

一种基于极大团的关键时间段挖掘方法

王宁^{1,2} 杨扬¹ 巩华荣² 赵耀培^{1,3} 孟坤¹

(北京科技大学计算机与通信工程学院 北京 100083)¹

(烟台工程职业技术学院机电工程系 烟台 264006)² (山东省工会管理干部学院科研处 济南 250100)³

摘要 针对带有时间属性的海量事务处理问题,提出了一种求最大相关性的最小时间区间(关键时间段 KTI)的算法。通过利用极大团把海量的数据项进行有效的划分,降低了后续数据挖掘和决策选择的复杂度。针对特定的含有时间参量的极大团,通过寻找关键时间段(KTI),提高了决策的准确度,同时可以减小分析数据的规模,降低对计算资源的需求。假设事务中各项出现的事件具有相同的概率分布,得到了一种寻找关键时间段(KTI)的算法。从理论上证明了算法的正确性,并对其进行了复杂度分析,通过实际数据验证了算法的可行性。

关键词 数据挖掘,关联规则,时序逻辑,极大团,关键时间段(KTI),概率

中图法分类号 TP311

文献标识码 A

Mining Method of Key Time Interval Based on Maximum Clique

WANG Ning^{1,2} YANG Yang¹ GONG Hua-rong² ZHAO Yao-pei^{1,3} MENG Kun¹

(College of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)¹

(Department of Electrical Engineering, Yantai Engineering and Technology College, Yantai 264006, China)²

(Research Department, College of Shandong's Union Management, Jinan 250100, China)³

Abstract To efficiently process mass transactions with temporal attribute, the key time interval(KTI), the minimum time interval reflecting correlation of items was introduced. A maximum clique based KTI mining algorithm was proposed, which reduces the complexity for information mining and decision making. Assuming the probability distribution of items is uniform for the target transaction, we stated an approach to find the KTI of the transaction and analyzed the correctness and complexity of the method. Experiments show that through considering the KTI of a clique, the candidates set impacting the accuracy of decision is reduced, and the resource consumption is saved mainly. In the end, we evaluated the feasibility of the proposed method in the real world use.

Keywords Data mining, Association rule, Time series logic, Maximum clique, Key time interval(KTI), Probability

1 引言

关联规则挖掘是数据挖掘中最活跃的研究方法之一^[1,5,7]。通过对这些海量数据的智能分析,可以获得有用的知识。其主要应用领域有目录设计、商品广告邮寄分析、追加销售、仓库规划、网络故障分析、市场规则、广告策划、分类设计等^[1,3,6,8,10]。

海量数据的处理对计算资源有较高的要求,尤其是内存资源往往面临极大考验。因此,对原始数据的预处理可以有效降低对计算资源的消耗,其中极大团是常采用的一种技术^[2,6,8]。根据出现的频率等标准,把关联性强的项构造成极大团并确定其共同特征,从而实现原始海量的数据项的有效划分和合理抽象,为后续的计算分析提供较小数量级的数据,以有效缓解内存不足的问题^[8,9]。

然而,极大团的生成往往忽略时间因素的影响^[2,4]。通过收集和分析尽量长时间段内(时间段并)的数据,保证了极大团的准确性。但是,据我们所知,关于从生成极大团的时间特性考虑的研究目前还比较匮乏。基于此,针对生成的极大团,求不连续时间段内的极大有序频繁项目集,并定义关联性超过一定阈值(可以根据具体情况恰当选择)的最小时间段(时间段交)为该极大有序频繁项目集的关键时间段(KTI)。本文探讨了关键时间段(KTI)的搜索算法,算法可应用于商家关于备货策略的选择,引入时间因素可以使得备货策略更具针对性和时效性。

本文在假设相同事务中数据项出现的事件具有相同概率分布的情况下,研究了关键时间段(KTI)的搜索算法。主要贡献包括以下几个方面:(1)给出了极大团关键时间段(KTI)的形式化定义;(2)得到固定时间段内极大团出现的概率组

到稿日期:2011-07-15 返修日期:2011-10-17 本文受国家自然科学基金(60873192,61070182),北京市教育委员会科技发展计划重点项目(KZ201010009008),山东省高等学校优秀青年教师国内访问学者项目经费资助。

