第 35 卷 第 5 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.5 Computer Engineering March 2009
·软件技术与数据库· 文章编号：1000—3428(2009)05—0061—03 文献标识码：A 中图分类号：TP311

多数据库中的负关联规则挖掘技术及发展趋势

尚世菊，董祥军，赵 龙
(山东轻工业学院信息科学与技术学院，济南 250353)

摘 要：负关联规则反映了数据项之间的互斥关系，能提供很多有用的信息，在决策支持中起重要作用，但现行的挖掘算法主要是针对单一数据库的挖掘，多数据库中负关联规则的挖掘还未引起重视。该文介绍负关联规则的研究现状、主要挖掘方法以及冗余正负关联规则的修剪方法，对多数据库中关联规则挖掘研究现状和主要技术进行论述，并展望多数据库中负关联规则挖掘的发展趋势。

关键词：负关联规则；数据挖掘；多数据库

Mining Technology and Development Tendency of Negative Association Rule in Multi-database

SHANG Shi-ju, DONG Xiang-jun, ZHAO Long
(School of Information Science and Technology, Shandong Institute of Light Industry, Jinan 250353)

【Abstract】 Negative association rules can catch mutually exclusive correlations among items and play an important role in decision-making. The current mining algorithm is mainly directed against mono-database, and mining negative association rules in multi-database do not arouse people's attention. This paper elaborates on the negative association rules of the status quo, mainly mining methods and redundant positive and negative association rules pruning methods, and then expatiates the present situation and main technology of association rules in multi-database, and developments tendency of negative association rules in multi-database is forecasted.

【Key words】 negative association rule; data mining; multi-database

数据挖掘，又称数据库中的知识发现。关联规则[1]的挖掘是数据挖掘的重要内容之一，多数研究工作都围绕 $A\Rightarrow B$ 的正关联规则形式，而对形如 $A\Rightarrow\neg B$, $\neg A\Rightarrow B$, $\neg A\Rightarrow\neg B$ 的负关联规则研究较少，但负关联规则在决策支持中同样起着非常重要的作用。随着信息产业和数据库技术的发展，很多情况下必须面对多个数据库，这就涉及多数据库挖掘问题。多数据库挖掘是解决基于全局企业分布数据状态下知识发现的决策支持问题的有效技术，现在多数据库挖掘已经成为新的研究热点。

1 负关联规则研究现状及主要技术

1.1 负关联规则的研究现状

负关联规则描述的是项集间的互斥关系，与传统关联规则不同，负关联规则研究“90%的客户在购买咖啡时不会购买茶叶”之类的问题。当决策者欲知“当某些有利因素出现时，哪些不利因素很少出现”的时候，负关联规则就变得非常重要。负关联规则挖掘就是在数据库 D 中筛选出所有满足用户指定的最小支持度 $minsupp$ 和最小置信度 $minconf$ 的负关联规则 $A\Rightarrow\neg B$(或$\neg A\Rightarrow B$, $\neg A\Rightarrow\neg B$)，其中，A 和 B 分别为频繁项集。

对于负关联规则的研究，文献[2]讨论了扩展型关联规则以及原关联规则及其若干性质，是国内较早讨论负关联规则的文章。文献[3]首次提到了 2 个项集间的相关性，项集 A 和项集 B 的 χ^2 值用于确定 A 和 B 是否相互独立，如果它们不相互独立，就用一个矩阵来确定它们之间是正相关还是负相关。文献[4]阐述了强负关联规则问题，它将特定领域知识与以前发现的正关联相结合，以分类法的形式来挖掘关联规则。Wu Xindong 等人提出一个 PR 模型[5-6]，给出了一个能够同时挖掘正关联规则和负关联规则的算法，该算法以传统的 Apriori 算法为基础来挖掘频繁项集和非频繁项集，在挖掘频繁项集中正关联规则的同时，能挖掘非频繁项集中的 $A\Rightarrow\neg B$, $\neg A\Rightarrow B$ 以及 $\neg A\Rightarrow\neg B$ 型负关联规则。文献[7]提出一种 PNARC(Positive and Negative Association Rules on Correlation)模型，该模型利用支持度——置信度框架，采用相关性检验方法，不仅能同时挖掘出频繁项集中的正、负关联规则，而且能检测并删除相互矛盾的规则。文献[8]中给出一种基于多置信度和 $\chi 2$ 检验的挖掘正负关联规则的方法，提出一种 PNARMC 算法，该算法不仅能够正确地产生正负关联规则，而且能灵活的控制关联规则的数量。文献[9]提出一种从频繁项集和非频繁项集中挖掘正负关联规则的方法，采用一种新的测量方法 VRRCC(Valid Association Rule based on Correlation Coefficient and Confidence)，并提出一种 PNAR_MLMS 算法。

