

基于规则模式 LRU 优化的交互式关联规则挖掘

张亚鹏, 刘 燕

(赤峰学院 计算机科学与技术系, 内蒙古 赤峰 024000)

摘 要: 将规则模式应用到交互式关联规则挖掘算法中,能有效提高用户的参与程度,然而如果规则模式数目较多,需要花费大量精力处理规则模式及其模式级别。通过 LRU 算法优化规则模式,提出 IAFBLRU 算法。在该算法中,用户可主动提高感兴趣的规则模式的级别;也可采用 LRU 策略自动降低用户不感兴趣规则的模式级别。实验证明 IAFBLRU 算法有效提高了关联规则挖掘算法的质量。

关键词: 规则模式;规则模式级别;交互式;LRU 策略

中图分类号: TP312

文献标识码: A

文章编号: 1000-7180(2011)08-0231-03

Interactive Association Rule Mining Based on Rule Schema LRU Optimization

ZHANG Ya-peng, LIU Yan

(Department of Computer Science and Technology, Chifeng University, Chifeng 024000, China)

Abstract: Applying rule schemas to interactive association rule mining can improve the participation of users effectively; however if the schemas are larger, much effort has to be spend on rule schemas and their levels. LRU is introduced to optimize rule schemas, and IAFBLRU is proposed in this paper, in which users can proactively improve high interesting rule schemas' level; and also LRU policy can reduce the uninterested rule schemas' level automatically. Experiments show that the algorithm improves the quality of association rule mining algorithm effectively.

Key words: rule schema; rule schema level; interactive framework; LRU

1 引言

交互式关联规则挖掘是数据挖掘的热点^[1],国内学者在交互式关联规则挖掘算法上做了很多研究,很大程度上改进了挖掘效率和效果^[2]。

文献[3]引入规则模式和模式级别较好地表达了用户对领域的认知和规则的主观兴趣度,结果较好地符合用户要求。文献[4]中提出一种新的方法,让用户仅仅专注于自己感兴趣的子集,通过这种方式,使算法产生的规则具有较高的兴趣度。

本文通过规则模式^[5]表达用户对领域的认知程度,提出 IAFBLRU (iterative Apriori Frame Based on LRU) 算法。算法中,用户可主动提高感兴趣规则的模式级别,降低不感兴趣的模式级别;也可根据用

户的交互自动修改规则模式级别。

2 规则模式级别的 LRU 优化

用户可通过定义规则模式,表达用户在该领域已经具备的知识。而规则模式级别则体现了用户对于此种类型的规则的主观兴趣程度。

LRU 算法是为提高页面访问命中率,进行页面置换的一种算法,广泛应用到现代操作系统中。本文引入 LRU 算法动态调节规则模式的级别。

2.1 LRU 在 IAFBLRU 算法中应用

LRU 算法的基本思想是自动淘汰很没有使用的页面。这种思想可以应用在 IAFBLRU 挖掘算法中。如果用户在很长时间没有对某规则模式进行调节,则可认为用户对该规则模式的兴趣度在降低,应

自动降低该模式的兴趣度级别。

定义 1 USED: 用户在迭代过程中, 修改了规则模式的级别, 那该模式称之为 USED。

定义 2 M-TIMES: 用户定义的固定值, 描述用户从上一次改变某规则模式级别到现在算法所经历的最大迭代次数。

定义 3 L-WEIGHT: 规则模式级别权重, 从高到底包含四个权重等级 L-Max、L-3、L-2、L-1, 分别对应 VI、CI、LI、NI 规则模式级别权重。

算法在自动处理规则模式级别时, 需要考虑特定的情况, 假设用户定义规则模式 R1 的级别为 VI, 表明用户对此规则模式的兴趣度极高, 不应该主动降低该规则模式级别。

本文在文献[5]的基础上为规则模式设定权重, 表 1 所示为对 All Rules 规则模式级别设定的权重。

表 1 所有规则模式级别权重

级别	简称	L-WEIGHT
All Rules-VI	AllR-VI	L-Max
All Rules-CI	AllR-CI	L-3
All Rules-LI	AllR-LI	L-2
All Rules-NI	AllR-NI	L-1