王宁(1978—),女,博士生,讲师,主要研究方向为数据挖掘、网络服务、云计算, E-mail: wangning200658@126.com; 杨扬(1955—),男,教授,博士生导师,主要研究方向为图像处理与模式识别、智能控制、多媒体通信; 巩华荣(1969—),女,副教授,主要研究方向为计算机应用; 赵耀培(1974—),男,博士生,副教授,主要研究方向为服务组合和优化; 孟坤(1980—),男,博士生,主要研究方向为网络安全、性能评价。

合;(3)在得到的概率组合的基础上设计了时间复杂度为 $O(n * |L_{max T}|)$ 的搜索算法;(4)针对真实数据验证了算法的有效性。

本文第2节对基本概念和理论基础进行了阐述;第3节给出了一种搜索算法并进行了复杂度分析;第4节对给出的算法进行了实际数据验证和分析;最后总结全文,并指出了未来研究方向。

2 基本概念和理论基础

2.1 基本概念

为了叙述方便,这里仅考虑项目集为有序的情形。事实上,对于可数的无序项目集,存在多项式算法使得该集合中的元素有序排列。下面给出相关术语和基本概念。

设 $Item = \{I_1, I_2, \dots, I_m\}$ 是项的集合。设任务相关的数据 D 是数据库事务的集合,其中每个事务 M 是项的集合,使得 $M \subseteq Item$ 。每一个事务有一个标识符,称作 Tid 。

定义1(时间区间变量操作)^[6] 设 $I_1 = [I_1^-, I_1^+]$ 和 $I_2 = [I_2^-, I_2^+]$ 是定义在某个时态解释域上的两个时态区间变量,它们的时态交($\cap_T$)、时态并($\cup_T$)运算定义如下:

(1)时态交($\cap_T$):如果 $I_1^+ < I_2^-$ 或 $I_2^+ < I_1^-$, 则 $I_1 \cap_T I_2 = \phi$; 否则 $I_1 \cap_T I_2 = [I_3^-, I_3^+]$ 。其中 $I_3^- = \max\{I_1^-, I_2^-\}$, $I_3^+ = \min\{I_1^+, I_2^+\}$ 。

(2)时态并($\cup_T$):如果 $I_1^+ < I_2^-$ 或 $I_2^+ < I_1^-$, 则 $I_1 \cup_T I_2 = \phi$; 否则 $I_1 \cup_T I_2 = [I_3^-, I_3^+]$ 。其中 $I_3^- = \min\{I_1^-, I_2^-\}$, $I_3^+ = \max\{I_1^+, I_2^+\}$ 。

定义2(基于时序的事务模式 M) 带有时间约束的事务模式 M 表示为 $\{Tid, Item, Time\}$ 。其中 Tid 为事务标识号,全局唯一; $Item = \{I_1, I_2, \dots, I_m\}$ 为项集,其中 $I_n (1 \leq n \leq m)$ 称为该项集的项,项集中项的个数被称为事务模式 M 的长度; $Time$ 为有效时间,常用时间区间表示。

定义3(基于时序的事务模式 M 的频繁项目集) 已知基于时序的事务模式 M 的 n 条事务的集合为 $\{M_1, M_2, \dots, M_n\}$, 其中 $M_i = \{Tid_i, Item_i, Time_i\}$, $Item_i$ 是事务 M_i 中 m 个项的集合,即 $Item_i = \{I_{i1}, I_{i2}, \dots, I_{im}\}$ 。如果 k 个事务的时间并区间 $\cup_T = \bigcup_{i=1}^k Time_i$, 且 $\cup_T \neq \phi$, $k \in \{1, 2, \dots, n\}$ 。如果在 $\cup_T$ 内,项集 FI 的出现频率大于或等于最小支持度,那么该项集 FI 就是 $\cup_T$ 中的频繁项目集。

定义4(基于时序的事务模式 M 的有序频繁项目集) 设项集 $FI = \{I_1, I_2, \dots, I_n\}$ 为时间并 $\cup_T$ 中的频繁项目集,如果 FI 为有序项集,则称 FI 为有序频繁项目集。

