# 基于候选项集个数上阶的增量式关联规则更新算法

王云岚, 李增智, 屈科文
(西安交通大学计算机系统结构与网络研究所, 陕西西安 710049)

摘 要: 提出了一种有效的增量式关联规则挖掘算法 IAR, 算法的特点在于: 提出并采用了基于候选项集个数上阶的选择扫描数据库的机制, 可有效减少数据库的扫描次数; 算法是一种通用的增量式算法, 提出了最小支持度和数据库均改变时, 增量式挖掘中的重要性质, 从而可充分利用上一次挖掘的结果, 有效减少候选项集的数目. 并且提出了基于组合数学和项集等价类理论的计算候选项集个数的上阶的方法. 通过大量的数据实验, 表明算法的效率比已有的算法有了很大提高.

关键词: 关联规则; 数据挖掘; 知识发现; 频繁项集; 增量式挖掘
中图分类号: TP311113 文献标识码: A 文章编号: 037222112 (2004) 0520731204

# A General Incremental Algorithm for Mining Association Rules

WANG Yun2lan, LI Zeng2zhi, QU Ke2wen
(Institute of Computer Architecture and Networks, Xi. an Jiaotong University, Xi. an, Shaanxi 710049, China)

Abstract: Mining of association rules is one of the most important fields in data mining. In this paper, a new general incremen2tal algorithm IAR for mining association rules is presented. The disting uishing feature of IAR is as follows: First, the selective scan strategy is adopted, which is based on the upper bound on the number of candidate itemsets; Second, IAR can efficiently update the discovered rules when the value of support threshold is changed, new transactions are added to the database, and obsolete ones are re2moved from it. Furthermore, based on the Kruskal2Katona theorem and the itemset equivalence class theory, we devise the technique for calculating the maximal number of candidate itemsets. Experiments show the prominent perfomance of the algorithm IAR.

Key words: association rules; data mining; knowledge discovery; frequenct itemsets; incremental mining

## 1 引言

关联规则是KDD的重要研究领域, 而增量式关联规则挖掘算法的研究有着尤为重要的意义. Apriori[1] 算法是一种最有影响的挖掘关联规则的算法, Apriori 使用一种称作逐层搜索的迭代方法, 并且利用频繁项集的反单调性用于压缩搜索空间. DHP[2] 算法采用 hash2技术和数据库中事务的缩减技术来提高 Apriori 算法的效率. TreeProjection[3] 算法通过逐渐建立字典顺序的项集树来产生频繁项集. FP2树[4] 算法将数据库压缩到一棵频繁模式树, 并在频繁模式树的基础上挖掘频繁项集. 但上述算法均为非增量式的算法, 当数据库更新时或最小支持度改变时, 需要重新挖掘整个数据库. FUP2[5] 研究当有新的事物增加到数据库和旧的事务被淘汰时的挖掘, 但是 FUP2 的缺点是每一次迭代过程中都需要扫描数据库, 且不适用于最小支持度改变时的挖掘; SWF[6] 算法采用了减少扫描数据库次数的策略, 但它的缺点是会造成候选项集个数急剧增长, 在最小支持度较小时算法效率很低, 而且也仅适用于最小支持度不变时的挖掘; NEWIUA[7] 算法仅仅研究了最小支持度改变时的增量式挖掘, 没有考虑数据库更新的情况.

本文提出了一种高效通用的增量式关联规则挖掘算法 IAR, 其主要特点及创新点为: (1) 研究了通用的增量式关联规则挖掘中的性质, 提出了新的定理和推论, 从而使 IAR 可用于数据库和最小支持度均改变, 仅有数据库更新, 或仅仅最小支持度改变等各种情况下的挖掘. 并通过扫描数据库新增部分和淘汰部分, 对候选项集进行有效剪枝. (2) 提出了一种基于候选项集个数上阶的数据库选择扫描策略, 可有效减少数据库的扫描次数, 且候选项集的个数不会增长很快, 从而使数据挖掘的总体时间开销大大减少. (3) 根据候选项集等价类理论和组合数学中的 Kruskal2Katona 定理, 提出了计算下一次迭代中候选项集的个数的方法. 并通过数据实验, 证明算法的效率比已有算法有了较大提高.

