文章编号 1004-924X(2000)06-0555-04

基于约束的多维数据挖掘技术

王晓升
(长春工程学院计算机系,吉林 长春 130021)

摘要:从特定查询驱动的系统功能出发,讨论了面向多维数据挖掘的重点——五种约束,通过一个数据挖掘查询实例,进一步阐述了这些约束及所产生的关联规则,并用数据挖掘查询语言(DMQL)进行表达。介绍和讨论了关联规则的处理,在挖掘关联规则中,联合使用维/层约束和规则约束,能够带来高效的挖掘过程,一个规则约束如果能被较深地推入到分层结构里,进一步在当前提取层和较深层上挖掘,是非常有价值的。最后给出了运用本文思想的一种联机分析挖掘系统的结构,并对其组成、功能及特点进行了描述。

关 键 词:约束;多维数据;数据仓库;数据挖掘;OLAM
中图分类号:TP311 文献标识码:A

1 引 言

当前的数据挖掘模型比较孤立,即缺少人的导向与控制机制,从而挖掘性能及效率低下。而基于约束的挖掘(用户提供一种指导查找的约束),能够最佳实现人——机劳动的划分,极大地提高挖掘性能。当前数据仓库系统已为多维数据挖掘系统开发提供了丰富的土壤,基于约束的和多维的挖掘技术能够实现特定查询驱动的系统,它比当前孤立的数据挖掘系统能更加有效地开发语义。

2 特定查询驱动的数据挖掘系统

特定查询驱动的数据挖掘系统比较适合用户查询意图,使知识推理过程更加高效。该挖掘系统具有两个能力,第一,它能提供一种与SQL语言相媲美的、面向挖掘的查询语言(DMQL),这种语言能使用户表达:

· 要挖掘的部分数据库(叫做可挖掘的视图),

· 要挖掘的图表/规则的类型,

· 令人满意的图表属性。

这些图表应当不仅包含关于统计属性(象支持度、可信度和相关性)的数字约束,而且包含基于属性领域、种类和聚集上的约束,如"I. type=´snacks´ and avg(I. price) < 100";第二,数据挖掘系统能够通过提供一个精致的挖掘查询优化器,来支持有效处理和挖掘查询的优化,在查询中该优化器利用用户指明的各种约束及其属性,产生与约束条件相匹配的访问图表。

3 约束:特定数据挖掘的重点

通常把约束分成五类:

· 知识约束:指定要挖掘的知识类型,例如:概念描述、联合、分类、预测、聚簇或异态,这些约束不同于其他约束,通常在查询开始被指定。

· 数据约束:指定与挖掘任务相关联的数据集,在查询过程中我们经常用一种类似于SQL的查询和处理方式指定这种约束。

· 维/层约束:限定数据库或数据仓库中要检查的数据维/层,这种约束遵循多维数据库模型,并且体现了多维挖掘的实质。这样,多维挖掘能被平滑地与基于约束的挖掘结构溶合为一体。

· 规则约束:指定对要挖掘的规则的具体约束。

收稿日期:2000-05-23;修订日期:2000-06-15

· 趣性(Interestingness)约束:指定所发现的有关图表度量范围,从统计学的观点看,什么范围是有用的或有趣的。

用实例说明这五种约束,假定存在一个具有4个相互关联的销售多维数据库

· sales (customer- name, item- name, transaction- id),

· lives (customer- name, district, city),

· item (item- name, category, price),和

· transaction (transaction- id, day, month, year),

这里lives, item, 和transaction 是三维表,这些表通过三个关键字段customer- name、item- name 和transaction- id 与sales 表相关联。

"对1998年Vancouver(温哥华市)顾客,查找在同类产品中什么便宜物品(总价格在$100以下)可能促使什么贵重物品(最低价格为$500)的销售",这是一种关联挖掘查询,用数据挖掘查询语言(DMQL)表达,如程序1所示。在程序2中描述了此挖掘查询所产生的关联规则,此规则的意思是:如果Vancouver 的一个顾客购买了Census- CD 和MS Office 97,有68%的可能性他也将购买MS SQL Server,此规则更进一步指示了所有顾客的1.5%都履行这一规则。

在此查询中,知识类型的约束是association (关联),数据约束是lives (C,-, "Vancouver")。此查询涉及到所有的三维数据:lives, item 和transaction。

层的约束较多,对于lives,仅考虑customer- name,由于city= "Vancouver"仅被用于选择;对于item,考虑层items- name 和category,因为它们被用于查询;对于transaction ,仅考虑transaction- id,因为day 和month 不被参考,并且year 仅被用于选择。规则约束包括许多部分字段和语句,例如S. year= 1998, T. year= 1998, I. category= J. category, sum (I. price)< 100, min (J. price)≥500。最后有两个趣性约束(阀值),最小支持度= 0.01 和最小可信度= 0.5。

程序1

```
mine associations as
        lives (C,-, "Vancouver") and sales +
(C,? {I},{S}) = > sales+ (C,? {J},{T})
         from sales
where S. year= 1998 and T. year= 1998
and I. category= J. category
group by C, I. category
having sum (I. price)< 100 and min (J.
price)≥500
with min- support = 0. 01 and min-
confidence= 0. 5
```