挖掘过程中, 对每个规则模式都要定义规则模式级别, 也就定义了相应的规则模式级别权重。迭代过程中判断是否有模式的级别需自动调整。

2.2 规则模式级别的 LRU 优化过程

图 1 中描述了本文提出的结合 LRU 方式自动处理规则模式级别的过程。

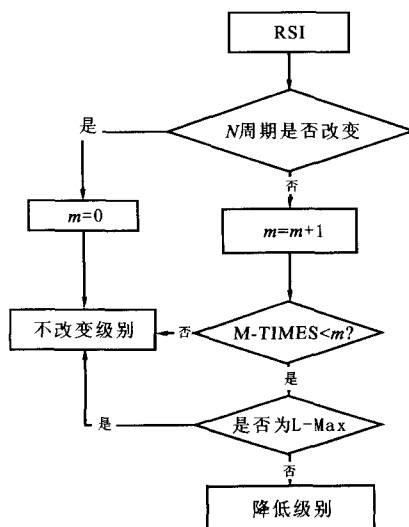


图 1 LRU 的处理过程

假设 IAFBLRU 算法进行到第 N 次的迭代, 对于存在的规则模式 RSI:

(1) 用户是否主动的改变了规则模式的级别, 如果是, 则不自动调整模式级别, 并将记录 m 清零, 否

则记录 m 加 1, 进行第二步判断;

(2) 判断 m 是否大于 M-TIMES, 如果是进入第三步判断, 否则不改变其规则模式的级别;

(3) 判断模式级别权重是否为 L-Max, 如果是则不修改规则模式级别, 否则降低模式级别。

3 基于规则模式 LRU 优化的交互式关联规则挖掘

面对大量的挖掘结果, 用户可能只会注意到特别感兴趣和特别不感兴趣的规则。兴趣度一般的规则, 既无法有效的去除, 又干扰用户的选择。在用户主导的关联规则挖掘中通过用户交互使挖掘结果更符合用户需求^[6]。IAFBLRU 能有效处理并适时自动调整模式级别, 减少不感兴趣模式产生规则的数目, 优化挖掘结果。IAFBLRU 交互式算法过程见图 2。

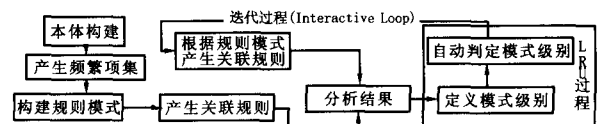


图 2 IAFBLRU 交互式算法过程

本文提出的交互式挖掘算法分为四个部分:

(1) 构建领域本体, 通过本体和用户定义的规则模式指导关联规则挖掘得到最初的结果;

(2) 通过分析和研究最初结果, 根据用户的兴趣程度定义规则模式级别;

(3) 动态修正规则模式的支持度、置信度, 通过 LRU 算法自动调整长时间没修改模式级别。

(4) 执行(2)和(3), 直到结果满足用户需求。

IAFBLRU 算法描述如下:

输入: 数据集 D ; 最小支持度 m_sup ; 最小置信度 m_conf ; 本体 O ; 空闲周期数 M

输出: 所有用户感兴趣规则的集合 R ;

```

[0] 构建规则模式集合  $S$ ;
[1] 初始化所有的规则模式级别为: AllR-CI;
[2] 频繁项集  $F = DS\text{-}Apriori\text{-}Gen()$ 
[3] for ( $F$  中每一个频繁项集  $f$ )
[4] for (由  $f$  产生的每一个规则  $r$ )
[5]   if ( $r$  Match  $S$  中的某个规则  $s$ )
[6]     判断  $s$  级别  $level(VI, CI, LI, NI)$ 
[7]      $f(r, sup \geq level * s, sup \&\& r, sup \geq m\_sup \&\&$ 
[8]        $r, conf \geq level * s, conf \&\& r, conf \geq m\_conf)$ 
[9]      $R = R \cup r$ ;
[10] if ( $R$  不满足用户需求)
[11]   主动修改规则模式, 及其级别;
[12]   LRU 处理规则模式;
[13] goto [3]
[14] else Return ( $R$ );
  
```

4 实验及分析

实验采用 Georgia Tech 大学^[7]提供的近 5 000 人互联网调查数据集,本文抽取数据集的 23 个属性进行试验.实验环境为 2 GB 内存,Inter(R) CPU P7350 2.0 GHz, Windows XP 系统.

首先设置最低最小支持度和置信度为 0.2;最小支持度和置信度为 0.2;M-TIMES=3;VI-Match=0.8;CI-Match=1.0;LI-Match=1.2;NI-Match= ∞ . 还需提供初始的规则模式,如表 2 所示.