## 2 增量式关联规则挖掘算法 IAR

本文采用了以下符号: 原数据库为 D, 数据库新增部分为

收稿日期: 2003203217; 修回日期: 2003207230
基金项目: 国家自然科学基金(No. 60173066)

$d^+$, 数据库淘汰部分为 $d^-$, 更新后的数据库为 DB, 则数据库保持不变的部分为 D- $d^-$; sc、s 分别表示前一次挖掘和本次挖掘时的最小支持度; $X_{sup}$, $X_{sup\_D}$, $X_{sup+}$, $X_{sup-}$ 分别表示项集 X 在 DB、D、$d^+$ 和 $d^-$ 中的支持计数, 显然 $X_{sup}= X_{sup\_D}+ X_{sup+} - X_{sup-}$; | DB| , | D| , | $d^+$| , | $d^-$ | 分别表示 DB、D、$d^+$ 和 $d^-$ 中的事物数, 显然, | DB| = | D| + | $d^+$| - | $d^-$ | ; $Lc_k$、$L_k$ 分别表示前一次挖掘和本次挖掘中的频繁 k 项集, $C_k$ 为数据库更新后的候选 k 项集. 另外, 有关项集, 频繁项集, 支持计数, 最小支持度等基本概念的定义同文献[8], 本文不再赘述.

当数据库和最小支持度均改变时, 增量式关联规则挖掘中有如下重要性质, 其中定理 1 和推论 1 与文献[5] 中的结果类似, 定理 2、3 和推论 2 为本文的研究结果.

定理 1 P X I $Lc_k$, 若 $X_{sup\_D}+ X_{sup+} - X_{sup-} < s^*$ | DB| , 则 X| $L_k$.

证明 根据频繁项集的定义, 立即可证.

定理 2 P X| $Lc_k$, 若 $X_{sup+} - X_{sup-}$ F ( s- sc)| D| + s* ( | $d^+$| - | $d^-$ | ), 则 X| $L_k$.

证明 用反证法. 假设 X I $L_k$, 则 $X_{sup}= X_{sup\_D}+ X_{sup+} - X_{sup-}$ E s* ( | D| + | $d^+$| - | $d^-$ | ), 那么, $X_{sup\_D}$ E s* ( | D| + | $d^+$| - | $d^-$ | )- ( $X_{sup+} - X_{sup-}$ ) E sc* | D| , 根据频繁项集的定义, X I $Lc_k$, 矛盾, 因此假设不成立. 定理得证.

推论 1 P X I $Lc_k$, 若 $X_{sup\_D}+ X_{sup+} < s^*$ | DB| , 则 X| $L_k$.

推论 2 P X| $Lc_k$, 若 $X_{sup+}$ F ( s- sc)| D| + s* ( | $d^+$| - | $d^-$ | ), 则 X| $L_k$.

定理 3 若| $d^+$| = | $d^-$ | = 0, s> sc, 则 $L_i$= {X| X I $Lc_i$ C $X_{sup\_D}$ E s* | D| }, 其中 1F i F m, m= max{j| $Lc_j$ X <}.

当数据库没有更新, 仅仅最小支持度改变且 s> sc 时, 无需扫描数据库, 按照定理 3 即可计算出新的频繁项集. 以下仅讨论其它情况时的挖掘.

定义 1 (候选项集个数上阶) 令 $P_k= C_k$ H $Lc_k$, $Q_k= C_k- P_k$. 即 $P_k$( $Q_k$) 在原数据库中为(非)频繁项集. 若v $n_1$ I N, 满足| $C_k$| F $n_1$, 则称 $n_1$ 为| $C_k$| 的上阶. 若v $n_2$ I N, 满足| $Q_k$| F $n_2$, 则称 $n_2$ 为| $Q_k$| 的上阶.

IAR 也是一种逐次迭代产生频繁项集的过程. 初始化 $C_1$ = I( I 为所有项组成的集合), $k_0$= 1(变量 $k_0$ 表示从第几次迭代没有扫描 D- $d^-$ ). 第 k 次迭代的过程是:

(1) 对于 X I $C_k$, 赋初值 $X_{sup+} = X_{sup-} = 0$; $P_k= C_k$ H $L_k$, $Q_k= C_k- P_k$.

(2) 若| $d^+$| > 0, 扫描 $d^+$, 得到{$X_{sup+}$ | X I $P_k$ G $Q_k$}, 根据推论 1, 对 $P_k$ 进行修剪, $P_k= P_k-$ {X| $X_{sup\_D}+ X_{sup+} < s^*$ | DB| }; 若( s- sc) * | D| + s* ( | $d^+$| - | $d^-$ | )> 0, 根据推论 2, 对 $Q_k$ 进行修剪, $Q_k= Q_k-$ {X| $X_{sup+}$ F ( s- sc)| D| + s* ( | $d^+$| - | $d^-$ | )}.