1.2 负关联规则的挖掘技术

PNARC 模型[7]利用已知的正关联规则的支持度和置信度来计算负关联规则的支持度和置信度：设 $A, B\subset I, A\cap B=\Phi$，则有：

(1)$supp(\neg A) = 1\text{-}supp(A)$；

基金项目：山东省自然科学基金资助项目(Y2007G25)；山东省优秀中青年科学家奖励基金资助项目(2006BS01017)

作者简介：尚世菊(1982－)，女，硕士研究生，主研方向：数据挖掘；董祥军，教授；赵 龙，硕士研究生

收稿日期：2008-06-14 **E-mail：**shiju82@163.com

(2) $supp(A\cup\neg B)=supp(A)-supp(A\cup B)$；

(3) $supp(\neg A\cup B)=supp(B)-supp(A\cup B)$；

(4) $supp(\neg A\cup\neg B)=1-supp(A)-supp(B)+supp(A\cup B)$；

(5) $conf(A\Rightarrow\neg B)=\dfrac{supp(A)-supp(A\cup B)}{supp(A)}=1-conf(A\Rightarrow B)$；

(6) $conf(\neg A\Rightarrow B)=\dfrac{supp(B)-supp(A\cup B)}{1-supp(A)}=\dfrac{supp(B)-supp(A)\times conf(A\Rightarrow B)}{1-supp(A)}$；

(7) $conf(\neg A\Rightarrow\neg B)=\dfrac{1-supp(A)-supp(B)+supp(A\cup B))}{1-supp(A)}=1-conf(\neg A\Rightarrow B)$

文献[10]对项集的相关性进行了定义，将项集 A 和 B 看作 2 个随机事件，项集 A 和 B 支持度 $supp(A)$, $supp(B)$就是它们发生的概率 $P(A)$和 $P(B)$，其相关系数为

$$\rho_{AB}=\frac{supp(A\cup B)-supp(A)supp(B)}{\sqrt{supp(A)(1-supp(A))supp(B)(1-supp(B))}}$$

如果 $\rho_{AB}>0$，则(1)$\rho_{\neg AB}<0$；(2)$\rho_{A\neg B}<0$；(3)$\rho_{\neg A\neg B}>0$。反之亦反之。

ρ_{AB} 有 3 种可能情况：

(1)如果 $\rho_{AB}>0$，则 A 与 B 正相关，事件 A 出现的越多，事件 B 出现的也越多；

(2)如果 $\rho_{AB}=0$，则 A 与 B 相互独立，事件 B 的出现与事件 A 无关；

(3)如果 $\rho_{AB}<0$，那么 A 和 B 负相关，事件 A 出现的越多，事件 B 出现的越少。

在挖掘正、负关联规则时只要对项集的相关系数进行判断即可避免矛盾规则的出现，即当 $\rho_{AB}>0$ 时仅挖掘规则 $A\Rightarrow B$ 和$\neg A\Rightarrow\neg B$，当 $\rho_{AB}<0$ 时仅挖掘规则$\neg A\Rightarrow B$ 和 $A\Rightarrow\neg B$，当 $\rho_{AB}=0$ 时不挖掘规则。

文献[8]中提出一种 χ^2 检验的方法，在关联规则中 χ^2 检验多是用在检验 2 个项集的情况，即二维相依表，服从自由度为 1 的 χ^2 分布。假设事务数据库中的事务总数为 n，对应其中的项集 A 和 B 的 χ^2 值为：

$$\chi^2=\frac{n\times[supp(A\cup B)-supp(A)supp(B)]^2}{supp(A)supp(B)(1-supp(A))(1-supp(B))}$$

