



DOI:10.13364/j.issn.1672-6510.20170254

不确定数据的约束频繁闭项集挖掘算法

牛浩浩, 李孝忠, 连春月

(天津科技大学计算机科学与信息工程学院, 天津 300457)

摘要: 针对当前数据挖掘方法没有考虑客观条件制约以及个人不同需求的问题, 在不确定数据的频繁闭项集挖掘算法中加入简洁性约束条件, 基于正态分布模型分别研究了在简洁反单调约束和简洁非反单调约束下, 对不确定数据库进行频繁闭项集挖掘的方法, 并给出了实例证明挖掘算法的可行性。

关键词: 不确定数据; 频繁闭项集; 简洁反单调约束; 简洁非反单调约束

中图分类号: TP311 **文献标志码:** A **文章编号:** 1672-6510(2018)04-0074-05

Mining Constrained Frequent Closed Itemsets From Uncertain Data

NIU Haohao, LI Xiaozhong, LIAN Chunyue

(College of Computer Science and Information Engineering, Tianjin University of Science & Technology,
Tianjin 300457, China)

Abstract: Because the environment of current data mining is too ideal and does not consider the constraints of objective conditions and individual needs, based on the normal distribution model, succinct constraints were added to the frequent closed itemsets mining algorithms for uncertain data, and the method of frequent closed itemsets mining for uncertain databases was studied separately under succinct anti-monotone constraints and succinct non-anti-monotone constraints. Examples were also given to prove the theoretical feasibility of the mining algorithm.

Key words: uncertain data; frequent closed itemsets; succinct anti-monotone constraints; succinct non-anti-monotone constraints

数据挖掘是从数据中获取有价值的潜在信息, 目的在于将海量的数据转换成有用的知识, 并用所得知识对未来的行为进行指引。因此, 通常也被称为数据库中的知识发现 (knowledge discovery in databases, KDD)^[1]。在数据挖掘领域, 关联规则挖掘是一种经典的挖掘方法, 旨在寻找数据库中有意义的关联, 在挖掘过程中需要找到需要的频繁项集^[2]。

在实际情况下, 很多数据的产生都带有不确定性, 导致原有的频繁项集挖掘算法无法直接应用于不确定数据中。目前, 关于不确定数据库的频繁项集挖掘已有许多研究, 如由确定数据挖掘算法 Apriori、FP-growth 发展而来的 U-Apriori, UF-growth 算法, 以及基于此的一系列改进算法。然而, 随着数据的大量增加, 挖掘所得频繁项集有过多冗余项集, 有些甚

至是毫无意义的。最大频繁项集虽然在很大程度上减少了冗余项集, 然而其并不包含项集支持度信息。而频繁闭项集^[3]很好地解决了这个问题, 频繁闭项集在不丢失所需信息的前提下, 其数量远小于频繁项集, 并包含了所有频繁项集的支持度信息^[4]。

目前, 对于不确定数据库的频繁闭项集挖掘, 主要是在不确定数据频繁项集挖掘的基础上, 对频繁项集中的子集与其直接超集的支持度进行比较, 由此得到频繁闭项集。对不确定数据的频繁项集挖掘, 其关键步骤之一在于如何确定不确定数据的支持度信息。除了上述 U-Apriori, UF-growth 等算法将期望概率作为项集支持度处理之外, 文献[5]给出了概率分布的方法来近似项集的支持度计数, 在已有成果中, 将支持度近似为正态分布来进行挖掘的结果准确度

收稿日期: 2017-09-20; 修回日期: 2018-05-03

基金项目: 国家自然科学基金资助项目(61603273); 天津市自然科学基金资助项目(16JCYBJC18500)

作者简介: 牛浩浩(1993—), 女, 河南人, 硕士研究生; 通信作者: 李孝忠, 教授, lixz@tust.edu.cn

较高^[6]. 本文在挖掘过程中也将采取此基于正态分布的方法.

然而,目前数据挖掘方法的环境过于理想,未考虑可能存在的一些针对性条件或者决策者决策需求的偏好. 因此,本文将不确定数据的支持度近似为正态分布,在其频繁闭项集挖掘过程中加入了简洁性约束条件,分别研究了在简洁反单调约束和简洁非反单调约束下,对不确定数据库进行频繁闭项集挖掘的方法.