(3) 若 0< | $d^-$ | < | D- $d^-$ | , 扫描 $d^-$, 得到 {$X_{sup-}$ | X I $P_k$G $Q_k$}, 根据定理 1, 对 $P_k$ 进行修剪, $P_k= P_k-$ {X| $X_{sup\_D}+ X_{sup+} - X_{sup-}$ F s* ( | DB| )}; 根据定理 2, 对 $Q_k$ 进行修剪, $Q_k= Q_k-$ {X| $X_{sup+} - X_{sup-}$ F ( s- sc)| D| + s* ( | $d^+$| - | $d^-$ | )}.

(4) 若| $Q_k$| > 0, 则采用基于候选项集数上阶的选择扫描策略, 首先计算| $Q_{k+1}$| 的上阶| $bQ_{k+1}$| , 若| $bQ_{k+1}$| F | $Q_k$| , 则不扫描 D- $d^-$, $C_{k+1}$= Apriori_gen($P_k$ G $Q_k$, s). 否则, 扫描 D- $d^-$, 得到{$X_{sup\_D-d^-}$ | X I G $\bigcup_{i=k_0}^{k} Q_k$}, 则 $L_i= P_i$ G {X| X I $Q_i$C $X_{sup+} + X_{sup\_D-d^-}$ E s* | DB| }, 其中 $k_0$F i F k, $C_{k+1}$= Apriori_gen( $L_k$, s), $k_0$= k+ 1.

(5) 若 $C_{k+1}$ 不为空, 令 k= k+ 1, 转式(1).

(6) 若 k E $k_0$, 扫描 D- $d^-$. 确定频繁项集 G $\bigcup_{i=k_0}^{k} L_i$.

对于 D- $d^-$, IAR 采用选择扫描的策略, 在| $bQ_{k+1}$| F | $Q_k$| 时, 不扫描 D- $d^-$, 可减少数据库扫描次数, 而在| $bQ_{k+1}$| > | $Q_k$| 时, 扫描 D- $d^-$, 可防止由于候选项集很快增长引起算法性能下降, 从而使算法具有好的性能. 下面阐述计算候选项集个数上阶的方法.

## 3 计算候选项集个数上阶的理论与方法

### 3.1 项集等价类理论和 Kruskal-Katona 定理对候选项集数的预测

为了计算候选项集个数上阶, 首先提出项集等价类概念并研究其性质.

在类 Apriori 算法中, 项集的各项之间保持着一定的偏序关系, 这里采用通常的字典序.

定义 2 (关系 $R_{k,m}$) $S_k$ 为 k 项集的集合, 对于项集 A, B I $S_k$, 当 1F mF k 时, 若对于 1F iF m, A[ i] = B[ i] ( A[ i] 表示项集 A 的第 i 项), 称 A 与 B 具有关系 $R_{k,m}$, 记作 A $R_{k,m}$ B. 当 m= 0 时, P A, B I $S_k$, A $R_{k,m}$ B.

显然, 关系 $R_{k,m}$ 为等价关系。因为它满足对称性、自反性和传递性。

定义 3 ($R_{k,m}$ 等价类) 若 $S_k$ 为 k2项集的集合, A I $S_k$, A 的 $R_{k,m}$ 等价类为 $[A]_{k,m}$= {X I $S_k$| X $R_{k,m}$ A}. 显然有 $[A]_{k,m}$ A $[A]_{k,m-1}$.

性质 1 k 项集的集合 $S_k$ 上的 $R_{k,m}$ 等价类形成 $S_k$ 的一个分划.

由等价关系的性质, 立即可证上述性质成立.

定义 4 (项集的大小) 设 $S_k$ 为 k 项集的集合, A, B I $S_k$, 若 A[ 1] > B[ 1] , 或v d, 1< d F k, A[ d] > B[ d] , 且当 1F i< d 时, A[ i] = B[ i] , 则称 A> B 或 B< A.