用 χ_α^2 来表示用户给定的显著性水平为α时的临界值，如果 $\chi^2\le\chi_\alpha^2$，则 A, B 相互独立，此时不挖掘关联规则，否则 A, B 相关，然后再根据相关系数的判定方法来得到正确的关联规则。

1.3 负关联规则的修剪技术

在负关联规则挖掘中，通常产生的规则数量很多，其中有很多不重要的、冗余的正负关联规则，使用户分析和利用这些规则变得十分困难，因此，要对挖掘出来的关联规则进行修剪。

1.3.1 基于相关强度的正负关联规则修剪方法

根据文献[10]中对相关系数和相关强度的介绍，相关强度是利用变量$\alpha(0\le\alpha\le1)$来表示项集 A 与 B 之间的相关的程度。若$\alpha>0.5$，则 A, B 为强相关；若 $0.3\le\alpha\le0.5$，则 A, B 为中度相关，$0.1\le\alpha\le0.3$，则为弱相关；若$\alpha<0.1$，则认为 A, B 的相关性可以忽略。可将相关系数 $\rho_{AB}\ge\alpha$作为一种约束来修剪那些作用不大的规则。

文献[9]将相关强度与最小支持度相结合，给出一种新的测量手段 $VARCC$：

$$VARCC(A,B,\alpha,minconf)=\frac{\rho_{AB}-\alpha+conf(A\Rightarrow B)-minconf+1}{|\rho_{AB}-\alpha|+|conf(A\Rightarrow B)-minconf|+1}$$

其中，α为相关强度，$minconf$ 为最小支持度。如果$|\rho_{AB}|<\alpha$，将不会从 $A\cup B$ 产生关联规则。

否则如果一条关联规则 $A\Rightarrow B$ 满足 $VARCC(A, B, \alpha, minconf)=1$，则它就是一条有效的关联规则，其他形式关联规则的判定与之类似。

1.3.2 基于多置信度的正负关联规则修剪方法

在购物篮分析中，设有商品 A, B，因 $supp(A)$和 $supp(B)$较小，则 $supp(A\cup B)$较小，而规则 $A\Rightarrow B$ 的置信度 $conf(A\Rightarrow B)$可能大，也可能小，但 $conf(\neg A\Rightarrow\neg B)$肯定较大，若对所有的规则采用统一置信度约束，就会出现这样尴尬的局面：若置信度较小，则会得到大量规则，导致用户无法从中选择真正需要的规则；若置信度较大，则可能会漏掉许多有价值的正关联规则。因此，文献[10]给出一种基于多置信度的挖掘正负关联规则的方法。此方法分析了 $conf(A\Rightarrow B)$值域与 $supp(A)$和 $supp(B)$的关系：

$$\max(0,\frac{supp(A)+supp(B)-1}{supp(A)})\le conf(A\Rightarrow B)\le\min(1,\frac{supp(B)}{supp(A)})$$

根据负关联规则置信度的计算方法和负关联规则的函数表示，同理可以计算出 $A\Rightarrow\neg B$, $\neg A\Rightarrow B$, $\neg A\Rightarrow\neg B$ 的值域。为了更直观，本文讨论当 $supp(A)$, $supp(B)$都很小时，$conf(A\Rightarrow B)$的变化对 3 种负关联规则置信度的影响，其他情况参照文献[10]。$supp(A)$, $supp(B)$很小时取值 0.1，很大时取值 0.9。$supp(A)supp(B)$都很小规则置信度的阈值如下：$conf(A\ \ B)$的值域是[0, 1]；$conf(A\ \ \neg B)$ 的值域是[0, 1]；$conf(\neg A\ \ B)$ 的值域是[0, 0.11]；$conf(\neg A\ \ \neg B)$ 的值域是[0.89, 1]。

从图 1 中可以看出置信度的取值对关联规则的影响非常大。当 $supp(A)$, $supp(B)$都很小时，$conf(\neg A\Rightarrow\neg B)$非常大，如使用一个置信度并设定为 0.6，结果中将会存在大量的$\neg A\Rightarrow\neg B$ 型规则，而不会有$\neg A\Rightarrow B$ 型的规则，这显然是不合理的，因此，应根据实际情况为不同形式的关联规则分别设定置信度，这样可以灵活地控制关联规则的数量，达到修剪规则的目的。