1 相关定义

1.1 不确定数据上的概率频繁闭项集

目前,在不确定数据库中挖掘频繁项集的方法分为两大类^[7]: 基于期望支持度模型的方法和基于概率频繁模型的方法.

基于期望支持度模型^[8]的方法最早是由 Chui 等提出的,这种方法使用期望支持度来对支持度进行近似计数,在计算某项集支持度时,把数据库中该项集在每个事务中对应的概率相加,相加的结果表示该项集的支持度的估计值. 因此,基于期望支持度的频繁项集有如下定义: 如果项集 X 是频繁项集,那么 X 的期望支持度 $esup(X)$ 必须大于等于用户给定的最小阈值.

基于概率频繁模型^[9]的方法由 Bernecker 在 2009 年提出,该方法认为在不确定性数据库中,项集的出现是不确定的,因此其支持度也是不确定的,可以使用适当的标准参数分布近似不确定项集的概率支持度计数,即支持度的概率分布,如泊松分布^[5]、正态分布^[6]等.

已有不确定项集支持度的近似模型中,近似效果最好的是正态分布模型^[6],其相关定义^[10]如下:

定义 1 频繁概率: 给定一个最小支持度阈值 c_{minsup} , 项集 X 的频繁概率是指 X 的支持度大于等于 c_{minsup} 的概率^[11], 记作 $P_{freq}(X)$.

定义 2 频繁概率项集: 给定一个最小支持度 $c_{minsup} \in [0, n]$ 和置信度(概率频繁阈值) $\tau \in [0, 1]$, 如果项集 X 的频繁概率大于概率频繁阈值, 则 X 是一个概率频繁项集, 即

$$P_{freq}(X) = P(\text{support}(X) \geq c_{minsup}) \geq \tau$$

定义 3 概率支持度^[12]: 项集 X 在置信度 τ 下的概率支持度 $ProbSup(X, \tau)$ 定义为

$$ProbSup(X, \tau) =$$

$$\arg_{i \in [0, n]} \max (P(\text{support}(X) \geq i) \geq \tau) \quad (1)$$

其中: $\arg_{\max}()$ 指使得括号内取得最大值的 i 值, 从定义中可知, 概率支持度是指满足项集 X 出现 i 次以上的概率大于等于 τ 的最大的 i . 也就是, 在具体过程中, 计算出 $P(\text{support}(X) \geq c_{minsup}) \geq \tau$ 后, 要继续计算 $P(\text{support}(X) \geq c_{minsup}+1) \geq \tau$, $P(\text{support}(X) \geq c_{minsup}+2) \geq \tau$, \dots , 直至 $P(\text{support}(X) \geq c_{minsup} + n) < \tau$, 则 $c_{minsup} + n - 1$ 就是概率支持度.

定义 4 概率频繁闭项集: 项集 X 是概率频繁闭项集, 当且仅当 $P_{freq}(X) \geq \tau$, 且找不到任何项集 X 的超集 Y , 满足 $ProbSup(Y, \tau) = ProbSup(X, \tau)$ ^[12].

1.2 约束条件

现有的约束型频繁项集挖掘, 允许用户使用一组 SQL(structured query language) 约束来规范其挖掘过程. 根据此约束, 可以挖掘满足用户需求的在事务中频繁发生的项目, 此类挖掘方法可以避免挖掘无意义的频繁项集所引起的不必要计算^[13].

其中, 大多数用户定义的约束是简洁的, 如 $C1: \max(X. Price) \leq \25 , 它表示用户所感兴趣的频繁项集 X 中, 最贵的物品的价格不超过 25 美元(即所需项集中每个项目的价格都不超过 25 美元); 同样的, $C2: \min(X. Price) \leq \30 表示所需项集中, 价格最低的项不超过 30 美元. 除了购物篮项目, 一套约束也可以针对其他领域的项目、事件或对象. 如, $C3: X.Location = \text{Winnipeg}$ 表示用户寻求的频繁项集 X 的所有事件都发生在加拿大温尼伯; $C4: X.Symptom = \{\text{drythroat, sneezing}\}$ 表示 X 中的每个人都患有喉咙干燥、打喷嚏中至少一种症状. 而非简洁约束, 如, $C5: \text{sum}(X. Price) \leq \100 表示挖掘所得的 X 中所有项集的价格之和不超过 100 美元, 本文对非简洁约束暂不作讨论.

