基于大型数据仓库的数据采掘:研究综述

胡 侃 夏绍玮

(清华大学自动化系 北京 100084)

E-mail:swxia@mail.tsinghua.edu.cn

摘要 本文介绍了数据采掘技术的总体研究情况,包括数据采掘的定义、与其他学科的关系、采掘的主要过程、分类和主要技术手段.作为例子介绍了关联规则采掘的研究,同时介绍了一些原型系统和商业产品以及主要应用领域,指出了数据采掘研究的挑战性以及目前的局限性.结合当前数据仓库的发展,本文探讨了数据仓库环境下数据采掘的特点和潜力.

关键词 数据采掘,数据仓库,信息仓库,数据库知识发现,决策支持.

中图法分类号 TP391

无论是商业企业、科研机构或者政府部门,在过去若干年的时间里都积累了海量的、以不同形式存储的数据资料.由于这些资料十分繁杂,要从中发现有价值的信息或知识,达到为决策服务的目的,成为非常艰巨的任务.数据采掘方法的提出,让人们有能力最终认识数据的真正价值,即蕴藏在数据中的信息和知识.数据采掘(Data Mining),指的是从大型数据库或数据仓库中提取人们感兴趣的知识,这些知识是隐含的、事先未知的潜在有用信息.数据采掘是目前国际上数据库和信息决策领域的最前沿研究方向之一,引起了学术界和工业界的广泛关注.一些国际上高级别的工业研究实验室,例如IBM Almaden和GTE,众多的学术单位,例如UC Berkeley,都在这个领域开展了各种各样的研究计划.研究的主要目标是发展有关的方法论、理论和工具,以支持从大量数据中提取有用的和让人感兴趣的知识和模式.

数据采掘的研究已经和数据仓库的研究结合起来.

数据仓库是近年来才提出的新概念.所谓数据仓库(Data Warehouse)是指这样一种数据的存储地,来自于异地、异构的数据源或数据库的数据经加工后在数据仓库中存储、提取和维护.传统数据库主要面向业务处理,而数据仓库面向复杂数据分析、高层决策支持.数据仓库提供来自种类不同的应用系统的集成化和历史化的数据,为有关部门或企业进行全局范围的战略决策和长期趋势分析提供了有效的支持.数据仓库使用户拥有任意提取数据的自由,而不干扰业务数据库的正常运行.

当前,一些企业已经在传统数据处理方面有了较丰富的经验.他们采用数据仓库希望能从中得到更多好处.例如,以合理的代价取得有效的决策支持、促进企业中业务处理过程的重组、改善并强化对客户的服务、强化企业的资产/负债管理、促进市场优化、加速资金周转、帮助实现企业的规模优化.

数据仓库的产生和发展为数据采掘技术开辟了新的战场,同时也提出了新的要求和挑战.目前的研究还主要着眼于数据仓库的构建和维护的基本理论、方法上,例如数据仓库更新问题的研究,因为这是迈向实用化的第一步的、首要的任务.下一步将把重点放在数据仓库的有效应用研究上.为高级的决策支持服务是数据仓库的最终目的,因此基于数据仓库的数据采掘理论和技术的研究,自然成为信息科学学术界的热点问题.

数据采掘在此前的研究主要是面向一般的数据库系统,并相应地发展起来一系列的理论和方法.现在将平台转移到数据仓库上来,由于数据仓库本身具有与一般数据库不同的特点,这就给基于大型数据仓库的数据采掘的研究提出了新的问题.本文将描述这些新的问题是什么,即基于大型数据仓库的数据采掘的实现方法和研究方向.

本文第1部分介绍数据采掘技术,从它的定义、发展及学科关系、学术挑战性和局限性、分类、主要实现技术和手

* 本文研究得到国家自然科学基金资助.作者**胡侃**,1970年生,博士生,主要研究领域为数据采掘理论与应用,数据仓库理论方法,智能决策系统,管理信息系统(MIS)的应用.**夏绍玮**,女,1932年生,教授,博士导师,主要研究领域为大系统理论与应用,智能决策系统.

本文通讯联系人:夏绍玮,北京 100084,清华大学自动化系

本文1996-11-18收到原稿,1997-04-28收到修改稿

段，到一些原型系统、应用情况和商业产品的介绍；第2部分介绍数据仓库的发展对数据采掘提出了哪些新的要求和挑战，提供了怎样的新的空间；第3部分是小结.

1 数据采掘技术

1.1 数据采掘的定义

数据采掘，英文是Data Mining，中文又译作数据挖掘. 一种比较公认的定义是W. J. Frawley, G. Piatetsky-Shapiro等人提出的[1-3]：数据采掘，就是从大型数据库的数据中提取人们感兴趣的知识. 这些知识是隐含的、事先未知的潜在有用信息，提取的知识表示为概念(Concepts)、规则(Rules)、规律(Regularities)、模式(Patterns)等形式. 这种定义把数据采掘的对象定义为数据库. 而更广义的说法是[4]：数据采掘意味着在一些事实或观察数据的集合中寻找模式的决策支持过程. 数据采掘的对象不仅是数据库，也可以是文件系统，或其它任何组织在一起的数据集合，例如WWW信息资源. 下文我们将提到，最新的对象是数据仓库.

