维普资讯 http://www.cqvip.com

利用领域知识改进数据仓库中知识发现查询的效率

王新军　洪晓光　王海洋　马绍汉
（山东大学计算机科学与技术学院　济南250100）

摘　要　随着数据库的规模和存储的数据量越来越大，许多知识发现方面的应用不得不面对大量的数据，这就迫切需要找出降低知识发现查询计算的复杂性的方法。以往的语义查询优化工作都是利用领域知识进行查询优化，主要是靠重写查询语句为一个等价的、高效的查询来降低费用。本文介绍了一种利用三类领域知识进行优化处理、缩小搜索范围、降低查询费用、提高被发现知识的质量的方法。

关键词　知识发现，数据仓库，领域知识，查询优化

Improving Query Efficiency of Knowledge Discovery in Data Warehouse by Using Domain Knowledge

WANG Xin-Jun　HONG Xiao-Guang　WANG Hai-Yang　MA Shao-Han
(School of Computer Science and Technology, Shandong University, Jinan 250100)

Abstract　As the scale and amount of data in database grow, many applications in knowledge discovery have to deal with it. It is necessary to find the way to simplify the query computation for knowledge discovery. All the works of semantic query optimization before are to decrease the cost by rewriting the query sentence to an equivalent and more efficient one. In this paper, a new method is proposed to improve the quality of discovered knowledge and decrease the cost by using three sorts of domain knowledge.

Keywords　Knowledge, Data warehouse, Domain knowledge, Query optimization

1　介绍

随着数据获取和存储技术的发展，数据库的规模和存储的数据量越来越大，同时由于人们对数据多方面、深层次的需求，使得数据仓库、知识发现和数据挖掘成为研究的热点。知识发现和数据挖掘就是从数据库中找寻隐含的、先前未知的、有潜在用途的模式，如银行根据储户的信贷历史情况提出更好的贷款建议和破产预警模式；超市根据消费情况找出更好的销售模式。

知识发现问题研究的重点是如何有效地对大量原始数据进行处理，生成针对某一用户需求的、具有特殊意义的模式。但是要在包含数以千兆、兆兆的历史数据的数据仓库中进行彻底的搜索几乎是不可能的，这种情况下如何通过对数据进行约束将搜索范围缩小到表中的某些特定属性和/或是某个子集上，从而快速地从数据中找到相关部分而又不丢失有用的知识就成为我们关心的重点问题。

本文给出了领域知识的分类及利用领域知识扩展语义查询优化技术，通过减少搜索空间和缩小搜索范围以提高知识发现的效率的方法。另外还提出一种不需要搜索全部领域知识而找到相关领域知识的方法。

本文第2节介绍了相关的基本概念和例子，第3节介绍了一些相关的工作，第4节描述了运用领域知识改进知识发现处理过程的方法，最后是总结和展望。

2　基本概念和例子

2.1　数据仓库

1993年W.H.Inmon基于传统的关系数据库管理系统，借鉴了历史数据库和知识发现技术，提出了数据仓库概念：是支持管理决策过程的、集成的、随时间而变的、持久的数据集合。数据仓库从运营系统及外部数据源提取相关的历史数据，经过归纳、集成，解决了数据冲突和表达不一致等问题，形成一个新的面向主题信息的、更适合于最终用户复杂业务分析的数据库，它能为管理者、信息分析师以及销售人员提供更有针对性、更具战略性的决策信息[1]。

2.2　基于数据仓库的知识发现系统

知识发现也称为数据挖掘，是从大量数据中筛选信息，发现对某种决策有价值的知识和规则。Fayyad等人对知识发现的定义为：从数据库中获取正确、新颖、有潜在应用价值和最终可理解的模式的非平凡过程。

基于数据仓库的知识发现系统，即KDDW(Knowledge Discovery in Data Warehouse)的输入主要来源于数据仓库的数据、专家指导以及存储在知识库中的知识和经验，通过一个知识发现引擎的处理生成有潜在应用价值的知识。知识发现引擎中提供了大量的挖掘算法，在知识库的指导下以合适的方式从数据仓库中抽取相关的数据及数据结构。

2.3　模式

知识发现的目的就是从数据仓库存储的大量数据中发现有用的模式(Pattern)以更好地利用已有数据。Piatetsky-Shapiro and Frawley在文[2]中收集了多种模式发现的方法，他对模式的定义为：

给定一个事实(数据)集合F、一种语言L和某个确定度C，一种模式就是语言L的一条语句，该语句描述了具有某个确定度C的F的一个子集F_s中数据之间的关系，S显然要比枚举F_s中的事实简单。