除了约束是否简洁这一特性之外, 还可以根据某些其他属性(如反单调性: 约束 C 是反单调的, 当且仅当满足 C 的项集的所有子集也满足 C) 对约束条件进行划分. 如, 根据反单调性, 简洁约束可以被划分为两类: 简洁反单调约束(succinct anti-monotone constraints, SAM), 简洁非反单调约束(succinct non-anti-monotone constraints, SUC)^[13].

上述所给简洁性约束例子中, $C1, C3$ 为 SAM; $C2, C4$ 为 SUC.

2 约束条件下的概率频繁闭项集挖掘

2.1 概率频繁闭项集及其概率支持度

在挖掘概率频繁项集时, 需要计算项集的概率支

持度. 假设项集 I 在数据库中出现的次数为 X_I , 由于事务数据库足够大(事务数为 n), 将 X_I 近似为连续的正态分布, 则项集 I 的频繁概率为

$$P_{\text{freq}}(I) \approx 1 - F(c_{\text{minsup}} - 1, \mu_I, \sigma_I^2) \quad (2)$$

其中, F 是概率密度函数的积分, 即

$$F(c_{\text{minsup}} - 1, \mu_I, \sigma_I^2) = \frac{1}{\sqrt{2\pi\sigma_I^2}} \int_{-\infty}^{c_{\text{minsup}} - 0.5} e^{-\frac{(x - \mu_I)^2}{2\sigma_I^2}} dx \quad (3)$$

而

$$\mu_I = \sum_{t_j \in D} P(I \subseteq t_j) = \sum_{j=1}^n \prod P(a \in t_j) \quad (4)$$

$$\sigma_I^2 = \text{Var}(X_I) = \sum_{t_j \in D} P(I \subseteq t_j) \cdot (1 - P(I \subseteq t_j)) \quad (5)$$

在实际情况下, 可以根据数据库计算 μ_I 和 σ_I^2 , 进而求出 I 的频繁概率 $P_{\text{freq}}(I)$, 若 $P_{\text{freq}}(I) < \tau$, 即项集 I 出现 c_{minsup} 次以上的概率小于 τ , 则认为项集 I 是不频繁的; 否则项集 I 为频繁项集.

根据定义 3, 为了得到项集的概率支持度, 在计算 $P(\text{support}(X) \geq c_{\text{minsup}}) \geq \tau$ 之后, 需要继续计算 $P(\text{support}(X) \geq c_{\text{minsup}} + 1) \geq \tau, \dots$, 直至 $P(\text{support}(X) \geq c_{\text{minsup}} + i) < \tau$, 则 $c_{\text{minsup}} + i - 1$ 就是其概率支持度^[10].

2.2 约束条件下的概率频繁闭项集

在约束条件下, 可根据约束条件将数据库中的项分为 Item^M 和 Item^O 两部分^[12]. 其中, Item^M 为强制性集合, 该集合具有强制性, 是因为它是约束条件下的必选项集合; Item^O 表示非必选项的集合.

则对于任意约束 SAM, Item^M 表示满足 SAM 约束的项集 X , Item^O 表示不满足该约束条件的项目集合, 由于其反单调性, X 不应包含 Item^O 中的项, 即 X 应由 Item^M 中的项组合而成. 相应地, 对于任意 SUC 约束, 满足该约束的频繁项集应包含 Item^M 中的至少一个中的项目, 即目标项集应由 N_1 个 $\text{Item}^M (N_1 \geq 1)$ 和 N_2 个 $\text{Item}^O (N_2 \geq 0)$ 组成.

下面以实例解释数据库处理过程. 表 1 为不确定数据组成的交易数据库, 表 2 为补充的商品价格信息. 其中, 表 1 是根据用户购买习惯及浏览记录得出的购买商品及其概率, “事务” 栏中的 “T1, T2, T3” 等表示每条成交记录, “项目” 栏表示可能购买的商品及购买的概率, 如 T1, {a, b, c, d, e} 为其可能购买的商品, 购买 a 商品的概率为 0.7, 购买 b 商品的概率为 0.8, ... 表 2 表示各商品价格, 如 a 商品

价格为 10, b 商品价格为 20 等.