1.2 数据采掘技术的定位

从数据采掘的定义可以看出，作为一个学术领域，数据采掘和数据库知识发现KDD(knowledge discovery in databases)具有很大的重合度，大部分学者认为数据采掘和知识发现是等价的概念，人工智能(AI)领域习惯称KDD，而数据库领域习惯称数据采掘. 也有学者把KDD看作发现知识的完整过程，而数据采掘只是这个过程中的一个部分.[1,5]我们倾向于前一种观点，认为数据采掘从理论和技术上继承了知识发现领域的成果，同时又有着独特的内涵. 数据采掘更着眼于设计高效的算法以达到从巨量数据中发现知识的目的. 数据采掘充分利用了机器学习、人工智能、模糊逻辑、人工神经网络、分形几何的理论和方法.

与数据采掘关系密切的的研究领域包括归纳学习(Inductive Learning)、机器学习(Machine Learning)和统计(Statistics)分析.[6,7]特别是机器学习被认为和数据采掘的关系最密切. 二者的主要区别在于：数据采掘的任务是发现可以理解的知识，而机器学习关心的是提高系统的性能，因此训练神经网络来控制一根倒立棒是一种机器学习过程，但不是数据采掘；数据采掘的对象是大型的数据库，一般来说机器学习处理的数据集要小得多，因此效率问题对数据采掘是至关重要的.

再来看看数据采掘在决策支持空间(Decision-Support Spaces)中处于何种地位. K. Parsaye把决策支持空间从应用层次上分成4个子空间[8]：数据空间(Data Space)、聚合(OLAP)空间(Aggregation Space)、影响空间(Influence Space)和变化空间(Variation Space). 如图1所示.

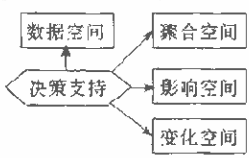

图1 决策支持空间

数据空间处理基于关键字(Key-Based)的决策查询，比如回答"产品123的价格是多少?"这类数据浏览式的查询. 数据空间中最典型的是联机事务处理(OLTP)系统. 对数据空间中数据元素进行聚合运算(如Sum, Average, Max, Min等)形成的空间就是聚合空间. 目前常见的提法有联机分析处理(OLAP)和多维空间(Multidimensional Spaces). 聚合空间处理诸如"某一商场在某月某种商品的销售额是多少?"这种关系到聚合运算的决策查询.

以上两个空间都是处理数值的计算，而影响空间处理逻辑性质的决策支持，比如回答"是什么因素影响在纽约的销售情况?"这样的问题. 这个空间能够提供比其它空间丰富得多的有用信息. 这些信息就是通过数据采掘而得到的.

变化空间负责回答某种变化的过程和速率问题，例如"在过去3个月中的'销售额增长'是怎样变化的?".

在以上4个空间中，数据采掘处于影响空间中. 从中我们可以看出数据采掘在整个决策支持空间中所处的重要地位. 如果一个企业的领导不仅仅满足于一些统计报表，那么数据采掘就是必要的. 它能提供非常重要的决策信息，而这些信息对于决策者可能是完全崭新的. 在当今高度复杂的社会，信息已经成为世上最有价值的商品，而数据采掘所能提供给我们的信息比其它财产更宝贵.

1.3 数据采掘技术的过程及分类

1.3.1 数据采掘的目的

前文提到，数据采掘的任务是从大量数据中发现知识. 那么，什么是知识? 这些知识是以何种形式表达出来? 又是怎样被利用的呢?

知识是人类认识的成果或结晶，包括经验知识和理论知识. 从工程角度定义，知识是有助于解决问题的有格式可复用的信息.

在传统的决策支持系统中，知识库中的知识和规则是由专家或程序人员建立的，是由外部输入的. 而数据采掘的

任务是发现大量数据中尚未被发现的知识,是从系统内部自动获取知识的过程.对于那些决策者明确了解的信息,可以用查询、联机分析处理(OLAP)或其它工具直接获取,比如"列出各子公司在上个月的销售情况".而另外一些隐藏在大量数据中的关系、趋势,即使是管理这些数据的专家也是没有能力发现的.这些信息对于决策可能又是至关重要的.现在就让数据采掘来对付这类问题吧.

数据采掘发现的知识通常是用以下形式表示:

- 概念(Concepts)
- 规则(Rules)
- 规律(Regularities)
- 模式(Patterns)
- 约束(Constraints)
- 可视化(Visualizations)

这些知识可以直接提供给决策者,用以辅助决策过程;或者提供给领域专家,修正专家已有的知识体系;也可以作为新的知识转存到应用系统的知识存储机构中,比如专家系统(Expert System)、规则库(Rule Base)等.

1.3.2　数据采掘的过程

数据采掘过程一般由3个主要的阶段组成,数据准备、采掘操作、结果表达和解释.知识的发现可以描述为这3个阶段的反复过程.[1,9]如图2所示.