模式可以理解为知识，即关于某个特定领域和任务的数据库中各数据元素关系中对我们有用的信息。

王新军　副教授，博士生，研究方向为数据库、工作流管理系统、软件工程、数据仓库。

维普资讯 http://www.cqvip.com

2.4 领域知识

领域知识(domain Knowledge)通常是指在数据库中没有被明确表达的信息(明确表达和存储的称为事实或数据),在知识发现系统中利用领域知识将大大减少数据库搜索的时间。

下面几个例子说明了领域知识在查找有用知识中的作用:

例1 有一个乒乓球运动员的数据库,存储包含当前世界排名、世界级比赛的个人最好成绩等在内的运动员基本信息。我们想找世界排名进入前16名的、世界级比赛的个人最好成绩进入前8名的运动员。假如我们具有以下领域知识:"所有世界排名进入前16名的运动员都称为国际级运动健将"、"所有世界级比赛的个人最好成绩进入前8名的运动员都称为国际级运动健将"。利用这些领域知识,我们只需考虑那些技术等级标准为"国际级运动健将"的运动员即可,因此可重写查询不必去判断世界排名和世界级比赛个人最好成绩属性了,如果在技术等级标准属性上有索引,就更节省时间了。

例2 有一个数码相机销售事务的数据库,我们想要分辨率在300万像素以上的、价格在1000元以下的进口相机。假如我们具有以下领域知识:"没有分辨率在300万像素以上的、价格在1000元以下的进口相机",则利用该知识我们就不用访问数据库了,从而省去了搜索空间的分配。

例3 再来看一个超市零售业务数据库的例子。超市经营者可能会查询"消费者的消费模式是什么?",如果我们掌握的领域知识是把消费者按年龄分成老年、中老年、中年和青年四类;按收入分成高、中、低三类,就可以进行更细化、更有意义的查询。例如可以查询"高收入的老年消费者的消费模式是什么?"

例4 对一个公司员工数据库,我们想找月工资高于3000元的员工。假定我们具有属性之间关系的领域知识,例如工资、职务、学历之间是有关系的,而工资和住址、身份证号之间是无关的。如果有了这些领域知识,我们在做上面查询时就没有必要考虑住址、身份证号等属性了。

通过上面例子可以看出在进行知识发现过程中,利用领域知识可以减少记录数或无关属性,或者是通过施加限制细化原来的查询从而达到降低数据库访问量的目的。我们可以将一个带条件的知识发现查询转化成一个含有领域知识的、更高效的查询。另外我们还能够指出与已知领域知识相矛盾的查询是不可满足的。

3 相关的工作

对于存储着大量(甚至是海量)数据的数据仓库来说,即便是最快速的算法要遍历所有数据开销也是很大的,因此需要利用一定的方法降低费用。最简单的一种方法是对数据仓库进行随机抽样,但缺点是可能会丢失有用数据,即便抽样方法是很合理的也很难确定抽样的数量。还有一种方法是使用另外的数据库工具,如 OLAP,定义数据的一个子集,但并不总是可行的。

我们更倾向于另外一种将语义查询优化技术应用并扩展于知识发现处理的方法。语义查询优化就是将一种查询转换成另一种等价的、高效的查询形式,在关系数据库和演绎数据库中都有对语义查询优化的研究。文[3]描述了一种利用启发式规则进行转换的算法,通过说明每条启发式规则如何用于转换一个查询以对限制转换数目有帮助。文[4]描述了利用一个知识库进行语义查询优化的方法。文[5]提出的方法通过与语义知识表达式进行比较而修改原有查询表达式以提高执行效率。

文[6]将语义查询优化的方法加以扩充并运用于知识查询处理中,介绍了一种特殊方法用以表达领域知识并在知识发现处理中运用领域知识提出控制数据的规模与质量的策略。文[6]将领域知识分成三类:属性内部(interfield)领域知识、分类(category)领域知识和关联(correlation)领域知识。

属性内部领域知识描述了一个关系内部的各个属性之间的关系,形如:

If (A_1 op B_1)and/or … and/or (A_{n-1} op B_{n-1})then (A_n op B_n)and/or … and/or (A_m op B_m)

其中 $A_i(1\leqslant i\leqslant m)$是属性名,$B_j(1\leqslant j\leqslant m)$是属性名或是 A_j 值域中的一个常量,op 为算术比较运算符。

如对例1的乒乓球运动员数据库,相关的属性内部领域知识可表示为:

If (世界排名)≤16 or (世界级比赛个人最好成绩≤8) then 技术等级标准="国际级运动健将"

例2中用到的也是属性内部领域知识,可表示为:

If (分辨率>300万像素)then 价格>1000

通常 if 和 then 之间的条件作为该条知识的体,then 后面的是该条知识的头。

分类领域知识表达了某个属性的值域上的有用的分类,形如:

If (A_1 op B_1)and/or … and/or (A_{n-1} op B_{n-1})then An="category name"

其中 A_i、op 的含义同上,$B_j(1\leqslant j\leqslant m)$是 A_j 值域中的一个常量或是其他分类领域知识已经定义了的一个类名。值域中不同的值可以对应同一个类名。如例3中的消费者可以根据年龄分为如下四类:

If (年龄≥65)then 年龄="老年"

If (年龄≥45)and (年龄<65)then age="中老年"

If (年龄≥35)and (年龄<45)then age="中年"

If (年龄<35)then 年龄="青年"

类似地,可以定义另外的分类,如根据月薪把消费者分成三类:高、中、低。分类领域知识也可以由两个或多个类名合成,表明来自多个属性的分类之间的全局关系。例如可以基于年龄和月薪两个属性上的分类定义"重点的消费者"如下:

If 年龄="中年" and 月薪="高"then 消费者类型="重点"

关联领域知识是指属性间的关联关系,可以用集合形式表示。例如针对员工数据库来说,员工的学历、职务和收入之间关联关系的领域知识可以表示成{学历,职务,月薪}。因此寻找和月薪有关的模式时就会关心学历和职务,而舍弃和月薪无关的地址和身份证号等属性。

文[6]中提出了通过预处理和执行两个阶段利用领域知识改进知识发现的思想,本文在上述工作基础上给出了一个更加可行的执行方案。

4 引入领域知识之后的发现系统

引入领域知识之后,基于数据仓库的知识发现系统将按下面的步骤缩小搜索范围,提高知识发现的效率。

(1)领域知识的输入和存储

领域知识不属于数据,它可以有许多来源,数据辞典中的典型信息包括数据库表中每个字段的名字、类型和一些约束等,都是领域知识的最基本形式。领域知识的另一个来源是由领域专家提供的大量领域知识,在知识发现系统中新近发现的模式也可以被加入到领域知识集合中供以后使用。按上一

维普资讯 http://www.cqvip.com

节所述，领域知识分成属性内部领域知识、分类领域知识和关联领域知识三类。

领域知识将存储在知识库中，相应的表结构和一个实例如表1所示。表中 operand1、operand2和 operator 分别对应第3节给出的领域知识描述中的 A_i、B_i 和 op，对属性内部领域知识和分类领域知识通过 num 标识出有关该属性的、该类别的不同领域知识，由字段 head 标示出其是头(结论)还是体(条件)，同时该字段还可存储 and/or。关联领域知识表达的一组关联属性通过 num 值标识。还可以将集群存储在表中并在属性上创建索引以提高查找领域知识的速度。

表1 领域知识表的一个实例

Attribute	TYPE	Num	Head	Operand1	Operator	Operand2
Salary	interfiled	1	No	Salary	>	10000
Salary	interfiled		Yes	Position	=	Manager
Salary	interfiled	2	No	Salary	≤	10000
Salary	interfiled		Yes	Position	=	Employee
Salary	category	1	No	Salary	>	10000
Salary	category		Yes	Salary	=	"High"
Salary	category	2	No	Salary	>	3000
Salary	category		and	Salary	≤	10000
Salary	category		Yes	Salary	=	"Middle"
Salary	category	3	No	Salary	≤	3000
Salary	category		Yes	Salary	=	"Low"
Salary	correlation	1				
Education	correlation	1				
Position	correlation	1				

(2)利用领域知识进行查询优化改写

当 KDDW 接收到一个知识发现查询时，根据查询中涉及的属性在领域知识表(DKT，Domain Knowledge Table)中查询相关的领域知识，将查询转换成等价的、更高效的另外一种形式。我们将对找到的每一条领域知识做如下处理：

```
select * from DKT where attribute="attribute_name"
for 每条领域知识
switch TYPE
    {"interfiled":若 head="yes"的查询条件是原知识发现查询条件
                  的一个子集，则从原查询语句中去除其余不必要
                  的、冗余的条件；
                  若 head="yes"的查询条件是对原知识发现查询条
                  件的一个有用的补充，则将其加到原查询语句的条
                  件中；
     "category":将关于该属性的所有分类信息显示出来，由用户选
                择(可放弃选择)有用的、有意义的分类，并将选中
                的分类条件加到原查询语句的条件中；
     "correlation":在原知识发现查询语句基础上，增加在与当前属
                   性具有相同 num 值的属性列表上进行投影的运
                   算；
    }
```