表 1 交易数据库

Tab. 1 Transaction database

事务	项目
T1	a:0.7 b:0.8 c:0.8 d:1.0 e:0.2
T2	a:0.7 b:0.8 d:1.0 e:0.1 f:0.4
T3	a:0.8 c:0.5 e:0.3 f:0.4
T4	b:0.8 c:0.8 d:1.0
T5	c:0.8 d:1.0

表 2 商品价格表

Tab. 2 Price of commodity

商品	价格/美元	商品	价格/美元
a	10	d	100
b	20	e	75
c	50	f	25

对于 SAM 约束 C1: $\max(X. \text{Price}) \leq \25 , 根据约束的定义和要求, 所挖掘的频繁项集中每项的价格应该小于等于 25 美元, 即所求频繁项集为小于等于 25 美元的项的集合. 根据附加信息可得: $\text{Item}^M = \{a, b, f\}$ 和 $\text{Item}^O = \{c, d, e\}$. 则对于 SAM 约束, 挖掘所得频繁闭项集必须只包括 Item^M 中的项.

对于 SUC 约束 C2: $\min(X. \text{Price}) \leq \30 , 可以得到: $\text{Item}^M = \{a, b\}$ 和 $\text{Item}^O = \{c, d, e, f\}$. 则所需约束频繁闭项集必须包含至少一个 Item^M 中的项, 也可能还包含有一些额外的 Item^M 或 Item^O 中的项.

3 约束频繁闭项集挖掘过程

数据被分为 Item^M 和 Item^O 之后, 对于 SAM 约束, 挖掘所得项集不包含 Item^O 中的项, 则可删除原数据库中 Item^O 包含的项; 对于 SUC 约束, 则可根据 Item^M 和 Item^O 将原数据库分成两个子数据库后进行挖掘.

3.1 SAM 约束下频繁闭项集挖掘

对于 SAM 约束 C1: $\max(X. \text{Price}) \leq \25 , 可以得到 $\text{Item}^M = \{a, b, f\}$ 和 $\text{Item}^O = \{c, d, e\}$. 根据约束的定义和要求, 对数据库进行修剪处理. 处理之后的结果见表 3.

表 3 SAM 约束下的必选数据库

Tab. 3 Mandatory database of SAM

事务	项目
T1	a:0.7 b:0.8
T2	a:0.7 b:0.8 f:0.4
T3	a:0.8 f:0.4
T4	b:0.8

对修剪后的数据库进行概率频繁闭项集挖

掘. 首先进行 1-项集挖掘, 在进行挖掘之前, 先对数据库进行简单的剪枝. 由于任何一个非频繁项集的超集是非频繁的, 所以对于 1-项集, 暂时不考虑项集的概率, 只计算其出现的次数, 根据设定的 c_{\minsup} , 若其出现次数小于 c_{\minsup} , 则该项集一定不频繁, 该项集的超集也是非频繁的, 因此对项集进行剪枝, 以减少不必要的计算.

剪枝之后, 对表 3 中剩余的 1-项集分别进行频繁判断: 根据式(2)一式(5)可以计算数据库中各项集的频繁概率, 若项集 I 的频繁概率大于设定的阈值, 则为 1-频繁项集, 反之由于其单调性, 去掉不频繁的项. 对于 1-频繁项集, 为了进一步进行概率频繁闭项集判断, 可以根据式(1)求得项集的概率支持度.

然后采用相同的方法判断 2-项集、3-项集等是否是频繁项集, 并计算其概率支持度. 以例 1 为例说明具体过程.

例 1 在上述数据表中, 假设 $c_{\minsup} = 2$, $\tau = 0.3$, 判断 2-项集 $I = \{a, b\}$ 和 1-项集 $J = \{a\}$ 是否是频繁项集.

对于 2-项集 I , $\mu_I = 1.12$, $\sigma_I^2 = 0.4928$, $Pr_{\text{freq}}(I) = 1 - \Phi((c_{\minsup} - 0.5 - \mu_I)/\sigma_I) = 1 - \Phi(0.38/\sqrt{0.4928}) < 1 - \Phi(0.38/\sqrt{0.5041}) = 1 - \Phi(0.38/0.71) < 1 - \Phi(0.53) = 1 - 0.7019 < \tau$, 由于 $Pr_{\text{freq}}(I) < \tau$, 则 I 为非频繁项集, 无需再进行概率支持度的计算.