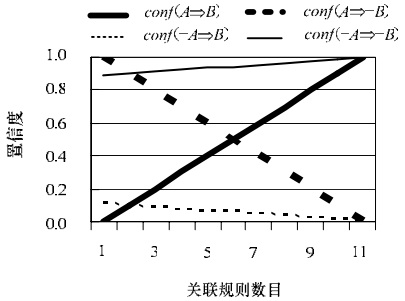

图 1 4 种置信度关系

2 多数据库中关联规则挖掘研究现状及主要技术

2.1 多数据库挖掘研究现状

对多数据库的挖掘可分为 3 个步骤：(1)对数据库进行分类；(2)挖掘同类数据库，即局部模式分析；(3)将同类数据库挖掘到的知识进行合成。

对于多数据库的挖掘，在文献[11]中根据用户需求提出了一种通过检索来挖掘多数据库中有趣知识的方法，但它只能挖掘所选择的数据库。对多数据库挖掘中数据库分类领域

的研究，文献[12-13]讨论了多种挖掘多数据库的方法，该方法能有效缩减大数据集的尺寸，去掉数据冗余和噪声，但只针对给定的应用，而在没有给定任何信息的情况下，不是很有效。文献[14]提出协同数据挖掘(CMD)模型，它利用一种合适的规范正交函数和计算标准系数来产生数据的全局模式。

2.2 多数据库中关联规则挖掘技术

在科学研究及应用中，通常用权值的方法来分析和合成不同数据库的信息，文献[15]也采用权值来合成多数据库的关联规则。一个规则的投票率指的是包含这个规则的子公司的数量，投票率越高，规则的权值越大，此规则也越有用。尽管每个子公司相对于总公司有平等的投票权，但由于子公司的销售额或在挖掘关联规则时设置的最小支持度和置信度不同，所以实际情况下，每个子公司对总公司的影响也是不同的。考虑到这些因素，应先合理确定规则的权值和数据库的权值，再对其进行合成。

(1)各个关联规则的权值

设 $S_1, S_2,\cdots, S_m$ 分别为各同类数据源的关联规则集，$S=\{S_1, S_2,\cdots, S_m\}$ 为总关联规则集，$R_1, R_2,\cdots, R_n$ 为总规则集 S 中具体的关联规则。R_i 的权值为

$$\omega_{R_i}=\frac{Num(R_i)}{\sum_{j=1}^{n} Num(R_j)}$$

其中，$i=1, 2,\cdots, n$，$Num(R)$表示具有规则 R 的数据库数目，即规则 R 的投票数。

(2)数据库的权值

设 $D_1, D_2,\cdots, D_m$ 为各分公司的数据库，S_i 为 D_i 中的关联规则集，而总规则集 $S=\{S_1, S_2,\cdots, S_m\}$，$R_1, R_2,\cdots, R_n$ 为 S 中具体的关联规则，数据库的权值为

$$\omega_{D_i}=\frac{\sum_{R_k\subset S_j} Num(R_k)\times\omega_{R_i}}{\sum_{j=1}^{m}\sum_{R_h\in S_j} Num(R_h)\times\omega_{R_h}}$$

(3)合成模式

设 $D_1, D_2,\cdots, D_m$ 为 m 个不同的数据库，规则集 S_i 是数据库 D_i ($i=1, 2,\cdots, m$)中的关联规则集。对于特定的关联规则“$A\Rightarrow B$”，假设 $\omega_1, \omega_2,\cdots, \omega_m$ 分别是数据库 $D_1, D_2,\cdots, D_m$ 的权值，则合成后的支持度和置信度分别为

$$supp_{\omega}(A\cup B)=\omega_1\times supp_1(A\cup B)+\omega_2\times supp_2(A\cup B)+\cdots+\omega_m\times supp_m(A\cup B)$$

$$conf_{\omega}(A\Rightarrow B)=\omega_1\times conf_1(A\Rightarrow B)+\omega_2\times conf_2(A\Rightarrow B)+\cdots+\omega_m\times conf_m(A\Rightarrow B)$$