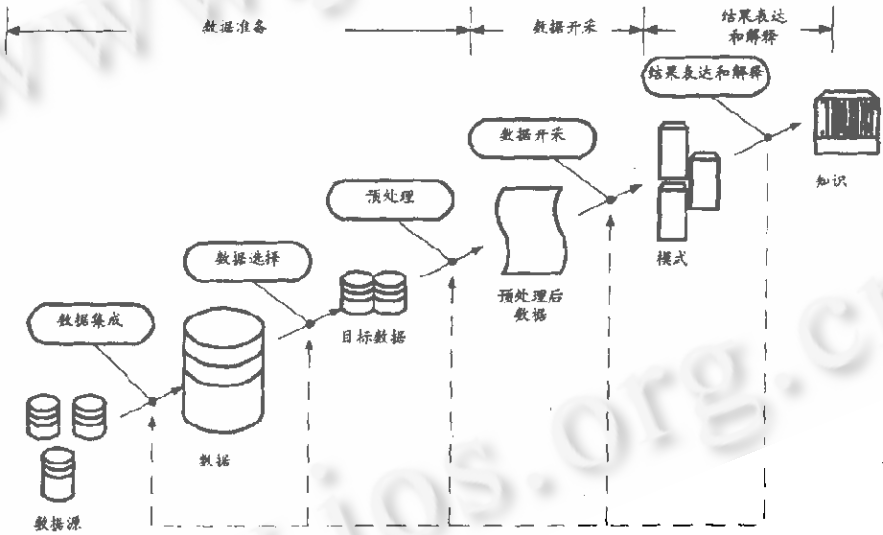

图2　数据采掘的过程

·数据准备　这个阶段又可进一步分成3个子步骤:数据集成、数据选择、数据预处理.数据集成将多文件或多数据库运行环境中的数据进行合并处理.解决语义模糊性、处理数据中的遗漏和清洗脏数据等.数据选择的目的是辨别出需要分析的数据集合,缩小处理范围,提高数据采掘的质量.预处理是为了克服目前数据采掘工具的局限性.

·数据采掘　这个阶段进行实际的采掘操作.包括的要点有:

(1) 要先决定如何产生假设,是让数据采掘系统为用户产生假设,还是用户自己对于数据库中可能包含的知识提出假设.前一种称为发现型(Discovery-Driven)的数据采掘;后一种称为验证型(Verification-Driven)的数据采掘;[10]

(2) 选择合适的工具;

(3) 发掘知识的操作;

(4) 证实发现的知识.

数据采掘操作所采用的主要技术将在1.3.3中提到.

·结果表述和解释　根据最终用户的决策目的对提取的信息进行分析,把最有价值的信息区分出来,并且通过决策支持工具提交给决策者.因此,这一步骤的任务不仅是把结果表达出来(例如采用信息可视化方法),还要对信息进行过滤处理.如果不能令决策者满意,需要重复以上数据采掘的过程.

1.3.3 数据采掘的分类

从不同的视角看,数据采掘技术有几种分类方法[2]:根据发现知识的种类分类;根据采掘的数据库的种类分类和根据采用的技术分类.

• 根据发现知识的种类分类 这种分类方法有:总结(Summarization)规则采掘、特征(Characterization)规则采掘、关联(Association)规则采掘、分类(Classification)规则采掘、聚类(Clustering)规则采掘、趋势(Trend)分析、偏差(Deviation)分析、模式分析(Pattern Analysis)等.如果以采掘知识的抽象层次划分,又有原始层次(Primitive Level)的数据采掘、高层次(High Level)的数据采掘和多层次(Multiple Level)的数据采掘等.

• 根据采掘的数据库分类 数据采掘基于的数据库类型有:关系型(Relational)、事务型(Transactional)、面向对象型(Objected-Oriented)、主动型(Active)、空间型(Spatial)、时间型(Temporal)、文本型(Textual)、多媒体(Multi-Media)、异质(Heterogeneous)数据库和遗留(Legacy)系统等.

• 根据采用的技术分类 最常用的数据采掘技术[11]是:

(1) 人工神经网络:它从结构上模仿生物神经网络,是一种通过训练来学习的非线性预测模型.可以完成分类、聚类、特征采掘等多种数据采掘任务;[12]

(2) 决策树:用树形结构来表示决策集合.这些决策集合通过对数据集的分类产生规则.典型的决策树方法有分类回归树(CART),典型的应用是分类规则的采掘;

(3) 遗传算法:是一种新的优化技术,基于生物进化的概念设计了一系列的过程来达到优化的目的.这些过程有基因组合、交叉、变异和自然选择.为了应用遗传算法,需要把数据采掘任务表达为一种搜索问题而发挥遗传算法的优化搜索能力;

(4) 最近邻技术:这种技术通过 K 个最与之相近的历史记录的组合来辨别新的记录.有时也称这种技术为 K-最近邻方法.这种技术可以用作聚类[13]、偏差分析[14]等采掘任务;

(5) 规则归纳:通过统计方法归纳、提取有价值的 If-Then 规则.规则归纳的技术在数据采掘中被广泛使用,例如关联规则的采掘;

(6) 可视化:采用直观的图形方式将信息模式、数据的关联或趋势呈现给决策者,决策者可以通过可视化技术交互式地分析数据关系.

1.3.4 关联规则(Association Rules)的采掘

关联规则表示数据库中一组对象之间某种关联关系的规则(例如"同时发生"或者"从一个对象可以推出另一个").

在数据采掘的研究领域,对于关联规则采掘的研究开展得比较积极和深入.[15,16,17]介绍一下关联规则采掘的研究情况,可以使人家对于数据采掘的研究有一定的感性认识.

关联规则采掘的一般对象是事务(Transactional)数据库.这种数据库的一个主要应用是零售业,比如超级市场的销售管理.条码技术的发展使得数据的收集变得更容易、更完整,从而存储了大量交易资料.关联规则就是辨别这些交易项目(Item,指交易中的内容.比如,面包、牛奶等都是项目)之间是否存在某种关联关系.例如,关联规则可以表示"购买了项目 A 和 B 的顾客中有 95% 的人又买了 C 和 D".这种关联规则提供的信息可以用作商品销售目录设计、商场布置、生产安排、针对性的市场营销等.