程序2

```
lives (C,-, "Vancouver") and sales (C, "
Census- CD", ) and
        sales (C, "MS/Office97",-) = > sales
(C, "MS/SQLServer",-) [0.015; 0.68]
```

知识类型约束和数据约束应用于数据挖掘之前,不与挖掘过程捆绑在一起。运用这两种约束后,挖掘程序可首先挖掘所有可能的规则,然后利用其它三种约束筛选出不满足这些约束的规则,但是这样产生的挖掘将是无效或有时极昂贵的,因此,分析这些约束和挖掘有价值的属性是非常必要的,这就要求把约束推到较深的挖掘过程内部——排除早先的无关项集合、最小化被检查的集合项数。

4 处理关联规则

关联规则以谓词的形式指定集合/子集关系、变量初值或聚集函数,在前面例子中,约束S. year = 1998, T. year = 1998, Icategory = J. category, sum (I. price)< 100, min (J. price)≥500是规则约束。

在经典的关联规则挖掘中,标准演绎算法为发现频繁项提供了一种趣性属性:每当项目集合S 的支持度违反频率约束(其支持度落在一个指定的阀值以下),则所有S 的超集肯定也违反此频率约束,这种属性称为非单调属性[1],与频率约束不同,许多规则约束满足非单调属性。

4.1 非单调约束

一种规则约束,如sum(I. price)≤100 是非单调约束,因为任何使总价格超过100的物品集合不能属于该组成员,若把更多的物品加到此物品集合中可能使总价格变高,违反了这种约束。同理,min (J. price)≥500 和S. year= 1998 也是非单调约束。

诸如此类的约束能够被较深地推进到挖掘过程里,因为若它们不满足于前层挖掘,则它们根本不能满足后层挖掘(因有更多项加到项目集合中)。

相反,一种约束如avg(I. price)≤100 不是非

单调约束, 因为增加更多的物品仍能使物品集合满足此约束。这种约束不能被推到挖掘过程内。

4.2 简要性(succinctness)约束

在每一次 Apriori- style (类型演绎)算法迭代之后, 就产生一次由非单调约束引起的"修剪", 约束的另一个属性—— "简要性(succinctness)"也能提供一种有效"修剪"方法。约束 min (J. price) ≤500 是简要约束, 因为我们能够明确地产生全部满足此约束的物品集, 而不依靠对每件物品的生成与测试算法。特别是, 此集合必须至少包含一个价格在 $ 500 元以下的物品, 因为存在一个严格的"公式"产生所有满足简要约束的集合, 在挖掘过程中没必要重复检查此约束。

相反, 约束 avg (I. price) ≤100 不是简要性约束, 因为平均值从本质上与集合中的全部项有关, 并且此约束不能被减小到只对单个物品的选择。

4.3 维/层约束

数据仓库和多维数据库在语义上被组织为多维和多层的, 在数据挖掘当中应用维/层约束, 能大大减小搜索空间。不同用户可能对不同提取层中物品之间的关联感兴趣, 例如: 有人也许对品种之间的销售感兴趣, 如软饮料和片类物品的销售, 而其他人也许对物品细类的销售感兴趣, 如可乐和傍晚土豆片的销售。有些挖掘方法首先查找什么种类的物品很可能被一起卖出, 然后沿着物品的维下查, 以便在被一起卖出的品种当中发现特殊的物品, 现在已开发出了便于共享层挖掘的有效算法[2], 该算法提供一种属性—— "如果一种高层物品很少出现, 则没有任何一个它的后代(子物品)会频繁出现", 此属性指明了有些非单调属性存在于两层之间并且能够对共享层的挖掘进行探查。

在挖掘关联规则当中, 能够联合使用维/层约束和规则约束。

检查一个规则是否能被较深地推进到分层结构里, 进一步在当前提取层和较深层上挖掘, 这是非常有价值的。

回到先前的实例上, 假设用户可能喜欢在具有同组约束的多个提取层上查找关联规则, 如果频率约束在不同层上保持相同的阀值, 则在高提取层上的非单调性将仍然有效, 象高层物品一样, 根据确定的非单调性和简要性能够观察到较深层的物品。例如: 约束 sum (I. price) ≤100 , 在多个提取层上是非单调的。因此最终获得的规则可以关联在"category (种类) "层或"item- name (物品名) "层上的物品, 在这两者任何一种情形中, 我们能把满足非单调和简要的同组约束较深地推进到关联挖掘过程里。

5 基于约束的多维数据挖掘系统结构

运用本文的思想, 可实现基于约束的多维大型数据库和数据仓库的挖掘。一种数据挖掘系统—— "联机分析数据挖掘系统(OLAM)"结构如图 1 所示, 它由四层构成, 最低层是数据储藏层, 由数据库和数据仓库组成; 第二层是多维数据层, 提供了一个用于联机分析处理(OLAP)和挖掘的多维数据视图; 数据挖掘最重要的是 OLAP/OLAM 层, 由两个引擎模块组成, 一个用于处理, 一个用于挖掘; 最后, 在 OLAP/OLAM 层的上部设置用户接口层, 用户通过这个接口, 构造数据仓库、建立多维数据库、选择要求的数据集、操纵基于约束的交互性的 OLAP 和挖掘、浏览并考查结果。

