文章编号: 1673-1549(2011)01-0066-05

关联规则挖掘的 Apriori算法综述

赵洪英[1], 蔡乐才[2], 李先杰[1]

(1.四川理工学院电子与信息工程学院, 四川 自贡 643000; 2 四川理工学院计算机学院, 四川 自贡 643000)

摘 要: 关联规则挖掘是数据挖掘研究领域中的一个重要任务,旨在挖掘事务数据库中有意义的关联。随着大量数据不停的收集和存储,从数据库中挖掘关联规则显得越来越有必要性,关联规则挖掘的 Apriori算法是数据库挖掘的最经典算法并得到广泛应用,在介绍关联规则挖掘和 Apriori算法的基础上,发现 Apriori算法存在着产生候选项目集效率低和频繁扫描数据等缺点。综述了 Apriori算法的主要优化方法,并指出了 Apriori算法在实际中的应用领域,提出了未来 Apriori算法的研究方向和应用发展趋势。

关键词: 数据挖掘; 关联规则; Apriori算法; 综述

中图分类号: TP391.4　　文献标识码: A

引 言

现在,数据挖掘作为从数据中获取信息的有效方法,越来越受到人们的重视。关联规则挖掘首先是用来发现购物篮数据事务中各项之间的有趣联系。从那以后,关联规则就成为数据挖掘的重要研究方向,它是要找出隐藏在数据间的相互关系。定义为,设 $I=\{I_1, I_2, \cdots I_m\}$是 m个不同项的项集, $X\in I$, $Y\in I$,并且 X和 Y是不相交的项集,即 $X\cap Y=\Phi$。关联规则的属性可以用以下三个参数描述: 一是支持度, (support)定义为全体事务集 T中有 s%的事务同时支持事务集 X和 Y,则称 s%为关联规则 $X\rightarrow Y$的支持度。支持度表示规则的频繁程度,用 $S(X\rightarrow Y)$表示。其中,最小支持度用 Minsup表示。二是置信度(confidence),定义为全体事务集 T中支持事务集 X的事务中,有 c%的事务同时也支持事务集 Y, c%为关联规则 $X\rightarrow Y$的置信度。置信度表示规则的强度,用 $C(X\rightarrow Y)$表示。其中,最小置信度用 Minconf表示。三是频繁项集,定义为支持度不小于最小支持度 (minsup)的事务集,称为频繁项集。

关联规则的挖掘问题就是在事务数据库 D中找出具有用户给定的满足一定条件的最小支持度 Minsup和最小置信度 Minconf的关联规则。关联规则的挖掘一般分为以下两个步骤:

(1)找出存在于事务数据库中的所有频繁项集。

(2)用频繁项集生成关联规则,即对于每个频繁项集 X,若 $Y\in X$, $Y\neq\Phi$,且 $c(Y\rightarrow(X-Y))\geqslant Minconf$,构成关联规则 $Y\rightarrow(X-Y)$。

本文分析了 Apriori算法,指出其存在的几个缺陷,提出了针对缺陷的主要改进优化的方法,列举了 Apriori算法的几个应用领域,展望了 Apriori算法的未来研究方向。

1 Apriori算法

1.1 算法概述

Apriori算法是第一个关联规则挖掘算法,也是最经典的算法。它利用逐层搜索的迭代方法找出数据库中项集的关系,以形成规则,其过程由连接(类矩阵运算)与剪枝(去掉那些没必要的中间结果)组成。该算法中项集(Itemset)的概念即为项的集合。包含 K个项的集合为 k项集。项集出现的频率是包含项集的事务数,称为项集的频率。如果某项集满足最小支持度,则称它为

收稿日期: 2010-07-21

基金项目: 四川省科技厅支撑计划项目(2008FZ0109); 四川省教育厅科技项目(2007ZL048)

作者简介: 赵洪英(1980-),女,河南驻马店人,硕士生,主要从事智能信息处理方面的研究。

频繁项集。

1.2 算法步骤

步骤如下:

(1)设定最小支持度 s和最小置信度 c。

(2) Apriori算法使用候选项集。首先产生出候选的项的集合,即候选项集,若候选项集的支持度大于或等于最小支持度,则该候选项集为频繁项集。

(3)在 Apriori算法的过程中,首先从数据库读入所有的事务,每个项都被看作候选 1-项集,得出各项的支持度,再使用频繁 1-项集集合来产生候选 2-项集集合,因为先验原理保证所有非频繁的 1-项集的超集都是非频繁的。

(4)再扫描数据库,得出候选 2-项集集合,再找出频繁 2-项集,并利用这些频繁 2-项集集合来产生候选 3-项集。

(5)重复扫描数据库,与最小支持度比较,产生更高层次的频繁项集,再从该集合里产生下一级候选项集,直到不再产生新的候选项集为止。

在此算法中要不断地重复两个步骤:连接和剪枝。具体内容如下:

(1)连接。为找 F_k,通过 F_{k-1}与自己连接产生候选 k-项集。该候选项集的集合记做 L_k。设 F_1和 F_2是 F_{k-1}中的项集。执行连接 $F_{k-1}\infty F_{k-1}$,其中 F_{k-1}的元素 F_1和 F_2是可以连接的。

(2)剪枝。L_k的成员不一定都是频繁的,所有的频繁 k-项集都包含在 L_k中。扫描数据库,确定 L_k中每个候选集计数,并利用 F_{k-1}剪掉 L_k中的非频繁项,从而确定 F_k。

2 分析 Apriori算法

2.1 算法代码

下面是 Apriori算法产生频繁项集的伪代码,令 C_k为候选 k-项集的集合,而 F_k为频繁 k-项集的集合:

```
1: k = 1
2: Fk = { i | i ∈ I ∧ σ({i}) ≥ N* minsup }
3: repeat
4:   k = k + 1
5:   Ck = apriori-gen(Fk-1)
6:   for 每个事务 t ∈ T do
7:     Ct = subset(Ck, t)
8:     for 每个候选项集 c ∈ Ct do
9:       σ(c) = σ(c) + 1
10:    end for
11:  end for
12: Fk = { c | c ∈ Ck ∧ σ(c) ≥ N* minsup }
13: until Fk = φ
14: Result = ∪F
```

Apriori算法基于一个频繁项集中任一子集也应该是频繁项集的性质,使用一种逐层搜索的迭代方法,k-项集用于 (k+1)-项集。其算法流程如下:首先遍历目标数据库一次,记录每个项目或属性的出现次数,即计算每个项目的支持度,收集所有支持度不低于用户最小支持度的项目构成频繁 1-项集 L_1,然后链接 L_1中所有的元素形成候选 2项集 C_2,再次遍历事务数据库,计算 C_2中每个候选 2-项集的支持度,收集所有支持度不低于用户最小支持度的项目构成频繁 2-项集 L_2,再链接 L_2形成 C_3,遍历数据库得 L_3,反复执行以上过程,直到没有候选项集为止。在整个过程中,多次循环,产生大量的候选集,验证环节需要反复扫描可能很大的交易数据库。

由上面的分析可知, Apriori算法要求多次扫描可能很大的数据库,如果频集最多包含 20个项,那么需要扫描数据库 20遍,这需要很大的 I/O负载。

2.2 算法的缺陷

在 Apriori算法中候选项集是逐层产生的,而产生此层的频集必须要扫描整个数据库一次,然后再结合频集产生下一层级的候选项集合,直到频集无法结合产生候选项集。该算法一定要等到扫描完整个数据库后才做结合,因为在扫描过程中,有些候选项集在若干的区段中的支持度已大于等于制定的最小支持度,因此在这些若干个区段后,便可以找出频集,并直接结合产生下一层级的候选项集。基于这些原因, Apriori算法要消耗许多时间,这些时间主要消耗在以下三个方面:

(1)利用 k频集连接产生 k+1候选项集,判断连接条件时比较次数太多。假设项集个数为 m的频集集合 L_k,判断连接条件时比较的时间复杂度是 $O(k^* m^2)$,并且其中 m的值会很大。因为,假设 1_项集集合中项集个数为 10^2,那将会产生 5000个候选 2_项集。最坏情况下即假设发现了一个长度为 100的频集,则将产生约 10^{30}个候选项集。

(2)对任意一个 $c \in C_k$ 判断 c的 k个(k-1)_子集是否都在 L_{k-1}中。在这个过程中,对于 c在最好情况下只需扫描一次 L_{k-1},即第一个(k-1)_子集不在 L_{k-1}中。在最坏情况下则需要一直扫描到第 k次时才发现第 k个(k-1)_子集不在 Lk-1中或 k个(k-1)_子集都在 Lk-1中。于是,在平均情况下,对任意一个 $c \in Ck$扫描 L_{k-1}次数为 $|L_{k-1}| \times k/2$,那么对所有候选 k_项集需要扫描次数为 $|C_k| \times |L_{k-1}| \times k/2$。

(3)为了得到所有 C_k(k=1, 2, …, n)候选频集的支持度,需要扫描数据库 n次。

3 Apriori算法的优化方法

虽然 Apriori算法是关联规则挖掘的最经典的算法,但由于其固有不足,许多学者就如何减少扫描数据库的次数以及减少 I/O的负载做出了研究,提出了一些优化的算法。

(1)基于用户感兴趣项集和项集重要性的方法。首先从数据库中利用某些用户感兴趣的项,从数据库的所有项的集合中选择出一个子集作为挖掘对象,然后对数据库进行一次扫描,实现用事务标识号来表示项目集。在产生项目集后,对项目集中的元素赋以权值,然后利用引入了权值的支持度函数计算项集的支持度以产生频集,最后的工作就是从这些频集中产生关联规则。

(2)基于划分的方法。算法先把数据库从逻辑上分成几个互不相交的块,每次单独考虑一个分块并对它生成所有的频集,然后合并产生的频集生成所有可能的频集,最后计算这些项集的支持度。这里分块的大小选择要使每个分块可放入主存,每个阶段只需被扫描一次。而算法的正确性是由每一个可能的频集至少在某一个分块中是频集保证的。

(3)基于矩阵的方法。它主要是将矩阵的思想应用到 Apriori算法当中,把事务数据库表示成矩阵的形式。具体方法为:对每一成员按一序列排列,事务集也按一序列进行排列。成员分别表示行向量,事务表示列向量,若第 m个成员在第 n个事务中,则矩阵的第 m行,第 n列的值为 1,否则为 0,称其为数据库的布尔矩阵。矩阵的行向量之和为成员出现的次数,则项集的支持记数可求出。对于二项集 $\{M_m, N_n\}$只需扫描第 m行与第 n行即可,它们同一列的值均为 1的个数,即为二项集 $\{M_m, M_n\}$的支持记数,依此类推。只需扫描矩阵的第 $m_1, m_2, \cdots, m_k$ 行,它们同一列的值均为 1的个数即为 k项集 $\{M_{m1}, M_{m2}, \cdots, M_{mk}\}$的支持记数。

(4)基于采样的方法。基于前一遍扫描得到的信息进行组合分析,得到一个改进的算法,即在计算 k-项集时,如果认为某个(k+1)-项集可能是频集时,就并行地计算这个(k+1)-项集的支持度,该算法需要的总的扫描次数通常少于最大的频集的项数。

(5)动态项集计数。该技术动态地评估已被计数的所有项集,不像 Apriori算法仅在每次完整的数据库扫描之前确定新的候选,它可以在任何点添加,一旦一个项集的所有子集被确定为是频繁的,就可以启动对该项集支持度的计算。因此,该算法所需的数据库扫描次数要比 Apriori算法少。

(6)压缩数据库事务集。有三个优化策略,一是,产生一项数据库 D的每个事务的项目计数项 n,每次扫描计数前比较 n与 k,如果 n小于 k,则可以忽略扫描本事务,同时置 n=0;二是,候选项计数过程中,如果数据库 D的某事务及其项目子集未被计数,则置 n=0;三是,首次支持度裁减后,比较非频繁项目集项目数和频繁项目集项目,取小值集进行剪枝操作。这样可以提高剪枝效率。