定义5(基于时序的事务模式 M 的极大有序频繁项目集) 已知基于时序的事务模式 M 的集合为 $\{M_1, M_2, \dots, M_n\}$, 其中 $M_i = \{Tid_i, Item_i, Time_i\}$, $Item_i$ 是事务 M_i 中 m 个项的集合,即 $Item_i = \{I_{i1}, I_{i2}, \dots, I_{im}\}$ 。设 $FI = \{I_{i1}, I_{i2}, \dots, I_{ij}\}$ 为 $\cup_T$ 中的有序频繁项目集,其中 $1 \leq j \leq m$, 且 $FI \subseteq Item_i$, $\cup_T$ 是 k 个事务的时间并区间,即 $\cup_T = \bigcup_{i=1}^k Time_i$, 且 $\cup_T \neq \phi$ 。①对于 $\forall I_k \notin FI$, $FI \cup I_k$ 均为非频繁项目集,则称 FI 为极大频繁项目集;②由于 FI 为有序频繁项目集,且满足 FI 为极大频

繁项目集,因此称 FI 为极大有序频繁项目集。

2.2 关键时间段(KTI)及其判别准则

由于本文最感兴趣的是在时间交中至少包含该频繁项目集的概率情况,因此首先定义基于时间交的关键时间段(KTI)。

定义6(关键时间段 KTI) 已知基于时序的事务模式 M 的集合 $\{M_1, M_2, \dots, M_n\}$ 中的 m 个事务,其中 $M_i = \{Tid_i, Item_i, Time_i\}$, $1 \leq m \leq n$, 而关键时间段(KTI)必须同时满足以下3个条件:

(1)这 m 个事务在时间上的交集不为空,即 $\cap_T = \bigcap_{i=1}^m Time_i \neq \phi$;

(2)这 m 个事务在时间并 $\cup_T$ 中对构成一个或多个极大有序频繁项目集都有贡献;

(3)在时间交 $\cap_T$ 中,至少包含该极大有序频繁项目集的概率和,用 P 来表示;而概率和的阈值即最小概率和用 θ_i 来表示,那么必须满足 $P \geq \theta_i$ 。

这样同时符合以上3个条件的的时间交 $\cap_T$ 就称为关键时间段(KTI)。

阈值 θ_i 的确定一般应满足以下原则:(1)阈值应根据实际分析的问题而定,不同的应用往往对应各异的阈值范围;(2)阈值的确定受处理能力的限制,若具有较多的能力或资源处理分析的目标问题,则选择较小的阈值,反之则选取较大的阈值;(3)阈值的确定应尽量避免主观因素的影响,可采用模糊决策和信息熵方法等;(4)阈值确定应综合考虑多方面因素影响,收集尽量多的信息,并采用多次计算的方法,降低选取值的误差。

阈值的确定方法多数采用决策选择的相关理论,形成了一系列较为有效的方法,如专家系统方法、统计方法及6西格玛准则方法等,并形成了另外专门的学科,其超出了本文讨论的范围,在此不再赘述。

2.3 时间区间内极大频繁项目集概率计算

关键时间段(KTI)的判断依赖于特定频繁项目集中所有元素出现的概率和。我们对此展开叙述,并在假设各项目的出现相互独立,且每个事务中所有项在时间交 $\cap_T$ 内出现的事件满足相同概率分布的条件下,给出频繁项目集中所有元素出现概率和的显式表达式,为关键时间段(KTI)的判断提供理论支撑。分析各异概率分布的情况,将是我们以后研究的一个方向。

这里考虑的分布可以是任意类型的分布,因此给出的结论具有较好的可扩展性,可以根据不同的应用选取相应的分布类型。直观地,在没有任何先知信息的条件下,可以选择所有项在其时间区间内服从均匀分布。

定理1 已知基于时序的事务模式 M 的集合 $\{M_1, M_2, \dots, M_n\}$ 中的 $r (1 \leq r \leq n)$ 个事务集合用 $\{M_1, M_2, \dots, M_r\}$ 表示,其中 $M_i = \{Tid_i, Item_i, Time_i\}$, 且所有事务相互独立。若用 $Item_{ij}$ 表示事务 M_i 中第 j 项在 $\cap_T$ 中出现的事件,令 $|Item_i| = L_i$, 满足 $\Pr(Item_{ij}) = \Pr(Item_{ij'})$, 其中 $j \leq L_i, j' \leq L_i, j \neq j', i \leq r$, 即任意 $Item_i$ 中项出现的事件相互独立且服从相同的概率分布,表示为 P_i , 这里 $i = 1, 2, \dots, r$; 用 P 表示在 $\cap_T$