表 2 规则模式表

ID	decription	Level
RS1	GI < Major Geographical Location, Race, Primary Language >	AllR-CI
RS2	GI < Race, Registered to vote >	AllR-CI
RS3	GI < Sector, Primary Language >	AllR-CI
RS4	RPC < Marital Status—> Number of children in Household >	AllR-CI
RS5	RPC < Education Attainment, Number of children in Household—> Household Income, Sector >	AllR-CI

由图 3(a)可以看出,IAFBLRU 算法在关闭 LRU 优化时,最初的挖掘的结果多达 1 244 条,繁多的规则使得用户很难从中得到对自己有用的部分,而且其中还包含了很多常识性的规则.当修改模式 RS1 的规则模式级别为 ConfmR-LI 后,从图 3(a)可知,关联规则的挖掘的数目降低到了 1 227 条,去掉了部分的已知而其无意义的规则.再将 RS1 的规则模式级别调整为 ConfmR-NI,挖掘结果的数目又减少到了 1 198 条.以后的迭代过程中,进一步降低 RS2,RS3,RS4,RS5 的规则模式级别,挖掘规则的数量进一步减少,规则的质量在逐步提高.在第十二步迭代之前,如果用户希望得到更多和 RS5 相关的未知结果的规则,则修改 RS5 的规则模式级别为:unexpConseqR—VI,由图 3(a)可以看出,挖掘的关联规则的结果数目有所提升,增加了更多有 RS5 导出的未知结果的规则.在第七步迭代之

前,修改 RS5 的规则模式级别为:AllR-VI,挖掘出来了更多关于 RS5 的规则.

第十二步迭代后,逐步提高 RS1,RS2,RS3,RS4 的规则模式级别,直至所有的规则模式的模式级别都为 AllR-VI,关联规则的挖掘数目在第十六步迭代后,达到了最高点 1 595. 随后进一步降低规则模式级别,规则数目逐步下降.

开启 LRU 优化策略后,通过对长时间没有处理的规则模式的级别进行自动降低,达到了优化挖掘结果的目的,如图 3(b)所示.在第四次迭代中,自动降低了规则模式的级别,因此第四次到第八次挖掘的规则数目,相对于没有开启 LRU 的算法都有所降低,起到了优化挖掘结果的目的.

由以上实验和分析可知,通过定义规则模式,可以有有效的表达用户具备的先验知识,通过对于规则模式级别的调整(用户调整和 LRU 调整),可以更有效的限制规则的产生方式,使关联规则挖掘的结果更加符合用户的需求,对现实具有更高的指导意义.

5 结束语

本文通过引入规则模式表示用户已有的知识,使用户可充分地参与到挖掘过程中;通过定义规则模式级别,表现用户对规则的主观兴趣度,使得挖掘的结果更加符合用户需求;随后本文提出交互式算法 IAFBLRU 采用 LRU 策略对规则模式级别优化处理,进一步优化挖掘结果.通过实验可以看出 IAFBLRU 算法挖掘效果比较理想.

参考文献:

- [1] Wache H, gele T V, Visser U, et al. Ontology based integration of information — a survey of existing approaches. [C]// Stuckenschmidt, H., ed., IJCAI — 01 Workshop, [s. l.]: Ontologies and Information sharing, 2001.
- [2] Lenca, P, Meyer, P, Vaillant B et al. On selecting interestingness measures for association rules; User oriented description and multiple criteria decision aid[J]European Journal of Operational Research, 2008, 184(2): 610 — 626.
- [3] Xia Shixiong, Wang Zhiyuan, Zhang Lei. An interactive algorithm research on ontology — based association rules mining[C]//ICCE2010. Chengdu, China: IEEE, 2010.
- [4] Blanchard J, Guillet F, Briand H. Interactive visual

(下转第 236 页)

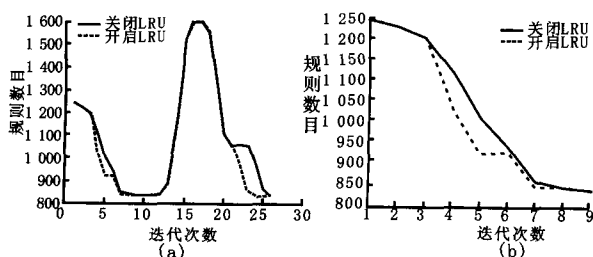


图 3 开启和关闭 LRU 对于算法的影响

机制,只需要重新选择其他节点完成下载任务.