(7)采用项编码方法。该算法的主要思想是对所有的项,根据它在交易中出现的记录进行编码,在编码的同时就可以统计出项的支持度并生成频繁 1-项集。然后通过对不同编码进行“与”的运算来得到频繁二项集,并根据 Apriori算法的大项目集性质修改简化编码。如此循环最终得到符合关联规则的频集。从以上描述可以看出该算法只需要扫描一遍数据库,并且大幅减少了侯选集数量。

此外,还有基于杂凑等优化方法。

4 Apriori算法的应用

4.1 农业病虫害分析

随着科技的发展,自然环境遭到不同程度的破坏,致使各种害虫繁衍很快,但不同的害虫对环境的要求是不一样的,有的适合在温度较低的环境里生存,有的则适合在较高温度的环境里生存,还有其它各种不同的生存环境,比如湿度等。为了了解各种害虫的生理特点,更好的除虫,有必要对各种害虫的数量和生存的环境条件做一下分析。Apriori算法能很好的解决这一问题。

例如对水稻二化螟

害虫的分析,能根据环境变化,对害虫更好的消除。

4.2 试卷成绩分析

将关联规则 Apriori算法应用于试卷成绩分析中,首先对数据进行预处理,然后使用 Apriori算法挖掘学生各科目试卷成绩的优良影响关系,最终产生关联规则,用以指导学生的学习及今后的工作。

4.3 英语教师课堂话语分析

为在教学过程中提高学生认知和语言习得能力,运用关联规则的经典挖掘算法 Apriori研究英语教师口语语料分布特点,建立教师提问语、指令语和母语使用之间的关联性,并结合 Bloom的认知发展类型理论分析学习者思维变化能力与人的认知能力之间的关系。

4.4 电子商务中的应用

随着数据库技术的迅速发展以及数据库管理系统的广泛应用,电子商务网站积累的数据越来越多,面对海量的存储数据,如何从中发现有价值的信息或知识是一项非常艰巨的任务。关联规则的发现是数据挖掘中最成功和最重要的一项任务,它的目标是发现数据集中所有的频繁模式。

4.5 科学数据分析

在地球科学数据分析中,关联模式可以揭示海洋、陆地和大气过程之间的有意义的关系。这些信息能够帮助地球科学家更好的理解地球系统中不同的自然力之间的相互作用。

5 未来研究方向及应用发展趋势

社会信息量在不断更新变化,潜在的规则也在不断变化着,算法的研究是一个十分复杂的问题。对于关联规则挖掘的 Apriori算法的未来研究方向,我们觉得可以在以下几个方向继续深入:(1)提高算法效率;(2)算法的进一步优化;(3)在关联规则挖掘的过程中,如何与用户进行交互,在挖掘的过程中结合用户的领域知识,生成可视化的结果。

对于关联规则挖掘的 Apriori算法的应用发展趋势,应该涉及到以下应用领域:(1)父母学历的高低与子女的个数之间的关联规则,有利于更好的制定计划生育政策,从而促进社会更好的发展;(2)智能化设备,如靠识别语音的自动门等;(3)工作效率与学历高低之间的关联规则;(4)人的血型与成功几率之间的关联规则等。

参 考 文 献:

[1] Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(6): 914-925.

[2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database[J]. In Proc. ACM SIGMOD Intl. Conf. Management of Data, 1993, 1(1): 207-216.

[3] 范 明,范宏建,译. 数据挖掘导论[M]. 北京:人民邮电出版社, 2006.

[4] 陈 安,陈 宁,周龙骧,等. 数据挖掘技术及其应用[M]. 北京:科学出版社, 2006.

[5] 范 明,孟小峰,译. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2001.

[6] 颜雪松,蔡之华. 一种基于 Apriori的高效关联规则挖掘算法的研究[J]. 计算机工程与应用, 2002, 10: 209-211.

