Han Jiawei,Kamber M. 数据挖掘：概念与技术[M]. 范明,孟小峰,译. 北京：机械工业出版社,2006.

[5] 鲍 静. 关联规则挖掘及其在图书流通数据中的应用研究[D]. 合肥：合肥工业大学,2007.

[6] 张 昀. 数据挖掘中一种改进的 Apriori 算法[J]. 软件导刊,2009,8(11):87-88.

[7] 陈文庆,许 棠. 关联规则挖掘 Apriori 算法的改进与实现[J]. 微机发展(现更名：计算机技术与发展),2005,15(8):155-157.

[8] Savasere A,Omieeinski E,Navathe S. An Efficient Algorithm for Mining Association Rules In Large Databases[C]//Proceedings of the 21st International Conference on Very Large Databases. New York:ACM,1995:432-443.

[9] Agrawal R,Srikant R. Fast Algorithms for Mining Association Rules[C]//Proceedings of the 20th International Conference on Very Large Databases. Santiago, Chile:Morgan Kaufmann Publisher,1994:487-499.

[10] Toiconen H. Sampling Large Databases for Association Rules [C]//Proceedings of the 22nd International Conference on Very Large Databases. Bombay, India: Morgan Kaufmann Publisher, 1996:134-145.

[11] 卢 露,丁才昌. 关联规则中 Apriori 算法改进的研究[J]. 长江大学学报(自然科学版),2009,6(2):241-243.

[12] 苟元琴,王钧玉. 关联规则在图书馆读者借阅记录中的挖掘应用[J]. 科技信息,2009(17):356-357.