定理 4 设 k 项集的集合 $S_k$ 有 c 个 $R_{k,k-1}$ 等价类, 第 i 个 $R_{k,k-1}$ 等价类表示为 $[A_i]_{k,k-1}$, | $[A_i]_{k,k-1}$| = $n_i$, $m_i$ = min{| $[X_j]_{k,k-2}$ | , j = 1, 2, , , k - 1}, 其中 $[X_j]_{k,k-2}$ = $\left\{ X \text{ I } S_k \middle| \begin{array}{l} X> A_i \text{ C } X \text{ I } [A_i]_{k,k-1} \\ X[p]= A_i[p], 1\text{F } p< j \\ X[p]= A_i[p+1], j \text{ F } p \text{F } k-2 \end{array} \right\}$. 若由 $S_k$ 产生候选 ( k+ 1) 项集, 则当 k E 2 时, | $C_{k+1}$| F $\sum_{1\text{F } i\text{F } c} \min\left\{ \binom{n_i}{2}, m_i \right\}$;

当 k= 1 时, $|C_{k+1}| \le \sum_{i=1}^{c}\binom{n_i}{2}$.

证明　设 $C_{k+1}(i)=\{X \in C_{k+1} \mid X[p]=A_i[p], 1 \le p \le k-1\}$

(1) 首先证明 $|C_{k+1}(i)| \le \binom{n_i}{2}$

$X \in C_{k+1}(i)$, X 的前(k- 1)个项为 $G_{p=1}^{k-1}\{A_i[p]\}$, 由类 Apriori 算法中候选项集产生过程可知, X 的后 2 个项在集合 $\{Y[k] \mid Y \in [A_i]_{k,k-1}\}$ 中, 该集合有 $n_i$ 个项, 而从 $n_i$ 个项中取 2 项有 $\binom{n_i}{2}$ 中取法, 因此 $|C_{k+1}(i)| \le \binom{n_i}{2}$.

(2) 接着证明, 当 k≥2 时, $|C_{k+1}(i)| \le m_i$

$X \in C_{k+1}(i)$, $j \in \{1,2,\ldots,k-1\}$, 令 $X^j=X-\{X[j]\}$.

因为候选项集的子集必须是频繁项集, 故 $X^j \in S_k$. 且容易证明, $X^j > A_i$, $X^j \mid [A_i]_{k,k-1}$; 当 $1 \le p < j$ 时, $X^j[p]=A_i[p]$; 当 $j \le p \le k-2$ 时, $X^j[p]=A_i[p+1]$. 从而 $X^j \in [X_j]_{k,k-2}$. 故 $|C_{k+1}(i)| \le \min\{|[X_j]_{k,k-2}|, j=1,2,\ldots,k-1\} = m_i$.

(3) 易证 $\forall i \le c$, 满足 $C_{k+1}(i) \subseteq C_{k+1}$; 且 $\forall 1 \le i < j \le c$, $C_{k+1}(i) \cap C_{k+1}(j)=\varnothing$. 故 $|C_{k+1}|=\sum_{i=1}^{c}|C_{k+1}(i)|$

综合(1), (2), (3), 可得到定理的结论.

根据 Kruskal2Katona 定理, 可以证明以下定理[9].

定理 5　若 $S_k$ 为 k2项集的集合, $|S_k|=\binom{n_k}{k}+\binom{n_{k-1}}{k-1}+\cdots+\binom{n_r}{r}$ 为 $|S_k|$ 的 k2规范表达式, 若候选(k+ p)项集 $C_{k+p}$ 由 $S_k$ 产生, 那么 $|C_{k+p}| \le \binom{n_k}{k+p}+\binom{n_{k-1}}{k-1+p}+\cdots+\binom{n_s}{s+p}$, 其中 s 为满足 $n_s \ge s+p$ 的最小整数.

312　候选项集上阶理论在增量式关联规则挖掘中的应用

定理 6　若 $C_{k+1}$ 由 $P_k \cup Q_k$ 产生, 则 $P_{k+1}=\{X \mid X \in Lc_{k+1}, P(k\text{ 项集}) Y < X, \forall Y \in P_k\}$.

证明　(1) $\forall X \in P_{k+1}$, 因 $P_{k+1}=C_{k+1} \cap Lc_{k+1}$, 故 $X \in Lc_{k+1}$, 且 $X \in C_{k+1}$. $P(k\text{ 项集}) Y < X$, 因 $C_{k+1}$ 由 $P_k \cup Q_k$ 产生, 则 $\forall Y \in (P_k \cup Q_k) \cap Lc_k \subseteq P_k$, 因而 $\forall Y \in P_k$.

(2)若 $X \in Lc_{k+1}$, 且 $\forall Y < X(Y$ 为k 项集), $Y \in P_k$, 那么必有 $X \in C_{k+1}$, 从而 $X \in C_{k+1} \cap Lc_{k+1}=P_{k+1}$. 根据定理 4、定理 5 和定理 6, 容易得到以下定理.