问题是这样描述的[16]:设 $I=\{i_1,i_2,\ldots,i_m\}$ 是 m 个不同项目的集合,D 是针对 I 的交易的集合,每一笔交易包含若干项目 $i_i,i_j,\ldots,i_k\subset I$.关联规则表示为 $X\Rightarrow Y$,其中 $X,Y\subset I$,并且 $X\cap Y=\varnothing$.X 称作规则的前提,Y 是结果.

一般把一些项目的集合称作 itemset.在 itemset 中项目的数量叫作 itemset 的长度.每一个 itemset 都有一个统计的度量称为"支持"(Support):对于 $X\subset I$,如果交易集合 D 中包含 X 的交易个数为 s,则 $support(X)=s$.而一个规则也有衡量的标准称为"置信度"(Confidence),定义为 $Confidence(X\Rightarrow Y)=support(X\cup Y)/support(X)$.

采掘关联规则的问题就是找出这样一些规则,它们的 support 和 confidence 分别大于用户指定的最小 support 和最小 confidence 限度,因此,该问题可以分解成如下两个子问题:

(1) 产生所有 support 大于指定的最小 support 值的 itemset,这些 itemset 称为 large itemset,而其它的称为 small itemset;

(2) 对于每个 large itemset,产生所有比最小 confidence 大的规则,如下:

对于一个 large itemset L 和任何 $S\subset L$,如果 $support(L)/support(L-S)\geq$ minimum-confidence,那么规则 $L-S\Rightarrow S$ 就是有效规则.例如,设数据库中有 4 次交易为 $T_1=\{A,B,C\}$,$T_2=\{A,B,D\}$,$T_3=\{A,D,E\}$,$T_4=\{A,B,D\}$.设

最小support为0.5,最小confidence为0.8,那么large itemsets为$\{A\},\{B\},\{D\},\{AB\},\{AD\},\{ABD\}$,而有效的规则为$B\Rightarrow A$和$D\Rightarrow A$.

这个问题的主要挑战性在于数据量巨大,因此算法的效率是关键.目前研究的重点在第1步,即找出large itemset,因为第2步相对来说是容易的.围绕这个问题,R. Agrawal等在1994年提出Apriori算法[18],基本的办法是重复扫描数据库,在第K次扫描产生出长度为K的large itemset,称为L_K.而在第$K+1$次扫描时,只考虑由L_K中的K-itemset产生的长度为$K+1$的备选集C_{K+1}.因此除了第1次扫描以外,每一次扫描要考察的并不是所有项目的组合,而只是其中的一部分,即备选集C_K.围绕着怎样精简备选集C_K的大小(特别是C_2的选择会大大影响采掘的性能)和减少对数据库的扫描遍数,又有不少改进方法.例如R. Agrawal提出的AprioriHybrid算法,Park等人提出的DHP算法[19],使用哈希(Hashing)技术有效地改进了备选集C_K的产生过程.Savasere等在1995年提出了一种把数据库分割(Partition)处理的算法[20],降低了采掘过程中I/O操作的次数,减轻了CPU的负担.H. Toivonen使用抽样(Sampling)的方法可以用较小的代价从大型数据库中找出关联规则.[21]更进一步的研究涉及分布和并行环境下采掘关联规则[22,23],例如,D. W. Cheung等提出了一种关联规则的快速分布式采掘算法(FDM).[24]

1.4 数据采掘研究和应用的挑战性

我们列出目前数据采掘研究和应用所面临的主要挑战.[1,6,25]数据采掘技术的研究还很不成熟,其应用还有较大的局限性.正是这些局限性,促使数据采掘研究进一步发展.

· 采掘的对象:更大型的数据库、更高的维数和属性之间更复杂的关系.　数据采掘要处理的数据量通常是十分巨大的.成百上千的表,上百万条记录,数据库容量达到若干GB(10^9)字节,甚至TB(10^{12})字节.更多的属性意味着高维的搜索空间,从而导致组合爆炸.属性值之间的关系变得更加复杂,比如表现为层次结构.这些因素使得搜索知识代价极高.目前的研究发展到用并行处理或抽样(Sampling)[21]的方法处理大规模数据,获得了较高的计算效率.根据问题的定义或相关知识可以选择出需要的属性从而降低维数.而处理属性之间的复杂关系,往往需要一些背景知识,比如不同层次的概念所构成的概念树.[26,27]

· 多种形式的输入数据　目前数据采掘工具能处理的数据形式有限.一般可以处理数值型的结构化数据,但大多不能对文本、图形、数学公式、图象或WWW资源等这些半结构、无结构的数据形式进行采掘操作.另外的挑战是数据本身存在缺损或噪声,特别是在商业数据库中.

· 用户参与和领域知识　有效的决策过程往往需要多次交互和多次反复.目前的数据采掘系统或工具很少能真正做到让用户参与到采掘过程中.用户的背景知识和指导作用可以加快采掘的进程,并且保证发现的知识的有效性.将相关领域的知识融入数据采掘系统中是一个重要但没有很好解决的问题.

· 证实(Validation)技术的局限　数据采掘使用特定的分析方法或逻辑形式发现知识,比如归纳或演绎.但是系统可能没有能力去交互证实(Cross-Validation)发现的知识,使得发现的知识没有普适性而不能成为有用的知识.另一种情况是待采掘的数据本身就可能是存在错误的,数据采掘技术必须具有足够的鲁棒性,能够确定结论具有何种程度的有效性.同样,还应该可以解释为什么存在与那些普适规则不一致的例外情况.