而对于 1-项集 J , $\mu_J = 2.2$, $\sigma_J^2 = 0.58$, $Pr_{\text{freq}}(J) = 1 - \Phi((c_{\minsup} - 0.5 - \mu_J)/\sigma_J) = 1 - \Phi(-0.7/\sqrt{0.58}) = \Phi(0.7/\sqrt{0.58}) > \Phi(0) = 0.5 > \tau$, 则 J 为 1-频繁项集, 再根据定义 3 计算可得到 J 的概率支持度为 3.

根据以上计算可得, 2-项集 $I = \{a, b\}$ 为非频繁项集, 1-项集 $J = \{a\}$ 为频繁项集, 其概率支持度为 3.

上述步骤中 n -项集的生成, 可以采用基于宽度优先的 Apriori 算法及其改进算法^[14]或者基于深度优先的 FP-Growth 算法及其改进算法^[15].

如在宽度优先算法中, 首先根据定义 2 挖掘 1-频繁项集, 并得到其概率支持度; 之后将 1-频繁项集链接生成 2-项集, 通过相同的计算方法确定 2-频繁项集及其概率支持度, 并根据定义 4 将 1-频繁项集 X 与 2-频繁项集中其对应的超集 Y 进行比较, 若不足定义 4, 则删去前者保留后者, 若满足, 则均保留; 再根据 2-频繁项集链接生成 3-项集..., 重复上述步骤直至挖掘结束, 可以得到所有满足约束条件的概率频繁闭项集及其概率支持度.

3.2 SUC 约束下频繁闭项集挖掘

对于 SUC 约束 $C2: \min(X. Price) \leq \30 , 按照约

束 $(X. Price) \leq \$30$ 和 $(X. Price) > \$30$, 将数据库的项分为两部分: $Item^M = \{a, b, f\}$ 和 $Item^O = \{c, d, e\}$, 则需要挖掘的频繁闭项集应该由 N_1 个 $Item^M (N_1 \geq 1)$ 和 N_2 个 $Item^O (N_2 \geq 0)$ 组成. 因此将数据库分为两个数据库: SUC 下的必选数据库(表 4)及可选数据库(表 5).

表 4 SUC 约束下的必选数据库
Tab. 4 Mandatory database of SUC

事务	项目
T1	a:0.7 b:0.8
T2	a:0.7 b:0.8 f:0.4
T3	a:0.8 f:0.4
T4	b:0.8

表 5 SUC 约束下的可选数据库
Tab. 5 Optional database of SUC

事务	项目
T1	c:0.8 d:1.0 e:0.2
T2	d:1.0 e:0.1
T3	c:0.5 e:0.3
T4	c:0.8 d:1.0
T5	c:0.8 d:1.0

首先根据剪枝条件对两个数据库分别进行剪枝, 之后根据上述方法分别得到两个数据库的 1-频繁项集及其概率支持度. 值得注意的是, 为避免数据损失, 在挖掘过程中, 并不对项集 X 及其超集 Y 进行概率支持度比较, 统一保留, 由此可以得到两个数据库中所有的频繁项集. 其中, 根据约束定义可知, 表 4 挖掘所得频繁项集为符合约束条件的频繁项集.

下一步, 由表 4 数据库挖掘所得的频繁项集 $M (M \in Item^M)$ 向表 5 数据库挖掘所得的频繁项集 $N (N \in Item^O)$ 进行扩展, 过程如例 2.

例 2 假设 $c_{\minsup} = 2$, $\tau = 0.3$, 从例 1 可知, 对于表 4 数据库, $Item^M$ 中的项集 $M = \{a\}$ 为频繁项集, 对于表 5 数据库, 计算 $N = \{c\}$ 是否为频繁项集.

对于 1-项集 N , $\mu_N = 2.9$, $\sigma_N^2 = 0.73$, $Pr_{\text{freq}}(N) = 1 - \Phi((c_{\minsup} - 0.5 - \mu_N)/\sigma_N) > \tau$.

则 N 为表 5 数据库挖掘的 1-频繁项集, 根据定义 3, 通过计算可得到 N 的概率支持度为 3.

