广西科学院学报
Journal of Guangxi Academy of Sciences
2006,22(2):125～128
Vol. 22,No. 2 May 2006

数据挖掘中的增量式关联规则更新算法*

An Efficient Incremental Updating Algorithm in Data Mining for Maintaining Association Rules

蒙 韧[1],苏毅娟[2],朱晓峰[3],张继连[3]
MENG Ren[1],SU Yi-Juan[2],ZHU Xiao-feng[3],ZHANG Ji-lian[3]

(1.广西师范大学财务处,广西桂林 541004;2.广西师范学院数学与计算机科学系,广西南宁 530001;3.广西师范大学数学与计算机学院,广西桂林 541004)
(1. Finance Department, Guangxi Normal University, Guilin, Guangxi, 541004, China; 2. Department of Mathematics and Computer Science, Guangxi Techers College, Nanning, Guangxi, 530001, China; 3. College of Mathematics and Computer Science, Guangxi Normal University, Guilin, Guangxi, 541004, China)

摘要:设计增量关联规则更新算法,用于解决数据挖掘中元组数增加而最小支持度不发生变化时关联规则增量式更新问题。该算法只须扫描原始数据库和新增数据库各一遍,能大大降低运算时间,加快速度,极大地提高关联规则的挖掘性能。

关键词:数据挖掘 关联规则 增量更新算法
中图法分类号:TP331 **文献标识码**:A **文章编号**:1002-7378(2006)02-0125-04

Abstract: Updating association rules is an inevitable yet important issue in data mining. This paper presents a highly efficient updating algorithm, referred to AIUA algorithm, for incrementally maintaining association rules with the same minimum support. This algorithm only takes one-scan on both the original database and the increased dataset. We experimentally evaluated our approach, and demonstrated the efficiency and promising.

Key words: data mining, association rule, incremental updating algorithm

数据挖掘,又名数据库中的知识发现,是从大量的数据中发现有效的、新颖的、潜在有用的知识[1]。在交易数据项目之间挖掘关联规则的问题是Agrawal在文献[2]中首先引入的。例如,"90% 的客户在购买面包和黄油的同时也会购买牛奶"是一条关联规则。其直观的意义是,客户在购买面包和黄油的时候有可能会购买牛奶。找出所有类似这样的规则,对于确定市场策略是很有价值的。关联规则的其他应用还包括附加邮递、目录设计、追加销售、仓储规划以及基于购买模式对客户进行划分等等,这些应用中的数据库都是极其庞大的。因此,不仅需要设计高效的算法来开采关联规则,而且也迫切需要设计高效的算法来更新、维护和管理已开采出来的关联规则。目前对更新算法方面学者们进行了大量的研究工作。Cheung[3]提出的FUP算法,首先考虑了关联规则的高效更新问题,他们考虑的问题是给定事物数据库DB,假定最小支持度不变,当一个新的事物数据集db添加到DB中去时,如何生成DB∪db中的关联规则。随后,Zhang等提出了增量式维护算法[4],冯玉才、冯剑琳提出了IUA算法[5],朱玉全提出了NEWFUP算法[6]和FUFIA算法[7]。

事实上,当一个新的事物数据库db添加到DB中去,对于任一项目集可能有4种情况:①在DB中频繁,db中频繁,则在DB∪db中频繁;②在DB中频繁,db中不频繁,则在DB∪db中不确定;③在DB中不频繁,db中频繁,在DB∪db中不确定;④

收稿日期:2005-09-21
修回日期:2006-02-16
作者简介:蒙 韧(1973-),男,广西玉林人,研究生,主要从事财务数据库研究。
*广西教育厅科学研究项目:区间值数据库中知识发现。

在DB中不频繁，db中不频繁，在DB∪db中不频繁。

其中①和④两种情况下项目集的性质不变，因此处理这两种情况比较简单。FUP算法有效的处理了情况①、②、④，但是对③却无能为力。IUA、NEWFUP算法分别使用了自己的方法解决了情况③，但是扫描DB的次数过多，无形中增加了算法的时间复杂度。本文主要研究的是最小支持度不变而数据库容量增加时关联规则更新问题，提出一种高效的增量式更新算法AIUA（advanced incremental updating algorithm）。这种算法把上述4种情况细化成6类，采用类似递归的方法有效解决了情况③。

1 增量关联规则更新算法(AIUA)