· 知识的表达和解释机制　许多应用中重要的是用户能够理解发现的知识.这要求知识的表达不仅限于数字或符合,而是更容易理解的方式,如图形、自然语言和可视化技术等.数据采掘系统指出它发现了新的知识,并且能以关系、规则和概念等形式把知识表达出来,但是用户不知道这种发现的基本原理.只有当数据采掘系统能提供更好的解释机制,用户才能更有效地评价这些知识,并且区分出哪些是真正有用的知识,哪些只是常识性的知识或异常情况.

· 知识的维护和更新　新的数据积累可能导致以前发现的知识失效.这些知识需要动态维护和及时更新.目前研究采用增量更新的方法来维护已有的知识,比如D. W. Cheung等提出了维护关联规则的增量算法.[28]

· 支持的局限、与其他系统的集成　目前的数据采掘系统尚不能支持多种平台.一些产品是基于PC的,一些是面向大型主机系统的,还有一些是面向客户机/服务器环境的.有的系统对于数据库中包含的域或记录是有限制的,例如要求数据文件为特定的大小,或者转化为特定的数据库管理系统(DBMS)识别的格式.但是,数据重定义的费用可能是十分昂贵的.另外的挑战是数据采掘系统和其他决策支持系统的有机集成,特别是和一些用户已经熟悉的系统结合在一起,这对于系统充分发挥作用是非常重要的.

1.5 一些原型系统和商业产品的介绍

数据采掘技术的强大生命力还体现在一大批有关的原型系统的建立上.一些著名的公司或研究机构纷纷推出自己的数据采掘商业产品.

1.5.1 数据采掘系统或系统原型[21]

· Quest 由IBM Almaden研究所的R. Agrawal等人研究开发,面向大型数据库,包括采掘关联规则、分类规则、序列模式和相似序列等;

· DBMiner 由加拿大Simon Fraser大学的J. Han等人研究开发.[29]这是一个交互式的、多层次采掘系统,主要采掘特征规则、分类规则、关联规则和预测等;

· KDW+ 由GTE的Piatetsky-Shapiro等人研究开发,采用多策略、统计方法等;

· Explora 由GMD的Klosgen研究开发,这是一个多模式、多策略发现的辅助发现系统;

· SKICAT 由日本的Kayyad等人研究开发,用于大规模天空测量数据分析;

· IMACS 由AT&T的Brachman等人研究开发,用于知识库的建构;

· INLEN 由乔治梅森大学的Michalski等人研究开发,这是一个多种学习策略集成的系统.

1.5.2 主要商业产品介绍

数据采掘技术的潜在应用是十分广泛的,从政府管理决策、商业经营、科学研究和工业企业决策支持等各个领域都可以找到数据采掘技术的用武之地.下面我们举出目前开展得比较活跃的数据采掘的应用方向[25,30]:

市场营销:预测顾客的购买行为;划分顾客群体.

银行业:侦测信用卡的欺诈行为;客户信誉分析.

生产、销售和零售业:预测销售额;决定库存量;批发点分布的规划、调度.

制造:预测机器故障;发掘影响生产能力的关键因素.

经纪业和安全交易:预测债券价格的变化;预报股票价格升降;决定交易的最佳时刻.

保险业:分析决定医疗保险额的主要因素;预测顾客保险的模式.

计算机硬件和软件:监测磁盘驱动故障;估计潜在的安全漏洞.

政府和防卫:估计军事装备转移的成本;预测资源的消耗;评估军事战略.

医药:验证药物的治疗机理;医药公司划分出哪部分大夫会再次购买某类药品.

交通:航空公司可以根据历史资料寻找乘客的旅行模式,改进航线的设置.

电信:电话公司评估哪一类客户会在短期内转向别的公司或其它服务项目,从而限制对这部分客户的广告投入.

公司经营管理:评价客户信誉;评估部门业绩;评估员工业绩;监测子公司或部门财务舞弊行为.

正是数据采掘技术巨大的商业潜力,吸引了众多公司从事数据采掘系统的研究和开发,而且有的已经商品化.在不久以前大部分数据采掘工具还只能为专门技术人员所操纵.但是,现在有更多的公司提供了更高级的数据采掘系统,使得非专业人士也能使用.表1归纳了一些主要的数据采掘产品.[4]

表1 数据采掘工具

产品	供应商	技术
Clementine	Integral Solutions	规则归纳
Darwin	Thinking Machines Corp.	神经网络、遗传算法等
Database mining Workstation	HNC Software Inc.	神经网络
DataEngine	MIT GmbH	模糊逻辑、神经网络、信号处理
IBM Intelligent Miner	IBM corp.	多种技术
F-DBMS	Cross/Z International Inc.	分数维
IDIS	Information Discovery Inc.	规则发现
Information Harvester	Information Harvesting	模糊专家系统规则
Knowledge Seeker	Angoss Software Int'l Ltd.	规则发现、决策树
NeuralWare	NeuralWare Inc.	神经网络
Prison	Nestor Inc.	神经网络
ReMind	Cognitive Systems	基于实例的推理、归纳逻辑

2 基于数据仓库的数据采掘技术

在1996年,信息技术领域最广泛的两个话题是Internet和数据仓库.[31~34]数据仓库的概念一出现,立即引起学术界和工业界的极大关注,厂商们争相展示出产品,虽然绝大多数是在其原有产品上稍加改进或附加新的功能.研究领域不甘落后,掀起了信息领域研究的一股热潮.

