一种发现函数依赖集的方法及应用*

张守志+, 施伯乐

(复旦大学 计算机与信息技术系,上海 200433)

A Method for Discovering Functional Dependencies and Its Application

ZHANG Shou-Zhi+, SHI Bai-Le

(Department of Computing and Information Technology, Fudan University, Shanghai 200433, China)

+ Corresponding author: Phn: 86-21-65642219, Fax: 86-21-65642219, E-mail: shouzhi_zhang@fudan.edu.cn

http://www.fudan.edu.cn

Received 2002-12-17; Accepted 2003-03-05

Zhang SZ, Shi BL. A method for discovering functional dependencies and its application. *Journal of Software*, 2003,14(10):1692~1696.

http://www.jos.org.cn/1000-9825/14/1692.htm

Abstract: An efficient method is introduced for discovering minimal functional dependencies from large database. It is based on the concept of agree sets. From agree sets, maximal sets and its complements are derived, and all minimal non-trivial functional dependencies can be generated. The computation of agree sets can be decreased by using stripped partition database. A levelwise algorithm is used for computing the left hand sides of minimal non-trivial functional dependencies. This method can be used to attribute reduction, clustering and mining associate rules, etc. in knowledge discovery as well as reorganization and design of databases.

Key words: minimal functional dependencies; agree set; hypergraph; attribute reduction

摘 要: 介绍了一种发现最小函数依赖集的方法.这种方法基于一致集的概念,根据一致集导出最大集及其补集,然后生成最小非平凡函数依赖集.通过使用带状划分数据库减少求一致集的运算次数,使用逐层求精的算法来计算最小非平凡函数依赖集的左部.其结果可用于数据库的重新组织和设计、属性约简、聚类、关联规则提取等知识发现工作中.

关键词: 最小函数依赖集;一致集;超图;属性约简

中图法分类号: TP311 **文献标识码**: A

函数依赖反映了现实世界中数据的完整性约束,对关系数据库的设计和分析起着重要的作用.从大量数据中发现的函数依赖反映了属性间的关联性和数据的完整性约束.数据库管理员通过对这些函数依赖的评价,能够极为方便地对数据库进行维护以及对它们的关系模式进行重新组织.已有多种方法[1~3]来发现一个数据库中的函数依赖.文献[1]中提出了一个发现最小非平凡函数依赖的架构.它基于一致集(agree set)的概念[4],根据一致集导出最大集(maximal set),然后生成最小非平凡函数依赖集.我们介绍一种新的方法来计算一致集、最大集和

* Supported by the National Natural Science Foundation of China under Grant No.69933010 (国家自然科学基金)

第一作者简介: 张守志(1965－),男,四川巴中人,博士,副教授,主要研究领域为数据库,知识库.

© 中国科学院软件研究所 http://www.jos.org.cn

最小非平凡函数依赖集的左部(LHS)[5],对其中的结论给出了证明.我们发现,对最小非平凡函数依赖集的发现方法可以用到对一致决策表的属性约简上.

本文第 1 节介绍关系数据库中的一些术语;第 2 节介绍新的方法,并对其中的结论给出证明;第 3 节介绍这种方法用在一致决策表的属性约简上;第 4 节是结束语.

1 基本定义

本节给出与第 2 节有关的一些术语.

设 R 是一个关系模式,R 上的一个函数依赖表示为 $X\to A, X\subseteq R, A\in R$.$r$ 表示 R 的关系,r 中的一个函数依赖 $X\to A$(表示为 $r\models X\to A$)成立 iff $\forall t,t'\in r, t[X]=t'[X]\Rightarrow t[A]=t'[A]$.若 $X\to A, X'\subset X$,不成立 $X'\to A$,则称 $X\to A$ 是最小的函数依赖.若 $A\in X$,则称 $X\to A$ 是平凡的函数依赖.

用 $dep(r)$表示 r 中成立的所有函数依赖集,即 $dep(r)=\{X\to A|X\cup A\subseteq R, r\models X\to A\}$.