内所有 M_i 中包含 k_i 个固定的项出现事件的概率和的乘积, 那么

$$P = P_1^{k_1} P_2^{k_2} \dots P_r^{k_r} (1 + P_1)^{L_1 - k_1} (1 + P_2)^{L_2 - k_2} \dots (1 + P_r)^{L_r - k_r}$$

证明: 已知基于时序的事务模式 M 的集合 $\{M_1, M_2, \dots, M_n\}$ 中的 $r (1 \leq r \leq n)$ 个事务集合用 $\{M_1, M_2, \dots, M_r\}$ 表示, 假设 M_1 在时间交 $\cap_T$ 内必须至少出现 $A_{i+1} A_{i+2} \dots A_{i+k_1}$ 这 k_1 个项, 其中 $1 \leq k_1 \leq L_1$ 。由于任意 Item_i 中项出现的事件相互独立且服从相同的概率分布 $P_i, 1 \leq i \leq r$, 因此在时间交 $\cap_T$ 内出现包含该 k_1 个项的事件的概率和为

$$P_1^{k_1} + C_{L_1 - k_1}^1 P_1^{k_1 + 1} + C_{L_1 - k_1}^2 P_1^{k_1 + 2} + \dots + C_{L_1 - k_1}^{L_1 - k_1 - 1} P_1^{k_1 + (L_1 - k_1 - 1)} + P_1^{L_1}$$

同理, 假设 M_2 在时间交 $\cap_T$ 内必须至少出现 $B_{j+1} B_{j+2} \dots B_{j+k_2}$ 这 k_2 个项, 其中 $1 \leq k_2 \leq L_2$, 那么在时间交 $\cap_T$ 内出现包含该 k_2 个项的事件的概率和为

$$P_2^{k_2} + C_{L_2 - k_2}^1 P_2^{k_2 + 1} + C_{L_2 - k_2}^2 P_2^{k_2 + 2} + \dots + C_{L_2 - k_2}^{L_2 - k_2 - 1} P_2^{k_2 + (L_2 - k_2 - 1)} + P_2^{L_2}$$

假设 M_r 在时间交 $\cap_T$ 内必须至少出现 $R_{t+1} R_{t+2} \dots R_{t+k_r}$ 这 k_r 个项, 其中 $1 \leq k_r \leq L_r$, 那么在时间交 $\cap_T$ 内出现包含该 k_r 个项的事件的概率和为

$$P_r^{k_r} + C_{L_r - k_r}^1 P_r^{k_r + 1} + C_{L_r - k_r}^2 P_r^{k_r + 2} + \dots + C_{L_r - k_r}^{L_r - k_r - 1} P_r^{k_r + (L_r - k_r - 1)} + P_r^{L_r}$$