因此, 以表 4 数据库挖掘所得的 $M = \{a\}$ 为基础, 与表 5 数据库所挖掘的频繁项集 $N = \{c\}$ 进行结合, 得到新的项集 $O = \{a, c\}$, 接下来使用概率频繁模型判断在原数据库表 1 中, 项集 O 是否频繁且是否与项集 M 的概率支持度相同, 若其不频繁则不是目标项集, 若项集 O 频繁则作为目标项集保留.

按照例 2 所述方法, 以表 4 数据库挖掘所得的频

繁项集为基础,与表5数据库挖掘所得的频繁项集进行结合,并判断是否频繁,可得到此结合方法下符合约束的频繁项集.则表4数据库挖掘的频繁项集与结合所得的频繁项集共同构成满足SUC约束的频繁项集集合.

最后,对所得到的所有符合约束的频繁项集根据定义4进行验证,若某一项集的直接超集与其概率支持度相同,则删去该项集保留其直接超集,否则二者均保留.由此,可得到所有的目标项集.

4 结 语

本文根据概率频繁模型进行数据挖掘,将项集的支持度近似为正态分布,避免了生成不确定数据库的所有可能世界,在一定程度上减少了计算量.另外,引入了约束条件,在两种约束限制下重新对数据库进行挖掘,使得挖掘结果更满足不同的个人实际需求.然而,随着不确定数据的增多,其复杂程度也随之增加,加之现实生活中对数据库挖掘的要求也多种多样,需要有更高效更准确的不确定数据挖掘方法来满足人们对数据库的挖掘需求.接下来,可以针对SUC约束下较复杂的挖掘情况进行研究,以期获得更高的效率.并把约束思想应用于其他挖掘算法中,以满足人们对数据挖掘的不同需求.

参考文献:

- [1] Hand D, Mannila H, Smyth P. 数据挖掘原理[M]. 张银奎, 廖丽, 宋俊, 等译. 北京: 机械工业出版社, 2003.
- [2] Han J W, Kamber M. 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, 2007.
- [3] Pasquier N, Bastide Y, Taouil R, et al. Discovering frequent closed itemsets for association rules//[C]Proceedings of International Conference on Database Theory. Berlin: Springer Berlin Heidelberg, 1999: 398-416.
- [4] 李学明, 杨阳, 秦东霞, 等. 基于频繁闭项集的新关联分类算法 ACCF[J]. 电子科技大学学报, 2012, 41(1): 104-109.
- [5] Wang L, Cheng R, Lee S D, et al. Accelerating probabilistic frequent itemset mining: A model-based approach [C]//Proceeding of the 19th ACM international conference on Information and knowledge management. New York: ACM, 2010: 429-438.
- [6] 沈盛霞. 频繁闭项集挖掘算法研究[D]. 合肥: 安徽大学, 2016.
- [7] 章淑云. 不确定性数据中的概率频繁项集挖掘算法的研究[D]. 上海: 复旦大学, 2013.
- [8] Chui C K, Kao B, Hung E. Mining Frequent Itemsets from Uncertain Data[M]//Advances in Knowledge Discovery and Data Mining. Berlin: Springer Berlin Heidelberg, 2007: 47-58.
- [9] Bernecker T, Kriegel H P, Renz M, et al. Probabilistic frequent itemset mining in uncertain databases[C]//Proceeding of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 119-128.
- [10] 刘慧婷, 沈盛霞, 赵鹏, 等. 不确定数据频繁闭项集挖掘算法[J]. 计算机应用, 2015, 35(10): 2911-2914.
- [11] Wang L, Cheung D W, Cheng R, et al. Efficient mining of frequent itemsets on large uncertain databases[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(12): 2170-2183.
- [12] Tang P, Peterson E A. Mining Probabilistic frequent closed itemsets in uncertain databases[C]//Proceeding of the 49th Annual Association for Computing Machinery Southeast Conference. New York: ACM, 2011: 86-91.
- [13] Cuzzocrea A, Leung C K S, MacKinnon R K. Mining constrained frequent itemsets from distributed uncertain data[J]. Future Generation Computer Systems, 2014, 37(7): 117-126.
- [14] 王伟. 关联规则中的 Apriori 算法的研究与改进[D]. 青岛: 中国海洋大学, 2012.
- [15] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[J]. ACM Sigmod Record, 2000, 29(2): 1-12.

责任编辑: 常涛