设 $t,t'\in r, X\subseteq R$;若 $t[X]=t'[X]$,则称 t 和 t'在 X 上一致.t 和 t'的一致集(agree set)定义为 $ag(t,t')=\{A\in R|t[A]=t'[A]\}$.关系 r 的一致集定义为 $ag(r)=\{ag(t,t')|t,t'\in r, t\neq t'\}$.

设 $A\in R$,属性 A 的最大集(maximal set)$\max(dep(r),A)$是 r 中不能决定 A 的最大属性集 X 的集合,即 $\max(dep(r),A)=\{X\subseteq R|r\not\models X\to A$ and $\forall Y\subseteq R, X\subset Y, r\models Y\to A\}$.

设 $A\in R, r$ 中函数依赖集 $dep(r)$的左部集定义为 $lhs(dep(r),A)=\{X\subseteq R|r\models X\to A$ and $\forall X'\subset X, r\not\models X'\to A\}$.显然,$\{X\to A|X\in lhs(dep(r),A), A\in R\}$与 $dep(r)$等价.

2 一种发现最小非平凡函数依赖集的方法

这种方法的步骤是计算一致集 $ag(r)$,根据 $ag(r)$计算最大集 $\max(dep(r),A)$和它的补集 $c\max(dep(r),A)$,由 $c\max$ 集算出 $lhs(dep(r),A)$,其中只有计算 $ag(r)$与关系 r 有关.因为 $ag(r)=\{ag(t,t')|t,t'\in r, t\neq t'\}$,若 r 中有 m 个元组,则不同元组的配对数达到 $O(m^2)$;计算每对元组 t,t'的 $ag(t,t')$需要 $O(n)$,n 为 R 的属性个数;这样,朴素计算 $ag(r)$的时间复杂度达到 $O(nm^2)$(没有考虑数据内外存交换的时间).所以,当 m 足够大时,这样计算 $ag(r)$就变得不实际了.这里介绍一种计算 $ag(r)$的新方法[5],对其结论进行了证明.在不影响计算结果的前提下,减少候选的元组对,把对计算 $ag(r)$无贡献的元组对排除掉;这种方法使用带状划分数据库(stripped partition database)[2,5]概念来实现.

2.1 带状划分数据库

设 $X\subseteq R, t\in r, [t]_x=\{t'\in r|t[A]=t'[A], \forall A\in X\}$,则 $[t]_x$ 是 X 的一个基本等价类;X 在 r 中形成的基本等价类的集合构成 r 的一个划分(partition),表示为 $X^*=\{[t]_x|t\in r\}$.若划分 X^* 中的一个等价类只有一个元组,则 r 中没有与它在 X 上一致的其他元组.所以,我们定义属性集 X 的带状划分(stripped partition) $\overline{X}^*=\{c\in X^*||c|>1\}$.

例 1:下面的关系表示职工分派到部门(系)的情况[5].

元组号	empnum	Depnum	year	depname	Mgr
1	1	1	1980	Mathematics	5
2	1	5	1992	Admission	7
3	2	2	1990	Computer Science	3
4	3	2	1996	Computer Science	3
5	4	3	1996	Physics	3
6	5	1	1978	Mathematics	5
7	6	5	1986	Admission	7

为了简便,我们把表中 5 个属性分别用 A,B,C,D,E 来代表.属性 A 的划分 $A^*=\{\{1,2\},\{3\},\{4\},\{5\},\{6\},\{7\}\}$,$A$ 的带状划分 $\overline{A}^*=\{\{1,2\}\}$.

设 r 是 R 的一个关系,r 的带状划分数据库 $\bar{r}$ 由 R 上每个属性 A 的带状划分 $\overline{A}^*$ 构成,即 $\bar{r}=\{\overline{A}^*|A\in R\}$,可直接从 r 计算出来,其最坏时间复杂度为 $O(mn)$.

© 中国科学院软件研究所 http://www.jos.org.cn

2.2 计算一致集$ag(r)$

我们先定义一个最大等价类的集合 MC=Max⊆{$c\in \overline{A}^* \mid \overline{A}^* \in \bar{r}$},其中的算子"Max⊆"是求集合中的按⊆最大的成员.

例 2 (Cont.):例 1 中的 MC={{1,2},{1,6},{2,7},{3,4,5}}.