数据仓库作为一种新型的数据的存储地,为数据采掘提供了新的支持平台.[9,35,36]可以预见,数据仓库以其内在的对决策的支持能力,将会成为数据采掘的主战场.数据仓库的发展不仅仅是为数据采掘开辟了新的空间,更对数据采掘技术提出了新的要求.

2.1 数据仓库提出的背景

从计算机应用初期的电子数据处理(EDP)到今天的执行信息系统和决策支持系统(EIS/DSS),都始终伴随着对数据仓库的探求.[37,38]从企业发展的角度来看,在不同历史阶段企业内部许多部门建立了各自的信息处理系统.这些系统之间相互隔离,结构各异.企业的决策者很难得到企业全局的决策信息.对这个问题的探求曾导致了多数据库系统(Multidatabases)的研究.企业的高层管理者还需要使用数据(历史的、现在的)进行各种复杂分析,如长期趋势分析和数据采掘等,以支持决策.从大量的历史数据中获取信息,要求系统保存大量的历史数据,而且还要进行复杂的分析处理(每次处理涉及大量数据).这些应用对于业务处理频繁的数据库系统而言,将成为沉重的负担.数据仓库面向复杂的数据分析以支持决策过程,而且可以集成企业范围内的数据.它把支持决策分析的数据事先收集、归纳、处理,使企业的业务操作环境和信息分析环境分离,从而有效地为决策提供实时的信息服务.

确切而言,我们现在称之为"数据仓库"的这一技术,发轫于80年代初W. H. Inmon的研究,即在其"记录系统"、"本原数据"(Atomic Data)、"决策支持数据库"等专题中.Devlin和Murphy在1988年曾披露过IBM的一项内部研究计划,这个计划的目的是构造一种"以关系数据库为基础的、公司数据的集成化仓储".这种仓储的使用者不是信息技术的专业人员,而是各级决策者,他们将使用"一组相容的工具"从仓库中提取有助于决策的信息,这组工具应当得到"业务数据字典"的有效支持,这个"业务数据字典"描述了决策者的可用信息.3年后,也就是1991年,IBM正式宣布了它的"数据仓库构架"INDEPTH,在信息产业界引起很大轰动.[38]

目前,由于相关领域的技术发展及相互协同,已使数据仓库成为一项可能的实用技术,例如,支持并行处理的分布式数据库管理系统(DBMS);为DSS准备数据的转换及加载工具;支持异构环境下分布式数据处理的客户机/服务器技术以及实现桌面信息系统集成的方案和工具等.

2.2 数据仓库的定义

斯坦福大学数据仓库研究小组是这样定义数据仓库[39]的:"数据仓库是集成信息的存储中心,这些信息可用于查询或分析".

W. H. Inmon曾对数据仓库作了这样的描述:"数据仓库是90年代信息技术构架的新焦点,它提供集成化的和历史化的数据;它集成种类不同的应用系统;数据仓库从事物发展和历史的角度来组织和存储数据,以供信息化和分析处理之用."由于Inmon本人在数据仓库发展中的作用,他的上述描述在技术性的文献中不断被引用,相对地成了一种权威的定义.

2.3 数据仓库的体系结构

数据仓库既是一种结构和富有哲理性的方法,也是一种技术.数据和信息从不同的数据源提取出来,然后把这些数据转换成公共的数据模型并且和仓库中已有的数据集成在一起.当用户向仓库进行查询时,需要的信息已经准备好了,数据冲突、表达不一致等问题已经得到了解决.这使得决策查询更容易、更有效.

作为一个系统,数据仓库至少应包括3个基本的功能部分[40]:

· 数据获取　这个部分负责从外部数据源获取数据.数据被区分出来,进行拷贝或重新定义格式等处理后,准备载入数据仓库.

· 数据存储和管理　这个部分负责数据仓库的内部维护和管理,提供的服务包括数据存储的组织、数据的维护、数据的分发、数据仓库的例行维护等.

· 信息访问　信息访问部分属于数据仓库的前端,面向不同种类的最终用户,这里主要由桌面系统的各种工具组成.数据仓库的最终用户在这里提取信息、分析数据集、实施决策,从而可望取得竞争优势.进行数据访问的软件工具,主要是查询生成工具、多维分析工具和数据采掘工具等.这里也是工具制造商们竞相争夺的地段.新的发展趋势是把信息访问工具紧密集成到数据仓库系统中.

虽然制造商们迫不及待地推出支持数据仓库的商业产品,但大多数只是原有技术或系统的稍加改进,这些系统都有局限性.一个真正通用、有效、灵活的数据仓库体系结构的建立被认为是十分必要的.这种体系结构的出现有赖于有关技术的进步.

在众多高校和研究所对数据仓库的研究中,斯坦福大学的数据仓库计划(Whips)处于领先地位.他们提出了一种有普遍代表性的数据仓库体系结构,并且围绕这个体系结构的各个环节,开展了深入的研究工作.