根据上述方法合成的知识是个大约值，对于多数据库中负关联规则的挖掘只在文献[16]中进行了讨论，发现在对多数据库正负关联规则进行挖掘时，产生的负关联规则可能会和其他数据库中的正关联规则产生矛盾，该文提出一种解决矛盾的方法，但并没有给出具体的算法。

3 多数据库中负关联规则挖掘的发展趋势

3.1 多数据库中的负关联规则研究

多数据库中包含很多不同类的数据库，怎样挖掘出不同类数据库间的关联关系非常重要。例如：当麦当劳在某一地点开设一家分店后，肯德基会在随后的 2 个月，在 1 英里范围内开设一家肯德基分店。由于时间保存在不同的数据库内，使事务间关联规则的挖掘比事务内关联规则的挖掘更富有挑战性，也更具有重要性。怎样挖掘出多数据库中不同类数据库间的负关联规则将会是一个新的研究方向。

3.2 多数据库中时态负关联规则的研究

由于时间是数据本身固有的因素，数据库中的每个事务均有其有效时间，但目前有许多研究工作都假定得到的规则是永远有效的，然而事实并非如此。在现实中，附加上某种时态约束的规则将可以更好地描述客观现实情况，因而也会更有价值。如何将时态关联规则与多数据库中负关联规则的研究相结合，将是一个新的研究热点。

3.3 多数据库中负关联规则的选取

对于多数据库的挖掘需要考虑的因素很多，例如商品的利润，商品的保质期，子公司的效益等因素。怎样根据某一具体的要求从多数据库的负关联规则中选取所需规则，有待进一步研究。

参考文献

[1] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Database[C]//Proc. of the ACM SIGMOD'93. New York, USA: ACM Press, 1993: 207-216.

[2] 李学明, 刘勇国, 彭 军, 等. 扩展型关联规则和原关联规则及其若干性质[J]. 计算机研究与发展, 2002, 39(12): 1740-1750.

[3] Brin S, Motwani R, Silverstein C. Beyond Market: Generalizing Association Rules to Correlations[C]//Proc. of the ACM SIGMOD'97. Tucson, Arizona, USA: ACM Press, 1997: 265-276.

[4] Savasere A, Omiecinski E, Navathe S. Mining for Strong Negative Associations in a Large Database of Customer Transaction[C]//Proc. of the 14th International Conference on Data Engineering. Orlando, Florida, USA: [s. n.], 1998: 494-502.

[5] Wu Xindong, Zhang Chengqi, Zhang Shichao. Mining both Positive and Negative Association Rules[C]//Proc. of the 19th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 2002: 658-665.

[6] Zhang Chengqi, Zhang Shichao. Association Rule Mining[M]. Heidelberg, Germany: Springer-Verlag, 2002: 47-84.

[7] 董祥军, 王淑静, 宋瀚涛, 等. 负关联规则的研究[J]. 北京理工大学学报, 2004, 24(11): 978-981.

[8] Dong Xiangjun, Sun Fengrong, Han Xiqing, et al. Study of Positive and Negative Association Rules Based on Multi-confidence and Chi-Squared Test[M]. Heidelberg, Germany: Springer, 2006: 100-109.

[9] Dong Xiangjun, Niu Zhendong, Shi Xuelin, et al. Mining both Positive and Negative Association Rules from Frequent and Infrequent Itemsets[M]. Heidelberg, Germany: Springer, 2007: 122-133.

[10] Cohen J. Statistical Power Analysis for the Behavioral Sciences [M]. New Jersey, USA: Lawrence Erlbaum Assotiates Publishers, 1988.

[11] Yao Jun, Liu Huan. Searching Multiple Databases for Interesting Complexes[C]//Proc. of PAKDD'97. Singapore: [s. n.], 1997: 198-210.

[12] Zhong Ning, Yan Yiyu, Ohsuga S. Peculiarity Oriented Multi-database Mining[C]//Proc. of PKDD'99. Berlin, Germany: Springer-Verlag, 1999: 136-145.

[13] Ribera J, Kanfman K, Kerschberg L. Knowledge Discovery from Multiple Databases[C]//Proc. of the 1st International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA, USA: AAAI Press, 1995: 240-245.

（下转第 93 页）