定理 1. $ag(r)=\cup_{c\in MC}ag(c)$,其中 $ag(c)=\{ag(t,t')|t,t'\in c,t\neq t'\}$.

证明:设 $t,t'\in r,t\neq t'$.

若 $ag(t,t')\neq\varnothing$,不妨设 $A\in ag(t,t')$.

$\therefore t[A]=t'[A],\therefore \exists c'\in \overline{A}^*,t,t'\in c'$.

根据 MC 的定义,$\exists c\in MC,c'\subseteq c,\therefore t,t'\in c$.

$\therefore ag(r)=\cup_{c\in MC}ag(c)$. □

根据定理 1,当计算 $ag(r)$时,只需考虑 MC 等价类中配对的元组.当 r 中属性的取值大不相同时,通过带状划分数据库 $\bar{r}$ 去除掉很多单元组组成的等价类,使其不与其他元组配成对,这样就使计算 $ag(r)$的时间大大减少了.

例 3 (Cont.):根据例 2 的 MC 产生的配对元组有(1,2),(1,6),(2,7),(3,4),(3,5),(4,5);$ag(r)$={A,BDE,E,CE}.

2.3 计算最大集$\max(dep(r),A)$和它的补集$c\max(dep(r),A)$

定理 2. 设 $A\in R$,$\max(dep(r),A)$=Max⊆{$X\in ag(r)|A\notin X$}.

证明:设 $X\in$Max⊆{$X\in ag(r)|A\notin X$}.

$\therefore \exists t,t'\in r,ag(t,t')=X,\therefore t[X]=t'[X]$.

$\because A\notin X,\therefore A\notin ag(t,t'),\therefore t[A]\neq t'[A]$.

$\therefore r|\neq X\rightarrow A$.

对 X 的任意一个超集 $X'\subseteq R$,若 $A\in X'$,则 $r|=X'\rightarrow A$;若 $A\notin X'$,则根据 X 的性质,$X'\notin ag(r)$,所以$\forall t,t'\in r,t\neq t',t[X']\neq t'[X'],\therefore r|=X'\rightarrow A$.所以,$X\in\max(dep(r),A)$.

另一方面,设 $X\in\max(dep(r),A)$={$X|r|\neq X\rightarrow A,X$ 的任意一个超集 $X'\subseteq R,r|=X'\rightarrow A$}.

$\because r|\neq X\rightarrow A,\therefore A\notin X$,并且$\exists t,t'\in r,X\subseteq ag(t,t')$.

若$\exists B\in ag(t,t'),B\notin X$,则 $r|\neq BX\rightarrow A$,这与 X 的性质矛盾,所以 $X=ag(t,t')$.同时,由 X 的最大的性质,不存在 X 的超集 $X'\subseteq R,X'\in ag(r)$且 $A\notin X'$,所以 $X\in$Max⊆{$X\in ag(r)|A\notin X$}. □

定义 $c\max(dep(r),A)$={$R-X|X\in\max(dep(r),A)$},$A\in R$.

例 4 (Cont.): $\max(dep(r),A)$={BDE,CE},$c\max(dep(r),A)$={AC,ABD}.

2.4 计算函数依赖集左部$lhs(dep(r),A)$

我们引进超图(hypergraph)概念.R 的属性集组成的 H 构成一个简单的超图,H 中的属性集称为超图的边,每个属性称为超图的顶点.H 的一个横截(transversal)T 是 R 的一个属性集,它与 H 的每一条边都相交.H 的一个最小横截 T 满足:T 是 H 的一个横截,$\forall T'\subset T,T'$不是 H 的横截.H 的所有最小横截集记为 $T_r(H)$.把 $c\max(dep(r),A)$看成一个超图,则有下面的结论:

定理 3: $T_r(c\max(dep(r),A))=lhs(dep(r),A)$.

证明:一方面,设 $T\in T_r(c\max(dep(r),A))$.

若 $r|\neq T\rightarrow A$,则$\exists X\in\max(dep(r),A),T\subseteq X$.

$\therefore T\cap\overline{X}=\varnothing$,其中 $\overline{X}=R-X\in c\max(dep(r),A)$,与 $T\in T_r(c\max(dep(r),A))$矛盾.所以,$r|=T\rightarrow A$.