图3说明了Whips的数据仓库系统的基本体系结构.[39]

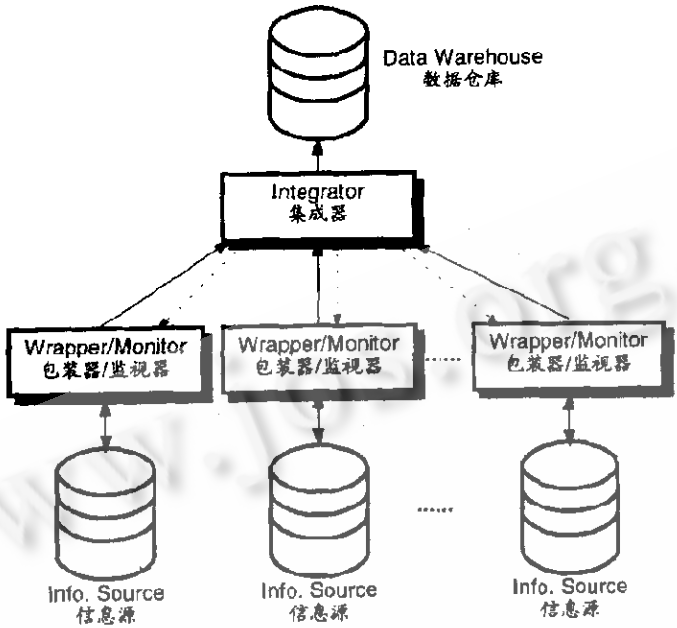

图3 数据仓库系统的基本体系结构

图中底部是信息源.这不但指那些常见的数据库,也包括文件、HTML文件、知识库、遗留(Legacy)系统等各种信息源.

连接着每个信息源的是包装器/监视器(Wrapper/Monitor).该模块的包装器(Wrapper)部分负责把信息从原信息源的数据格式转换成仓库系统使用的数据格式和数据模型,而监视器(Monitor)部分负责自动监测信息源中数据的变化并把这些变化上报给集成器(Integrator).

每当有新的信息源挂上仓库系统,或者每当信息源中相关信息发生变化时,这些新的或改变的数据就传送给集成器(Integrator),集成器对这些信息进行过滤、总结,或者和其它信息源的信息进行合并处理,再安置在仓库中.为了把新信息准确地集成到仓库中,集成器可能还要从原来或相关的其它信息源中获取进一步的信息.图中向下的虚箭头表示这种行为.

数据仓库本身可以使用现在流行的,或者是特别设计的数据库管理系统(DBMS).虽然图中表示的是单一、集中的仓库,但仓库能够以分布式数据库系统来实现.实际上,为了获得期望的性能,常常需要数据的并行和分布处理.

以上我们描述的数据仓库的结构和基本功能在通用性方面远远超过了当前大多数的商业系统.目前的一般系统都假定信息源和仓库属于单一的数据模型(一般是关系型的),信息从信息源到仓库的传送按批处理执行,可能还是离线的,更没有从集成器到信息源的逆向查询过程.

2.4 数据仓库中的数据采掘

作为数据仓库系统三要素之一的信息访问部分,是最终用户赖以从数据仓库中提取信息、分析数据、实施决策的必经途径.我们知道数据仓库建立的最终目的是面向高层的决策支持.虽然目前的研究还集中在数据仓库的创建和维护功能上面,因为这是迈向实用化的第一步的、首要的任务.可以预见,当数据仓库创建和维护的理论和方法基本成熟以后,研究的重心将放在数据仓库的有效使用上,即研究如何有效地进行决策支持.实际上许多商家已经在这个领域展开了竞争.图4描述了数据仓库环境中数据采掘的体系结构.

从前面分析的决策支持的4个空间(数据空间、聚合空间、影响空间、变化空间)来看,影响空间中能够提供比其它空间更丰富的有用信息,而这些信息是通过数据采掘才能获得的.

那么在数据仓库中进行数据采掘有什么新的特点呢?这些新的特点,不论是使数据采掘更容易或更困难,都体现在数据仓库与目前数据采掘的主要对象——数据库的本质区别上面.

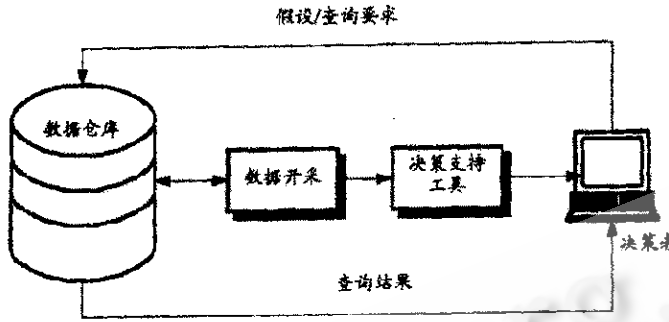

图 4　数据仓库环境下的数据采掘

• 规模　数据仓库中集成和存储着来自若干分布、异质的信息源的数据.这些信息源本身就可能是一个规模庞大的数据库,可以想见数据仓库会有比一般数据库系统更大的数据规模.目前数据仓库的规模一般都超过50GB,将来会更大.如何从如此巨量的数据中有效地提取有用的信息,需要各方面技术的进步.从当前发展来看,支持并行处理的分布式DBMS、具有大规模并行处理(MPP)能力的计算机、超大规模的存储机构等技术的发展和协同将使数据仓库走向实用.但要进行数据采掘我们还必须发展更有效、更快速的算法,因为我们面对的是更巨大的数据的"山脉",要从中找到有价值的矿藏必然难度更高.

• 历史数据　传统的数据库系统为了获得最大的执行效率,往往存储尽可能少的数据量.因为拥有的数据越多,数据组织、重构、浏览、索引和监控的难度越大.传统数据库系统在"时间"轴上的长度很有限.比较而言,数据仓库的根本特征之一就是进行长时间的历史数据存储——从5~10年,这使得我们可以进行数据长期趋势的分析.数据仓库为决策者的长期决策行为提供了独一无二的支持.数据仓库中数据在时间轴上大的纵深性是数据采掘不能回避的又一个新难点.