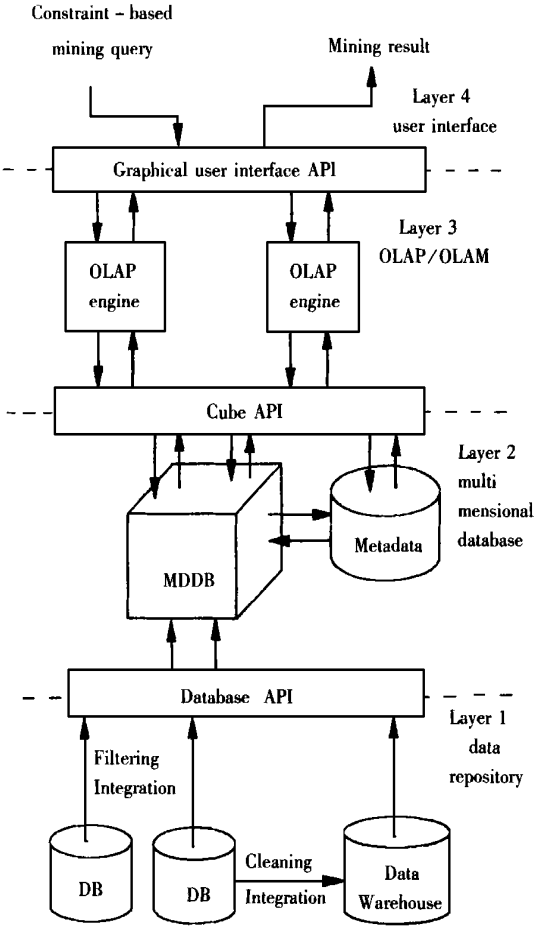

Fig. 1　Online analytical mining (OLAM) architecture

在 OLAM 引擎模块中,嵌入了一种用于约束的演绎算法 CAP(for Constrainted Apriori)。CAP 把规则约束分类为非单调约束和/或简要性约束,以多用户交互方式执行关联规则挖掘和其他相关形式规则的挖掘,把约束较深地推进到挖掘过程里。比那些直截了当的挖掘算法(不把约束较深地推进到挖掘过程里)速度快。

多维的、基于约束的关联挖掘也已结合在 DBMiner 系统(加拿大 Simon Fraser 大学开发的数据仓库与知识发现集成系统)中,其中关联器是主要数据挖掘模块。

这种 OLAM 结构的特点是:第一,适用面宽,具有广泛信息处理的底层结构,整个系统能够围绕关系数据库管理系统和数据仓库系统性地构建;第二,提供一种基于 OLAP 的数据探察分析环境,(1)能够挖掘不同的数据子集,在不同的提取层上通过下查,能够对一个多维数据库和中间挖掘结果进行旋转、过滤、切片、切块;(2)便于联机、交互式的选择数据挖掘功能和门限,提高了数据挖掘效率和灵活性。

6 结 束 语

开发集成的、基于约束的联合挖掘环境,还需要大量的研究。怎样把基于约束的多维数据挖掘应用到其他类型的知识中去,如特征描述、分类、聚类和异态分析,需要进一步的研究和开发。我们相信,新一代数据挖掘系统将会成功地集成传统的数据库管理系统的性能和挖掘性能,把传统的数据库管理与高效的数据分析和挖掘工具紧密地结合在一起。

参考文献:

[1] Ng R, et al. Exploratory Mining and Pruning Optimization of Constrained Association Rules[A]. Proc. ACM SIGMOD Int'l Conf. Management of Data[C]. New York: ACM Press, 1998.

[2] Han J. Discovery of Multiple-Level Association Rules from Large Databases[A]. Proc. 21st Int'l Conf. Very Large Databases[C]. Morgan Kaufman, San Francisco, Calif: 1995. 420-431.

Constraint-based multidimensional data mining technology

WANG Xiao-sheng

(*Dep. Computer, Changchun Institute of Technology, Changchun* 130021, *China*)

Abstract: Based on the functions of ad hoc and query-driven system, the essentials-five constraints for multidimension data mining are discussed. The constraints and the association rules are interpreted using an example of data mining query, and it is expressed with DMQL. Association rules handling is discussed. It produces efficient data mining that dimension/level and rule constraints are used together in mining association rules. It is interesting whether a rule constraint can be pushed deeply into the hierarchy to facilitate mining at the current level of abstraction and deeper levels. In the end, this paper provides a OLAM architecture, its composition, function and benefits are presented.

Key words: constraint; multidimension data; data warehouse; data mining; OLAM

作者简介:王晓升(1964-),男,山东青岛人,1988 年毕业于吉林大学计算机科学系系统结构专业,学士。现工作于长春工程学院,讲师,在读硕士研究生,主研方向为数据库、知识工程。