定理 7　若 $|P_k \cup Q_k| \le \binom{n_k}{k}+\binom{n_{k-1}}{k-1}+\cdots+\binom{n_r}{r}$, $P_k \cup Q_k$ 有 c 个 $R_{k,k-1}$等价类, 第 i(1≤ i≤ c)个 $R_{k,k-1}$等价类表示为 $[A_i]_{k,k-1}$, $|[A_i]_{k,k-1}|=n_i$, $m_i=\min\{|[X_j]_{k,k-2}|, j=1,2,\ldots,k-1\}$, 其中 $[X_j]_{k,k-2}=\left\{X \in S_k \;\middle|\; \begin{array}{l} X > A_i \cap X \mid [A_i]_{k,k-1} \\ X[p]=A_i[p], 1 \le p < j \\ X[p]=A_i[p+1], j \le p \le k-2 \end{array}\right\}$. 若 $C_{k+1}$ 由 $P_k \cup Q_k$ 产生, 那么 $|Q_{k+1}| \le \min\left\{\sum_{i=s}^{k}\binom{n_i}{i+1}, \sum_{i=1}^{c}\min\left\{\binom{|[A_i]_{k,k-1}|}{2}, m_i\right\}\right\} - |P_{k+1}|$, 其中 s 为满足 $n_s \ge s+1$ 的最小整数, $P_{k+1}=\{X \mid X \in Lc_{k+1}, P(k\text{ 项集}) Y < X, \forall Y \in P_k\}$.

根据定理 7, 在算法 IAR 第 k 次迭代中, 根据 $P_k$ 和 $Q_k$, 即可计算出 $|Q_{k+1}|$ 的上阶, 用于本文提出的启发式选择扫描策略.

## 4　算法性能实验

实验在 2. 0GHz, 512M 内存的计算机上进行, 数据的产生方法与文献[5]相同, 算法采用 VC+ + 6. 0 实现. 在算法运行时间、扫描数据库的次数和候选项集数等方面对 IAR、Apriori 和 FUP2 算法进行了全面比较. 以下用 Tt. Ii. Dm- x+ y 表示数据库中事务的平均长度为 t, 频繁项集的平均长度为 i, 更新前数据库中有 m 条事务, 数据库更新后有 x 条事务被淘汰, 有 y 条事务增加到数据库中.

图 1 研究数据库中增加的事务数对算法性能的影响. 数据库为 T5l I4l D 10K- 1 K+ x, $|d^+|$ 的变化范围从 500 到 4000, s= sc= 110%. 在这种情况下, IAR 的速度比 Apriori 快 2. 06 到 2. 92 倍, 比 FUP2 快 28% 到 37%.

图 2 研究数据库中减少的事务数对算法性能的影响, 数据库为 T5. I4. D10K2x+ 1K, $|d^-|$ 的变化范围从 500 到 4000, s = sc= 1. 0%. IAR 的速度比 Apriori 快 0. 71 到 4. 11 倍, 比 FUP2 快 26% 到 38%. 随着减少事务数的增加, 更新后数据库中的事务数减少, 因而 Apriori 的运行时间减少; 而 IAR 和 FUP2 的运行时间有所增加. 但即使在数据库减少数为原数据库事务数 40% 的情况下, IAR 和 FUP2 的速度依然比 Apriori 快.

图 3 研究算法 IAR 是否适合大规模数据库的挖掘, 数据库中的事务数从 5 万条到 100 万条不等, 数据增加量和减少量均为原数据库的 10%, 最小支持度为 1. 0%, 从图中可看出, 算法 IAR 比 FUP2 快 30%, 比 Apriori 快两倍, 且算法的运行时间与数据库中的事务数呈线性关系, 说明算法具有很好的扩展性.

图 4 研究最小支持度改变量即 $\Delta s= s- sc$ 对算法性能的影响, 因 FUP2 不能用于此种情况下的挖掘, 因此仅比较了 IAR 与 Apriori 算法的性能. 数据库为 T5. I4. D20K- 2K+ 2K, 上一次挖掘支持度为 sc= 210%, 本次挖掘支持度 s 从 1. 0% 变化到 4. 0%, 从图中可以看出, IAR 的速度比 Apriori 快 1. 04 到 6. 29 倍, 且随着 $\Delta s$ 增大, 算法 IAR 对于算法 Apriori 的加速比也随之提高, 也就是说在本次挖掘的最小支持度比上一次大的情况下, 算法 IAR 具有更好的性能, 原因在于若上一次挖掘最小支持度较小, 则频繁项集较多, 从而为本次挖掘存储了更多信息.