因为 $M_1, M_2, \dots, M_r$ 之间也是相互独立的, 所以

$$\begin{aligned} P &= (P_1^{k_1} + C_{L_1 - k_1}^1 P_1^{k_1 + 1} + C_{L_1 - k_1}^2 P_1^{k_1 + 2} + \dots + C_{L_1 - k_1}^{L_1 - k_1 - 1} P_1^{k_1 + (L_1 - k_1 - 1)} + P_1^{L_1}) \times \\ &\quad (P_2^{k_2} + C_{L_2 - k_2}^1 P_2^{k_2 + 1} + C_{L_2 - k_2}^2 P_2^{k_2 + 2} + \dots + C_{L_2 - k_2}^{L_2 - k_2 - 1} P_2^{k_2 + (L_2 - k_2 - 1)} + P_2^{L_2}) \times \dots \times \\ &\quad (P_r^{k_r} + C_{L_r - k_r}^1 P_r^{k_r + 1} + C_{L_r - k_r}^2 P_r^{k_r + 2} + \dots + C_{L_r - k_r}^{L_r - k_r - 1} P_r^{k_r + (L_r - k_r - 1)} + P_r^{L_r}) \\ &= P_1^{k_1} (1 + C_{L_1 - k_1}^1 P_1 + C_{L_1 - k_1}^2 P_1^2 + \dots + C_{L_1 - k_1}^{L_1 - k_1 - 1} P_1^{L_1 - k_1 - 1} + P_1^{L_1}) \times \\ &\quad P_2^{k_2} (1 + C_{L_2 - k_2}^1 P_2 + C_{L_2 - k_2}^2 P_2^2 + \dots + C_{L_2 - k_2}^{L_2 - k_2 - 1} P_2^{L_2 - k_2 - 1} + P_2^{L_2}) \times \dots \times \\ &\quad P_r^{k_r} (1 + C_{L_r - k_r}^1 P_r + C_{L_r - k_r}^2 P_r^2 + \dots + C_{L_r - k_r}^{L_r - k_r - 1} P_r^{L_r - k_r - 1} + P_r^{L_r}) \end{aligned}$$

因为 $(a+b)^n = C_n^0 a^0 b^n + C_n^1 a^1 b^{n-1} + \dots + C_n^n a^n b^0$, 所以 $P = P_1^{k_1} P_2^{k_2} \dots P_r^{k_r} (1 + P_1)^{L_1 - k_1} (1 + P_2)^{L_2 - k_2} \dots (1 + P_r)^{L_r - k_r}$ 。

3 算法设计与分析

3.1 设计思路

由定义 6 可知, 关键时间段(KTI)的确定依赖于预先设定的阈值和极大团在时间并 $\cup_T$ 内的关联程度。假设事务模式 M 的集合中所有事务相互独立, 那么在 $\cap_T$ 内所有 M_i 中包含 k_i 个固定的项出现事件的概率和的乘积可用来度量项目集元素间的关联程度。基于此, 针对事务集合 $\{M_1, M_2, \dots, M_r\}$ 中频繁项目集, 可以得到一种判断关键时间段(KTI)的算法, 其基本思路如下:

(1) 根据定义 6(关键时间段 KTI), 求 r 个事务的时间交

$\cap_T = \bigcap_{i=1}^r \text{Time}_i$, 若 $\cap_T = \phi$, 可判断所有项目集不存在关键时间段(KTI), 否则得到了该算法的研究对象 $\cap_T$;

(2) 遍历所有的极大有序频繁项目集, 根据定理 1 计算至少包含该项目集的事件概率和的乘积, 得到该项目集的关联程度;

(3) 根据定义 6(关键时间段 KTI), 比较设定的阈值可以判断该 $\cap_T$ 是否为所搜索的关键时间段(KTI)。

3.2 设计实现

依据上面的思路可以得到关键时间段(KTI)的搜索算法, 描述如下:

算法 $\text{MaxF} \cap_T()$: 在时间并 $\cup_T$ 中, 对构成极大有序频繁项目集做贡献的事务集合在时间交 $\cap_T$ 中至少出现该极大有序频繁项目集的概率和与关键时间段(KTI)。

输入: 极大有序频繁项目集 L_{maxT}

输出: 在时间并 $\cup_T$ 中, 对构成极大的有序频繁项目集做贡献的事务集合在时间交 $\cap_T$ 中至少出现该极大有序频繁项目集的概率和与关键时间段(KTI)

步骤: (1) 分别求 L_{maxT} 中的每个极大有序频繁项目集中事务个数 r 和项的个数 L ;

(2) 利用 r 个事务集合, 可以找到每个事务的时间区间, 求 r 个事务的时间交 $\cap_T = \bigcap_{i=1}^r \text{Time}_i$, 若 $\cap_T = \Phi$, 可判断该极大有序频繁项目集不存在关键时间段(KTI), 否则执行(3);

(3) for all 极大有序频繁项目集 do{ 根据定理 1 计算至少包含该项目集的事件概率和的乘积 p };