1.1 算法主要思想

增量关联规则更新问题的形式化描述如下：设原事务数据库为DB，其包含的事务总条数为$|DB|$，S_0为用户给定的最小支持度，L为数据库DB的频繁项集集合，db是增加的事物数据库，其所包含的事务总条数为$|db|$，L'为db最小支持度$S_{d-\min}$的频繁项集集合；DB∪db为DB添加了db后的数据库，L''为DB∪db的频繁项集集合。$X.SupD$，$X.Supd$，$X.SupUD$分别表示项集X在DB、db及DB∪db中的支持度，$t=|DB|/|db|$。解决数据库增加数据时的关联规则更新问题最简单的方法是直接把剩下的数据库DB∪db挖掘一遍。但显然这从时间和空间来说都不划算的，并且我们希望在挖更新的事物数据库时要用到原来DB所产生的频繁集。

定义　对于项集$X\subseteq L$，如果有$X.count\ L\geqslant S_0\times|DB|$，且$X.count\ L\geqslant S_0\times(|DB|+|db|)$成立，则称$X$为DB中的强频繁项集。同样定义db中的强频繁项集。

定理1　若X为L中的强频繁项集，则X必为DB∪db中的频繁项集。

定理显然可以由定义得证。

定理2　设$S_{D-\max}=\{L$中项集的支持度的最大值，且为L中非强项频繁集$\}$，若要使L中的非强项频繁集为DB∪db中的频繁集，则在db中的这个项集的支持度的最小值$S_{d-\min}\geqslant S_0-t\times(S_{D-\max}-S_0)$。

证明　因为要使$(S_{D-\min}\times|DB|+S_{d-\max}\times|db|)/(|DB|+|db|)\geqslant S_0$，

必有$S_{d-\max}\times|db|\geqslant|DB|\times S_0+|db|\times S_0-|DB|\times S_{D-\max}$，

所以$S_{d-\min}\geqslant S_0-t\times(S_{D-\max}-S_0)$。

定理3　设$S_{d-\max}=\{L'_d$中非强项频繁集支持度的最大值，且L'_d中所有项集的支持度$\geqslant S_0\}$，若要使L'_d中的项集为DB∪db中的频繁集，则在L_D中的这个项集的支持度的最小值$S_{D-\min}\geqslant S_0-1/t\times(S_{d-\max}-S_0)$。

证明　因为要使$(S_{D-\min}\times|DB|+S_{d-\max}\times|db|)/(|DB|+|db|)\geqslant S_0$，

必有$S_{d-\max}\times|db|\geqslant|DB|\times S_0+|db|\times S_0-|DB|\times S_{D-\min}$，

所以$S_{d-\min}\geqslant S_0-t\times(S_{D-\max}-S_0)$。

定理4　项集X要在DB∪db中频繁，则它必须在DB或db中频繁。

证明　用反证法。

假设$X.SupD<S_0$且$X.Supd<S_0$，则

$(X.SupD\times|DB|+X.SupdX\times|db|)/(|DB|+|db|)<(S_0\times|DB|+S_0\times|db|)/(|DB|+|db|)<S_0$，

所以，命题成立。

有了以上定义和定理，我们的算法可以，第一步：先在L中找出强项频繁集，这些强项频繁集由定理1知属于L''，同时经过扫描计算出L中除去强项频繁集的项集中的$S_{D-\max}$。第二步：计算出$S_{d-\min}$，以为最小支持度扫描db，得到db中的强项频繁集（属于L''）和L'（L'最小支持为$S_{d-\min}$），由定理1保证在这个支持度以下的项集不可能和L中的相同项集累加为DB∪db中的频繁集。第三步：逐个扫描L中的每个项集，对L中的每个项集扫描L'集，计算出DB∪db中的频繁集，把这样的项集累加到L''；同时，令$L_F=L-L''$，由定理2知，L中剩下的项集绝对不可能成为DB∪db中的频繁集。目前的进程，就是前言中的情况①和②，L''还剩下的部分在L_D中。第四步：令$L'_d=L'-\{L'$中的强项频繁集$\}-\{L'$中的DB∪db中的频繁集$\}-\{L'$中的最小支持度$S_{d-\min}\leqslant s<S_0\}$，由定理4知，要使$X\subseteq L''$，必须$X\subseteq L'_d$。第五步：扫描$L'_d$，通过公式$S_{D-\min}=S_0-1/t\times(S_{d-\max}-S_0)$计算出$S_{D-\min}$，一般情况下db远远小于DB，这使得$S_{D-\min}$非常接近于$S_0$，这使扫描时产生的频繁集不多并且时间复杂度也大大减少。第六步：在DB中可以用各种产生频繁项集的方法产生L_D：{最小支持度为$S_{D-\min}$}. 由定理3知，L''中的剩余部分只能在这找到，这里，我们使用apriori算法。第七步：逐个扫描L'_d中的每个项集，对L'_d中的每个项集扫描L_D，计算出DB∪db中的频繁集，这样