[7] 胡吉明,鲜学丰. 挖掘关联规则中 Apriori算法的研究与改进[J]. 计算机技术与发展, 2006, 16(4): 99-101.

[8] 刘君强,孙晓莹,潘云鹤. 关联规则挖掘技术研究的新进展[J]. 电子科技大学学报, 2004, 31(1): 110-113.

[9] Ng R T, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining[J]. In: Proc. of the 18th Intl. Conf. on Very Large Data Bases, 1994, 1(1): 144-155.

[10] Chen M S, Han J W, Yu P S. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866-883.

[11] Park J S, Chen M S, Yu P S. Using a hash-based method with transaction trimming for mining association rules[J]. Knowledge and Data Engineering IEEE Transactions, 1997, 9(5): 813-825.

[12] 崔立新,范森森,赵春喜. 相联规则发现方法及算法[J]. 计算机学报, 2000, 23(2): 216-220.

[13] 赵春玲,宁红云. Apriori算法的改进及其在物流信息挖掘中的应用[J]. 天津理工大学学报, 2007, 23(1): 30-33.

[14] 姜红艳. Apriori关联算法在学生成绩中的应用[J]. 鞍山师范学院学报, 2007, 9(2): 48-50.

[15] Toivonen H. Sampling Large Databases for Association Rules[C]. Proceedings of the 22nd International Conference on Very Large Data-base[M]. Bombay, India, 1996.

[16] 牛丽敏. Apriori算法分析与改进综述[J]. 桂林电子

科技大学学报, 2007, 27(1): 27-30.
[17] 高振中, 杨小劲. 利用项编码方法改进 apriori算法[J]. 计算机时代, 2009(1): 27-29.
[18] 徐章艳, 刘美玲, 张师超, 等. Apriori算法的三种优化方法[J]. 计算机工程与应用, 2004, 40(36): 190-193.
[19] JiaweiHan, MichelineKamber. DataMining: Concepts and Techniques[M]. Beijing: ChinaMachine Press, 2001.
[20] 陈敏艳. 基于矩阵方法优化 Apriori算法[J]. 内蒙古科技与经济, 2008, 16(17): 69-70.
[21] Chen F, Drezner Z, Ryan J K. Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Leadtime and Information[J]. Management Science, 2000, 46(3): 436-443.
[22] 吴 斌, 肖 刚, 陆佳炜. 基于关联规则挖掘领域的Apriori算法的优化研究. 计算机工程与科学[J]. 2009, 31(6): 116-118.
[23] 袁万莲, 郑 诚, 翟明清. 一种改进的 Apriori算法[J]. 计算机技术与发展, 2008, 18(5): 52-53.
[24] 庄晓毅, 张忠能. 一种改进的关联规则挖掘算法[J]. 计算机工程, 2004, 30(14): 128-129.
[25] 李志勇. 深入分析关联规则 Apriori算法[J]. 现代计算机, 2009(5): 48-49.
[26] SayginY, Verykios V S, Clifton C. Using Unknowns to Prevent Discovery of Association Rules[J]. ACM SIGMOD Record, 2001, 30(4): 45-54.

Overview of Association Rules Apriori Mining Algorithm

ZHAO Hong-ying[1], *CAI Le-cai*[2], *LI Xian-jie*[1]
(1. School of Automation Electronics and Information, Sichuan University of Science & Engineering, Zigong 643000, China;
2. School of Computer Science, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract Mining association rules, designed to tap the fun associated which obtained the transaction database, is an important task of data mining research field. With the kept capture and storage of large amount of data, mining association rules from the database plays more and more important role, the Apriori algorithm of mining association rules is the most classic one in database mining algorithms and widely used. On the base of description of mining association rules and the Apriori algorithm, Apriori algorithm is found to have drawbacks: the rate of generating candidate item sets is low and frequently scan data, and so on. The main optimization methods of the Apriori algorithm are overviewed, and practical applications of the Apriori algorithm are pointed out, the research directions and application trends of the Apriori algorithm in the future are proposed.

Key words data Mining; association rules; apriori algorithm; review