又 T 是最小的横截,$\therefore\forall T'\subset T,\exists\overline{X}\in c\max(dep(r),A),T'\cap\overline{X}=\varnothing$.

$\therefore T'\cap X=T',X=R-\overline{X}\in\max(dep(r),A),\therefore r|\neq T'\rightarrow A,\therefore T\in lhs(dep(r),A)$.

另一方面,设 $Y\in lhs(dep(r),A)$.

若 $Y\notin T_r(c\max(dep(r),A))$,则$\exists\overline{X}\in c\max(dep(r),A),Y\cap\overline{X}=\varnothing$.

$\therefore Y\subseteq X,X=R-\overline{X}\in\max(dep(r),A),\therefore r|\neq Y\rightarrow A$,与 $Y\in lhs(dep(r),A)$矛盾.

© 中国科学院软件研究所 http://www.jos.org.cn

$\therefore Y\in T_r(c\max(dep(r),A))$. □

下面给出计算 $lhs(dep(r))$的算法.

算法. 计算 $lhs(dep(r))$.

输入:$c\max(dep(r))$;

输出:最小函数依赖集的左部 $lhs(dep(r))$.

for $\forall A\in R$ do

begin i:=1;

L_i:={{B}|$B\in X, X\in c\max(dep(r), A)$};

While $L_i\neq\varnothing$ do

begin

$LHS_i[A]$:={$l\in L_i|l\cap X\neq\varnothing, \forall X\in c\max(dep(r),A)$};

L_i:=L_i–$LHS_i[A]$;

L_{i+1}:={l'||l'|=i+1 and $\forall l\subset l', |l|=i\Rightarrow l\in L_i$};

i:=i+1;

end

$lhs(dep(r),A)=\cup LHS_i[A]$

end

例 5 (Cont.):$lhs(dep(r),A)$={A,BC,CD},$lhs(dep(r),B)$={AC,AE,B,D}等.

由 $lhs(dep(r),A)$得到关于 A 的最小函数依赖集:

$$r|=BC\rightarrow A, r|=CD\rightarrow A.$$

文献[2]中提出的 Tane 算法是基于逐层求精的算法,按从小到大顺序搜索函数依赖集的左部,通过验证在左部一致的元组在其右部是否也一致来确定一个函数依赖是否成立.该算法最坏时间复杂度为 $O((m+n^{2.5})2^n)$.在本文介绍的方法中,设 s 为由 MC 中配对的元组对的数目,$k=|ag(r)|$,则该方法的时间复杂度的上界为 $O(mn+sn+n^2k)$.若 $s<<mn, k<<2^n$,则算法的时间复杂度较优.

3 一致决策表上属性约简

一个决策表 $T=(U,A,C,D)$,其中 U 是非空、有限的个体集合,A 是非空、有限的属性集合,$A=C\cup D, C\cap D=\varnothing$;$C,D$ 分别称为 A 的条件属性集和决策属性集.

设 $k=|POD_C(D)|/|U|$,其中 $POD_C(D)=\cup_{X\in D^*}\underline{C}X, \underline{C}X=\cup\{Y\in C^*|Y\subseteq X\}$.在一个决策表 T 中,决策属性 D 依赖于条件属性 C 的程度以 $k(0\leq k\leq 1)$来反映.若 $k=1$,则 D 完全(totally)依赖于 C,记为 $T|=C\rightarrow D$.这意味着:在已知条件 C 下,可将 U 上全部个体分成 D 基本等价类.当 $T|=C\rightarrow D$ 时,我们称 T 是一致的.设 $E\subseteq C$, $POD_E(D)=POD_C(D)$ 且 E 中每个属性都不可被约去,即$\forall E'\subset E$, $POD_{E'}(D)\neq POD_E(D)$,则称 E 是 $D_$约简.

从上面分析可以得出,在一致决策表 $T=(U,A,C,D)$中寻找 C 的 $D_$约简,就是从 T 中找出满足 $T|=E\rightarrow D, E\subseteq C$,且$\forall E'\subset E, T|\neq E'\rightarrow D$ 的 E.由于 $T|=E\rightarrow D$ 就是 T 中满足的函数依赖,所以在一致决策表中求 E 就变成了求 $lhs(dep(T),D)$.