的项集一定属于 L'';此时,L'' 的产生过程结束。第八步:输出 L''。

1.2 AIUA 算法描述

算法中一些相关符号:$X.count\ L$、$x.count\ db$、$X.count\ L_F$、$x.count\ L'$、$X.count\ L_D$ 和 $x.count\ L'_d$ 分别表示项集 X 在 L、db、L_F、L'、L_D 及 L'_d 中的支持数。

输入:DB:原事务数据库
　　L:DB 的频繁项集集合
　　db:从 DB 中减去的事务数据库
　　S_0:最小支持度
输出:L'':DB∪db 的频繁项集集合

(1) L''=φ
　for each X∈L do begin
　　if X.count L≥S_0×(|DB|+|db|) then
　　L''=L''+{X};
　　L_F=L−{X};
　　Else S_{D-max}= max{x_i.suppD|x_i∈L}
　end
//搜索 L,找出其中的强项频繁集,求出 S_{D-max}
(2) S_{d-min}=S_0−t×(S_{D-max}−S_0);
　　for each x∈db do begin
　　if x.count db ≥S_0×(|DB|+|db|) then
　　L''=L''+{x};
　　db=db−{x};
　　Else L'={x_I∈db|x.count db≥S_{d-min}×(|DB|+|db|)}
　　end
//扫描 db,找出其中的强项频繁集,求出 L'
(3) for each X∈L_F do begin
　　for each x∈L' do begin
　　if X.count L_F+X.count L'≥S_0×(|DB|+|db|) then
　　L''=L''+{X}
　　L'=L'−{X}
　　else　L_F=L_F−{X}
　　end
　end
//在 L_F 和 L'中找出 DB∪db 的频繁项集
(4) L'_d=L'−{x|S_{d-min}≤x.Supd≤S_0}
(5) if L'_d=φ then output L''
　　else for each x∈L'_d　do　begin
　　S_{d-max}=max {x_i.suppd|x_i∈L}
　　S_{D-min}=S_0−1/t×(S_{d-max}−S_0)
　　end
　end
//如果 L'_d=φ 说明 L'_d 中的支持度<S_0,即程序结束,否则继续,求出 S_{D-min}
(6) L_D=aproiri-gen(L_k,S_{D-min})
//用 aproiri 算法求 L_D
(7) for each x∈L'_d do begin
　　for each X∈L_D do begin
　　if X.count L_D+X.count L'_d≥S_0×(|DB|+|db|) then
　　L''=L''+{X}
　　else delete X
　　end
　end
//在 L'_d 和 L_D 中产生 DB∪db 中的频繁项集
(8) output L''

2　实验结果与分析

在 P 350 和 Windows NT4.0 环境下对 AIUA 算法进行实验。算法用 VC++6.0 实现,测试数据库的有关参数采用与文献[2]相同的记号:$|D|$ 表示交易数据记录的数目,$|dk|$ 表示新增数据库,$|T|$ 表示交易数据记录的平均长度,$|I|$ 表示最大的潜在频繁项目集的平均长度,$|L|$ 表示最大的潜在频繁项目集的数目,N 表示交易项目的个数,采用文献[8]的数据生成程序生成实验所需要的合成数据集,取 $|D|=1105$,$|dk|=500$,$|T|=10$,$|I|=4$,$|L|=2000$,$N=1000$,minsupp=20%。由图 1 可知,IUAR 算法极大地提高了关联规则的挖掘性能。

AIUA 算法高效的关键在于如何减少对数据库的扫描次数和生成较小的候选项目集,与重新运行 Apriori 算法相比,AIUA 算法在以下几方面提高了关联规则的发现效率。

(1)文献[3]中增加新数据 db 后新的候选频繁项目集 L 的生成,是由扫描 db 直接得到的,但在多次增加 db 时每次扫描所得到的 L' 之间含有重复的频繁项目,AIUA 算法则是通过理论计算并且经过了证明得出最小的支持度去扫描 db,使得产生的频繁项集不会很大而且没有冗余。

(2)文献[5]采用类似 Apriori 算法扫描 DB 多次,文献[6]虽然只扫描了 DB 一次,但是由于使用了后备支持度(小于最小支持度),产生了很多无用的后选频繁集,AIUA 算法仅扫描 DB 一次,并且通过计算和证明保证了扫描的支持度与最小支持度尽