3.2 RS 纠删码在云存储中的优化

在对文件进行编码的过程中由于用户待上传文件的大小是不一定的,在上传的文件过大的情况下,需要读入内存的数据过多将导致编码效率的下降.因此,设置缓冲区分段读入文件数据是有必要的.然而如何在文件大小不断变化的情况下合理设置缓冲区大小,使得算法在编码任何大小的文件时都能保持最高效率,是缓冲区设置算法需要解决的问题.

在设计缓冲区自适应算法时,首先对两个不同大小的文件进行多组测试,测试结果如图 3 所示.

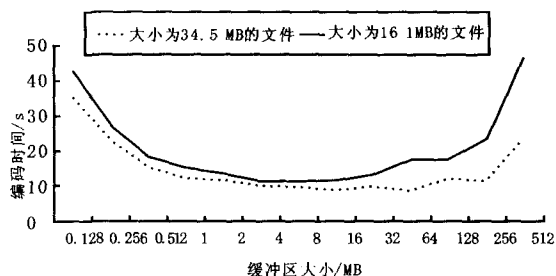


图 3 不同缓冲区大小设置下编码时间对比

易见在缓冲区大小为 2~16 MB 的时候,编码效率最高,因此设置缓冲区初值为 2 MB. 纠删码在编码过程中,涉及到的参数有文件读写单位 W 和文件分片数量 k ,同时要求每个进行计算的数据分片的大小能够被上述两个参数的乘积整除,因此缓冲区的设置需要考虑到这点.综上所述,缓冲区自适应算法的描述如下:

输入: k 值和 w 值;

输出: 缓冲区大小 buffersize 值;

步骤 1: 初始化 buffersize, 设值其初值为 2MB;

步骤 2: 如果 buffersize 不能被 $w * k$ 整除, 设置两个信号变量 up 和 down, 初值大小与 buffersize 相等; 步骤 3: 信号变量 up 和 down 同时分别自增和自减, 当其中一个能被 $w * k$ 整除时, 停止当前操作, 设置 buffersize 为该值.

应用该算法后, n 个分片中数据内容不是连续的, 而是来自不同的数据段, 这使得在数据恢复时缓冲区大小 (buffersize) 的值是必须获得的. 这就使得在无法获得编码缓冲区大小算法时, 数据恢复无法进行, 在一定程度上也保证了用户数据的私密性.

4 结束语

本文将 RS 纠删码引入云存储中, 在客户端采用基于 RS 纠删码的分片聚合机制对文件进行编码处理, 弱化了云端存储节点的不可靠性对数据安全性的影响, 保证了用户数据的隐私性、可靠性和完整性. 同时这种方法也存在着不足, 算法中没有对云端存储节点信息进行及时的检查和更新, 一旦大量存储节点信息丢失或遭篡改, 用户数据将无法恢复.

参考文献:

- [1] Hafner J L. WEAVER codes: highly fault tolerant erasure codes for storage systems[C]// Proceedings of the 4th Usenix Conference on File and Storage Technologies (FAST'05), San Francisco, CA: IEEE, 2005.
- [2] 张焕国, 刘玉珍. 密码学引论[M]. 武汉: 武汉大学出版社, 2004.
- [3] Bowers K D, Juels A, Oprea A. HAIL: A high-availability and integrity layer for cloud storage[C]// Proceedings of the 16th ACM conference on Computer and communications security (CCS'09), New York, NY: IEEE, 2009.
- [4] Wang C, Wang Q, Ren K, et al. Ensuring data storage security in cloud computing[C]// Proceeding of the 17th International Workshop on (IWQoS), Charleston, sc: IEEE, 2009: 1-9.

作者简介:

余林琛 女, (1977-), 讲师. 研究方向为网络安全、对等计算.

(上接第 233 页)

exploration of association rules with rule-focusing methodology[J]. KAIS, 13(1): 43 - 75.

- [5] Liu B, Hsu W, Wang K, et al. Aided exploration of interesting association rules[C]// proceedings of the 3rd Pacific-Asia Conference on Methodologies for knowledge Discovery and Data Mining. London, UK: springer-verlay, 1999.
- [6] Claudia Marinica, Fabrice Guillet. Knowledge-based

interactive postmining of association rules using ontologies [J]. IEEE Transactions on Knowledge and Data Engineering, 2010: 784-797.

作者简介:

张亚鹏 男, (1970-), 硕士研究生, 副教授. 研究方向为数据库、数据挖掘.

刘燕 女, (1963-), 硕士研究生, 教授. 研究方向为数据挖掘、智能教学软件.