(4) if $p \geq \theta$, then $\cap_T$ 是关键时间段(KTI)。

(5) 输出 p 和相应的关键时间段 $\cap_T$ 。

3.3 复杂度分析

算法 $\text{MaxF} \cap_T$ 的复杂度分析: 分析该算法之前, 假设该数据库中所有的项目数是 n 。该算法实质上是一个二重循环, 算法的时间复杂度是 $|L_{\text{maxT}}| \times k$, 其中 $|L_{\text{maxT}}|$ 指的是在极大有序频繁项目集中项目集的总个数, $1 \leq k \leq n$, 其中 $|L_{\text{maxT}}|$ 与有序频繁 2-项集的个数成正比。 $|L_{\text{maxT}}|$ 最糟糕的情况是 $C_2^{[n/2]}$, 该值趋向于 2^n 。由于本文要求的是极大有序频繁项目集, 故在实际的求解过程中会不断地大量删除项目集中的所有子集, 因此实际的 $|L_{\text{maxT}}| \ll 2^n$ 。综上所述, 算法 3 的时间复杂度是 $O(n * |L_{\text{maxT}}|)$ 。

4 实验验证

为验证所提方法的有效性, 以教材^[6]中的测试数据为蓝本, 通过对所有数据随机生成时间区间, 构造了本文的数据集。为保证数据的时间属性的一般性, 采用多次实施有差别筛选的方法来减少误差。为叙述方便, 采用前述的符号格式, 数据的结构表示为 $(\text{Tid}, \text{Item}, \text{Time})$, Tid 是事务标识, Item 是该事务的项目集合, Time 是该事务发生的时间区间。

这里主要是验证算法的正确性, 采用如下的试验设置和实施环境。

试验设置: 15 条事务、9 个项进行测验; 最小支持度为 10%, 极大团阈值为 5, 最小置信度为 60%; 关键时间段(KTI)概率和阈值为 1.17。实施环境: CPU: Intel® Pentium®

表 1 样本数据

Tid	Item	Time	Tid	Item	Time
14	I1,I2,I5,I8	[10,75]	4	I1,I2,I4	[90,170]
8	I1,I2,I3,I5,I8	[10,200]	15	I1,I2,I3	[200,300]
3	I2,I3,I7,I9	[15,75]	6	I2,I3,I6,I9	[200,750]
1	I1,I2,I5,I6,I9	[20,105]	10	I1,I3,I6,I9	[200,750]
2	I2,I4,I8	[30,180]	7	I1,I3,I6,I7	[230,800]
9	I1,I2,I3,I9	[35,97]	11	I2,I5,I7,I9	[250,300]
5	I1,I3,I4,I6,I8	[45,135]	13	I1,I2,I4,I7	[270,900]
12	I3,I5,I7	[60,300]			

将该数据进行预处理之后,先求出频繁有序 2-项集,由于频繁有序 2-项集的个数是 9(9>5),因此要对这 9 项进行极大团的划分,并针对每个极大团求解时间并内的极大有序频繁项目集和时间交内至少出现该极大频繁项目集的概率和;对所有极大团求得的时间并内的极大有序频繁项目集和时间交内至少出现该极大频繁项目集的概率和进行有效的合并,即删除子集,得到的最终结果如表 2 所列。

表 2 满足关键时间段(KTI)概率和阈值的数据

Tid	Item	时间并	时间交	概率和
12,8	I3,I5	[10,800]	[60,200]	1.532625
13,7	I1,I7	[10,900]	[270,800]	7.725795
1,5	I1,I6	[10,900]	[45,105]	5.089506
10,7	I1,I3,I6	[10,900]	[230,750]	2.387152
10,6	I3,I6,I9	[10,800]	[200,750]	1.000000
1,9	I1,I2,I9	[10,900]	[35,97]	1.160687

从表中可以得到关键时间段(KTI)分别是[60,200]、[270,800]、[45,105]、[230,750]和[35,97]。如果把关键时间段(KTI)的思想应用到超市,那么在这些关键时间段(KTI)内,商家可根据不同的情况,更有效地指导自己科学地安排进货、库存、货架设计并更准确地准备充足的货源,以供顾客购买等。