可能的接近，这样产生的频繁集无效的不多，量也不是很大。

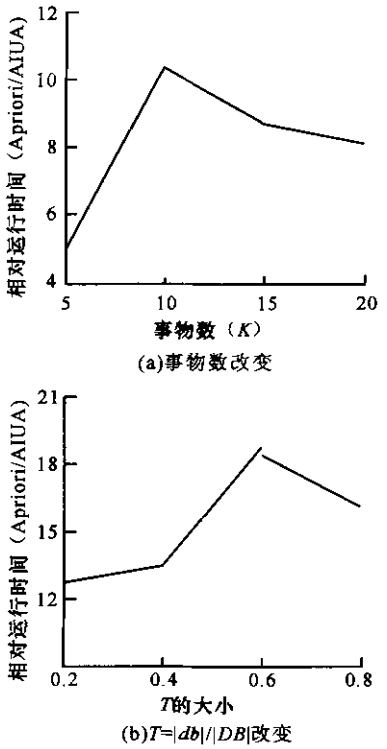

图1 IUAR 算法与 Apriori 算法的效率比较

(3)文献[6]扫描了多次 db，时间的花销和存储空间上的花销必会很大，AIUA 算法通过理论的支持只扫描一遍 db，有效了节省了时间和空间。

(4)文献[7]虽然在对 DB 和 db 扫描的次数上有优势，但是它的前提是 DB 中的频繁项集数目较少和 db 远远小于 DB，但事实上，往往有时数据库产生的频繁项集非常多，甚至可以超过事物数的个数，而 db 的大小也经常与 DB 差不多，AIUA 算法在这两个方面都有优势。

3 结束语

本文提出的 AIUA 算法只扫描了 DB 和 db 各一次，大大的降低运算时间，加快了速度。用 AIUA 算法对数据库进行增量式的关联规则更新，效率比 Apriori 算法高。但是，由于有时数据库中产生的频繁项集的数量有可能很大，甚至有时比原数据库量还大，因此扫描 DB 中频繁项集的花销较大，这是我们即将要进行研究的问题。

参考文献：

[1] HAN JIAWEI, MICHELINE KAMBER. Data mining: Concepts and techniques[M]. Beijing: Higher Education Press, 2001.

[2] AGRAWAL R. Mining association rules between sets of item s in large databases[C]//Proceedings of ACM S IGMOD Conference on Management of Data, Washington, DC, 1993.

[3] CHEUNG DAVID W. Maintenance of discovered association rules in large databases: an incremental updating technique [C]//Proceedings of the 12th International Conference on Data Engineering. New Orleans, Louisiana, 1996: 106-114.

[4] ZHANG SHICHAO, ZHANG CHENGQI, YAN XIAOWEI. Postmining: maintenance of association rules by weighting[J]. Information Systems, 2003, 28(7): 691-707.

[5] 冯玉才，冯剑琳. 关联规则的增量式更新算法[J]. 软件学报，1998，9(4)：301-306.

[6] 朱玉全，汪晓刚. 一种新的关联规则增量式更新算法[J]. 计算机工程，2002，28(7)：106-107.

[7] 朱玉全，孙志挥，赵传申. 快速更新新频繁项集[J]. 计算机研究与发展，2003，40(1)：94-100.

[8] IBM. Computer science: Intelligent information systems [EB/OL]. [2005-12-15]. http://www.almaden.ibm.com/cs/quest/data/assoc.gen.tar.Z.

（责任编辑：邓大玉）

减轻拔牙痛苦的新生物技术

美国佛罗里达州立大学的科研人员研究出一种减轻拔牙痛苦的新生物技术。他们利用一种叫做“耻骨松弛激素”的人类荷尔蒙来实现减轻拔牙过程的痛苦。耻骨松弛激素是由女性卵巢分泌的雌性荷尔蒙，能帮助生产时放松子宫颈并且放松骨盆韧带。女性在怀孕末期，耻骨松弛激素能够放松联合耻骨，这块骨头是骨盆前端的接合部，这样一来能够在分娩时给宝宝以更大的空间通过骨盆。而佛罗里达州立大学的研究人员正是利用了这一点，把它注射入牙龈或者牙齿根部，使胶原质和弹性蛋白的弹性回复能力被弱化，牙齿于是变得容易拔出。目前，耻骨松弛激素作为一种药物已经获得了美国食品和药物管理局的许可，相信不久就能够帮助牙病患者从拔牙的痛苦中摆脱出来。

（据《科学时报》）