例 6:考察下面的一致决策表,其中 C={a,b,c},D={d,e}.我们求得 $lhs(dep(T),D)$={ac,bc},所以 T 中的两个约简是{a,c}和{b,c},其核心为属性 c.

U	a	b	c	d	e
1	1	0	2	1	1
2	2	1	0	1	0
3	2	1	2	0	2
4	1	2	2	1	1
5	1	2	0	0	2

© 中国科学院软件研究所 http://www.jos.org.cn

4 结 语

本文介绍了一种在大量的数据中发现最小函数依赖集的方法以及在一致决策表上属性约简的应用.依赖集反映了数据中存在的完整性约束,便于DBA对数据库进行重新组织和设计;属性的约简广泛地应用于数据挖掘的预处理、决策规则提取等过程中.

References:

[1] Mannila H, Räihä K-J. Algorithms for inferring functional dependencies from relations. Data and Knowledge Engineering, 1994, 12(1):83~99.

[2] Huhtala Y, Kärkkäinen J, Porkka P, Toivonen H. Tane: An efficient algorithm for discovering functional and approximate dependencies. The Computer Journal, 1999,42(2):100~111.

[3] Savnik I. Bottom-Up induction of functional dependencies from relations. In: Piatetsky-Shapiro G, ed. Proceedings of the AAAI'93 Workshop on Knowledge Discovery in Databases. 1993. 174~185.

[4] Beeri C, Dowd M, Fagin R, Statman R. On the structure of Armstrong relations for functional dependencies. Journal of the ACM, 1984,31(1):30~46.

[5] Lopes S, Petit J-M, Lakhal L. Efficient discovery of functional dependencies and Armstrong relations. In: Proceedings of the EDBT 2000. LNCS 1777, Heidelberg: Springer -Verlag, 2000. 350~364.

2004年全国理论计算机科学学术年会

征 文 通 知

由中国计算机学会理论计算机科学专业委员会主办，海军工程大学信息与电气学院承办的“2004年全国理论计算机科学学术年会”将于2004年10月在武汉召开。会议录用论文将收录在正式出版的论文集中，欢迎大家积极投稿。现将有关征文要求通知如下：

1．应征论文应未在其他刊物或学术会议上正式发表过。特别欢迎有创见的论文和有应用前景的论文。

2．稿件要求用计算机打印，格式为38行×38字，字体为5号宋体。稿件中的图形要求画得工整、清晰、紧凑，尺寸要尽量小；图中字体要求为六号宋体。稿件正文不超过六千字。标题、作者姓名、作者单位、摘要、关键词采用中英文间隔行文。稿件各部分依次为：一、引言；二、...；最后是结束语。附录放在参考文献之后；参考文献限已公开发表的，文中最好不要出现文献序号。参考文献的格式为：

序号 作者·书名·出版社所在地：出版社名，出版年代

序号 作者·论文名·出处，年代·卷号（期号）：起迄页码

务必附上第一作者简历（姓名、性别、出生年月、职称、学位、研究方向等）、通信地址和联系电话。并注明论文所属领域。请提供打印稿和电子稿各一份。来稿一律不退，请自留底稿。

3．征文范围

程序理论（程序逻辑、程序正确性验证、形式开发方法等）；计算理论（算法设计与分析、复杂性理论、可计算性理论等）；语言理论（形式语言理论、自动机理论、形式语义学、计算语言学等）；人工智能（知识工程、机器学习、模式识别、机器人等）；逻辑基础（数理逻辑、多值逻辑、模糊逻辑、模态逻辑、直觉主义逻辑、组合逻辑等）；数据理论（演绎数据库、关系数据库、面向对象数据库等）；计算机数学（符号计算、数学定理证明、计算几何等）；并行算法（分布式并行算法、大规模并行算法、演化算法等）。

4．征文截止日期：2004年5月1日

5．论文投寄地址：（430033）武汉 海军工程大学信息与电气学院 张志祥 收

联系电话：027-83443985,83443984（张志祥，贲可荣）

电子信箱：tcs2004@vip.sina.com; hgzzx@163.com

© 中国科学院软件研究所 http://www.jos.org.cn