数据试验还比较了数据库扫描次数和候选项集数, 对数据库 T5. I4. D10K21K+ 2K 的挖掘中, 当最小支持度为 1. 0%, IAR 扫描数据库的次数为 5 次, 而 Apriori 扫描数据库的次数为 8 次; IAR 产生的候选项集总数为 5837, 而 Apriori 和 FUP2

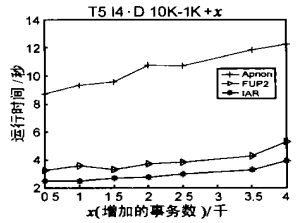


图 1 数据库中增加的事务数对运行时间的影响

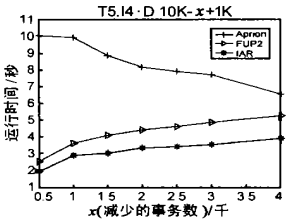


图 2 数据库中减少的事务数对运行时间的影响

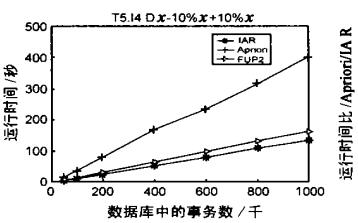


图 3 数据库大小对运行时间的影响

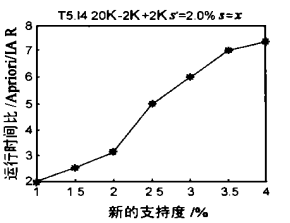


图 4 最小支持度改变量对算法性能影响

的候选项集的总数为5646. 可见, 算法 IAR 可减少数据库的扫描次数, 同时, 不会引起候选项集数的过分增多.

## 5 结束语

提出了一种通用的增量式关联规则挖掘算法 IAR, 该算法可用于数据库和最小支持度均改变时的挖掘, 也可用于仅仅数据库更新或仅仅最小支持度改变时的挖掘. 研究并提出了增量式关联规则挖掘中的重要性质, 充分利用上一次挖掘出的知识, 对候选项集进行修剪. 提出了基于组合数学和项集等价类理论的计算候选项集个数上阶的方法, 以此为基础确定了一种启发式的数据库选择扫描策略, 在保证候选项集数不会增长很快的情况下, 减少数据库扫描次数. 并通过大量数据试验说明了本文提出的算法 IAR 具有优越的性能.

参考文献:

[1] R Agrawal, T Imielinski, A Swami. Mining association rules between sets of items in large databases [A]. Proceeding of 1993 SIGMOD International Conference on Management of Data [C]. New York: ACM press, 1993. 207- 216.

[2] J S Park, M S Chen, P S Yu. Using a hash2based method with transaction trimming for mining association rules [J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9: 813- 825.

[3] R Agarwal, C Aggarwal, et al. A tree projection algorithm for generation of frequent itemsets [J]. Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining), 2001: 1- 23.

[4] Jiawei Han, Jian Pei, Yiwen Yin. Mining frequent patterns without candidate generation [A]. Proceeding of 2000 ACM SIGMOD International Conference on Management of Data [C]. New York: ACM press, 2000. 1- 12.

[5] D W Cheung, S D Lee, Benjamin Kao. A general incremental technique for maintaining discovered association rules [A]. Proceedings of the Fifth International Conference on Database Systems for Advanced Applications [C]. World Scientific Press, Singapore, 1997. 185- 194.

[6] Chang2Hung Lee, Cheng2Ru Lin, Ming2Syan Chen. Sliding window filtering: An efficient algorithm for incremental mining [A]. Proceedings of the ACM tenth International Conference on Information and Knowledge Management [C]. New York: ACM press, 2001. 263- 270.

[7] 周海岩. 关联规则的开采与更新 [J]. 软件学报, 1999, 10 (10): 1078- 1084.

[8] Han Jiawei, Micheline Kamber. 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, 2001.

[9] F Geerts, B Goethals, J Van den Bussche. A tight upper bound on the number of candidate patterns [A]. IEEE International Conference on Data Mining [C]. USA: IEEE Computer Society Press, 2001. 155- 162.

作者简介:

王云岚 女, 1970 年 11 月生于山西省垣曲县, 博士生, 主要研究领域为知识发现, 信息融合及计算机网络管理. Email: wangyunlan@263. net.

李增智 男, 1938 年 4 月生于陕西省浦城县, 教授, 博士生导师, 主要研究领域为计算机网络及应用.