结束语 本文考虑了带有时间属性数据的关联规则挖掘方法。首先分析了目前这方面的研究现状,发现只在时间并范围内考虑数据项的关联规则挖掘,会导致伪关联规则大量出现;然后,通过实例说明了考虑时间交范围内数据项关联规则挖掘的重要性,并首次提出了关键时间段(KTI)的定义,给出了相应的寻找极大团关键时间段(KTI)的算法;最后,针对

所有事务中的每一个项在时间交内出现的事件具有相同概率分布的情形,给出了一种度量关联程度的指标,在此基础上给出了计算该指标的方法,并设计了多项式复杂度的搜索算法。

本文只针对所有项相互独立且具有相同概率分布的情形给出了多项式搜索算法,而针对非独立或不具有相同概率分布的情形是否存在多项式的搜索算法,还是个公开问题,是未来一个有意义的研究方向。

参 考 文 献

[1] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[C]//Proceeding the ACM SIGMOD International Conference on Management of Data. Washington, 1993;207-216

[2] Stix V. Finding All Maximal Cliques in Dynamic Graphs[J]. Computational Optimization and Applications, 2004, 27:173-186

[3] Skrikumar K, Bhasker B. Efficiently Mining Maximal Frequent Sets For Discovering Association Rules[C]//Proceedings of the Seventh International Database Engineering and Applications Symposium(IDEAS'03). 2003

[4] Adibi J, Shen Wei-min. Time-invariant Sequential Association Rules; Discovering Interesting Rules in Critical Care Databases [C]//Conference'KDD-01. San Francisco, CA, August 2001

[5] Yang Gui-zhen. The Complexity of Mining Maximal Frequent Item Sets and Maximal Frequent Patterns[C]//KDD'04. Seattle Washington, USA, August 2004; 22-25

[6] 毛国军,段立娟,王实,等. 数据挖掘算法与研究[M]. 北京:清华大学出版社,2005

[7] 陈安龙,唐常杰,陶宏才,等. 基于极大团和 FP-Tree 的挖掘关联规则的改进算法[J]. 软件学报, 2004, 15(8):1198-1207

[8] 韩蒙,张炜,李建中. 一种高效的不确定图 K-极大频繁模式挖掘算法[J]. 计算机学报, 2010(8):1387-1395

[9] 陆介平,杨明,孙志挥,等. 快速挖掘全局极大频繁项目集[J]. 软件学报, 2005, 16(4):553-560

[10] Yun Zhai, Ru Yang-bing, Wu Qu, et al. Study on source of classification in imbalanced datasets based on new ensemble classifier[J]. Systems Engineering and Electronic, 2011, 33(1):196-201

[11] 雷文章,陈延兵. 陆地石油企业 HSE 应急预案建立研究[J]. 中国职业安全卫生管理体系认证, 2004, 24(4):26-29

[12] 王文俊,扬鹏,董存洋. 应急案例本体模型的研究及应用[J]. 计算机应用, 2009, 7(5):1437-1440

[13] 谭酒桥,袁哲明,柏连阳,等. 基于支持向量机回归与 k-最近邻法的组合预测用于除草剂 QSAR 建模[J]. 农药学报, 2007, 9(4):324-329

[14] 柴根象,吴月琴. 相依样本下污染性模型的最近邻估计[J]. 应用数学学报, 2006, 29(3):542-554

[15] 冯凯,徐志胜,冯春莹,等. 小城镇基础设施防灾减灾决策支持系统的研究与开发[J]. 中国安全科学学报, 2004, 14(2):74-77

(上接第 165 页)

[6] 田依林. 城市公共安全应急管理信息系统建设模型[J]. 武汉理工大学学报, 2007, 29(3):68-71

[7] 张英菊,仲秋雁,叶鑫,等. CBR 的应急案例通用表示与存储模式[J]. 计算机工程, 2009, 35(17):28-30

[8] 廖光焯,翁韬,朱舜平,等. 城市重大事故应急辅助决策支持系统研究[J]. 中国工程科学, 2005, 7(7):7-13

[9] 李仕明,刘娟娟,王博,等. 基于情景的非常规突发事件应急管理研究[J]. 电子科技大学学报, 2010, 1(1):12-14

[10] 郑晓军,王奕首,滕弘飞,等. 应急系统的开发与应用[J]. 计算机应用研究, 2006, 1(1):33-35