(3)新领域知识的积累

对在知识发现系统中新近发现的模式按(1)的方法追加到领域知识表中。

通过上述三种情况的处理，可分别以减少关系内部扫描次数、细化查询条件和削减属性的个数的方式降低数据集合的大小、缩小搜索范围，从而提高知识发现查询的时间开销。

如针对乒乓球运动员的数据库用户提出的查询是"找出那些世界排名进入前16或是世界级比赛的个人最好成绩进入前8名的运动员"，假设我们具有前面例1中给出的领域知识，则按上述处理方法可以去除其中一个查询条件或是用"技术等级标准="国际级运动健将""替换原查询。在例3中可按年龄的四种分类或收入的三种分类将"消费者的消费模式是什么?"的查询细化。在例4中对与收入有关的模式查询，通过投影运算将教育程度和职位以外的、与收入无关的属性去除。

通过以上描述和例子分析不难看出，运用我们的方法可以在查找与某一个查询相关的领域知识时节省大量的时间。

总结 本文介绍了一种在数据仓库的知识发现查询中利用领域知识进行优化处理、缩小搜索范围、使发现的数据更合乎用户要求的方法。下一步，还需要寻找某种启发式的规则以挑选最有用的领域知识，另外如何衡量每条领域知识的质量也是一个有趣的研究课题。

参考文献

1 Hummergren T. 数据仓库技术. 曹增强，王备战，岳晓奎等译. 北京：中国水利水电出版社，1998

2 Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Databases. AAAI/MIT Press, 1991. 229～248

3 Kaufman K A, et al. Mining for Knowledge in Database: Goals and General Description of the INLEN System, Knowledge Discovery in Database. In: G. Piatetsky, W. J. Frawley, eds. AAAI/MIT Press, 1991. 449～462

4 Han J, et al. Data-Driven Discovery of Quantitative Rules in Relational Databases. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(1)

5 Charkravarthy US, et al. Logic-based Approach to Semantic Query Optimization. ACM Transactions on Database Systems, 1990, 15(2): 162～207

6 Yoon S-C, et al. Using Domain Knowledge in Knowledge Discovery. In: Proc. of CIKM'99, ACM Press, 1999. 243～250

(上接第112页)

rule1': $(a,\succ)\wedge(b,\succ)\Rightarrow(e,\prec)\vee(e,\succ)$

accuracy =0.58 OR 0.42 coverage =0.45 OR 0.35

rule2': $(a,\succ)\wedge(b,\prec)\Rightarrow(e,\prec)$

accuracy =1.0 coverage =0.11

rule3': $(a,\prec)\wedge(b,\prec)\Rightarrow(e,\succ)\vee(e,\prec)$

accuracy =0.55 OR 0.45 coverage =0.56 OR 0.43

rule4': $(a,\prec)\wedge(b,\succ)\wedge(b,\prec)\Rightarrow(e,\succ)\vee(e,\prec)$

accuracy =0.8 OR 0.2 coverage =0.09 OR 0.02

结语 本文以采用 Vague 值属性描述的决策系统为例，给出了有效的 Vague 值属性决策系统的知识获取方法。首先根据样本对于决策者需求的适合程度构造 Vague 值之间的一个序关系，将 Vague 决策表转化为二元决策表，然后利用粗糙集理论进行分析并推理出最优规则，最后再将二元决策表的决策规则转化为 Vague 决策表的有序规则。实验分析表明了该方法的有效性。

参考文献

1 Zadeh L A. Fuzzy Sets. Information and Control, 1965, 8(3): 338～353

2 Gau Wen-Lung, Danied J B. Vague Sets. IEEE Transactions on Systems, Man, and Cybernetics, 1993, 23(2): 610～614

3 李凡，徐章艳，饶勇. Vague 集. 计算机科学，2000，27(9)：12～14

4 Chen S M, Tan J M. Handling Multicriteria Fuzzy Decision-making Poblems Based on Vague Set Theory. Fuzzy Sets and Systems, 1994, 67: 163～172

5 Yao Y Y, Sai Y. Mining Ordering rules using rough set theory. Bulletin of International Rough Set Society, 2001, 5: 99～106

6 王国胤. Rough 集理论与知识获取. 西安交大出版社，2001