• 数据集成和综合性　从一个企业的角度看,数据仓库集成了企业内各部门的全面的、综合的数据.数据采掘要面对的是关系更复杂的企业全局模式的知识发现.从这一点上讲,基于数据仓库的数据采掘能更好地满足高层战略决策的要求.而且,数据仓库机制大大降低了数据采掘的障碍.一般进行数据采掘要花大量的力量在数据准备阶段,而在数据仓库中数据已经被充分收集起来,进行了整理、合并,并且有些还进行了初步的分析处理.这样,注意力更集中于数据采掘的核心处理阶段.另外,数据仓库中对数据不同粒度的集成和综合,更有效地支持了多层次、多种知识的采掘.

• 查询支持　数据仓库面向决策支持.数据仓库的体系结构努力保证查询(Query)和分析的实时性.而一般的联机事务处理(OLTP)系统主要要求更新(Update)的实时性,对查询的性能要求相对较弱.一般的数据仓库设计成只读方式,最终用户不能直接更新数据仓库.数据更新由专门的一套机制保证,通常由系统自动更新和管理员控制来协同完成.数据仓库对查询的强大支持使数据采掘效率更高,采掘过程可以做到实时交互,使决策者的思维保持连续,有可能采掘出更深入、更有价值的知识.

从以上分析可以看出,数据仓库在纵向和横向都为数据采掘提供了更广阔的活动空间.数据仓库完成了数据的收集、集成、存储、管理等工作,数据采掘面对的是经初步加工的数据,使得数据采掘能更专注于知识的发现;另一方面,由于数据仓库所具有的新的特点,又对数据采掘技术提出了更高的要求.可以说,数据采掘技术要充分发挥潜力,就必须和数据仓库的发展结合起来.

3 小　结

大量数据的产生和收集导致了信息爆炸.现代社会的竞争趋势要求对这些数据进行实时的和深层次的分析.虽然现在有了更强大的存储和检索系统,但是使用者发现在分析和使用所拥有的信息方面变得越来越困难.数据仓库提供了容纳大量信息的场所,但只有和数据采掘技术的应用结合起来才能最终解决用户的困惑,使用户能够从大量繁杂的数据中找出真正有价值的信息和知识.随着数据采掘和数据仓库集成的进一步深化,必然给用户带来更大的利益.

参考文献

1 Usama M Fayyad, Gregory Piatetsky-Shapiro *et al*. Advances in knowledge discovery and data mining. California: AAAI/MIT Press, 1996

2 Han J. Conference tutorial notes: data mining techniques. In: Proceedings of ACM SIGMOD International Conference'96 on Management of Data (SIGMOD'96). Montreal, Canada, June 1996

3 Piatetsky-Shapiro G, Fayyad U, Smith P. From data mining to knowledge discovery: an overview. In: Fayyad U M, Piatetsky-Shapiro C *et al* eds. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. 1～35

4 Jay-Louise Weldon. Data mining and visualization. Database Programming and Design, May 1996,9(5):21～24

5 Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, Dec. 1993,5(6):914～925

6 Heikki Mannila. Data mining: machine learning, statistics and databases. In: Proceedings of the 8th International Conference on Scientific and Statistical Database Management, Stockholm, June 18～20, 1996. 1～8

7 Ruth Dilly. Data mining: an introduction student notes. WWW files, Queens University Belfast, Dec. 1995 (http://www-pcc.qub.ac.uk/tec/courses/datamining/stu_notes/m_book 1.html)

8 Kamran Parsaye. Surveying decision support. Database Programming and Design. Apr. 1996,9(4):27～33

9 Data mining: extending the information warehouse framework. White Paper, IBM Corporation, 1996 (http://www.almaden.ibm.com/cs/quest/papers/whitepaper.html)

10 IBM's Data Mining Technology. WWW Book, IBM Corporation, 1996 (http://booksrv2.raleigh.ibm.com:80/cgi-bin/bookmgr/bookmgr.cmd/BOOKS/datamine/CCONTENTS)

11 Data mining: discovering hidden value in your data warehouse. Pilot Software. WWW files, 1996. (http://www.pilotsw.com/dmpaper/dmindex.html)

12 Lu Hongjun, Rudy Setiono, Liu Huan. Effective data mining using neural networks. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):957～961

13 Fisher D. Optimization and simplication of hierarchical clustering. In: Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD'95). Montreal, Canada, Aug. 1995. 118～123

14 Arning A, Agrawal R, Raghavan P. A linear method for deviation detection in large databases. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96). Portlan, Oregon, Aug. 1996

15 Mannila H, Toivonen H, Inkeri Verkamo A. Efficient algorithms for discovering association rules. In: Proceedings of AAAI Workshop on Knowledge Discovery in Database. July 1994. 181～192

16 Srikant R, Agrawal R. Mining generalized association rules. In: Proceedings of the 21th International Conference on Very Large Data Bases. Sept. 1995. 407～419

17 Hannu Toivonen, Mika Klemettinen, Pirjo Ronkaine *et al*. Pruning and grouping discovered association rules. In: MLnet Workshop on Statistics, Machine Learning and Discovery in Databases. Heraklion, Crete, Greece, April 1995

18 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases. Santiago, Chile, Sept. 1994

19 Park J S *et al*. An effective hash-based algorithm for mining association rules. In: Proceedings of the ACM SIGMOD. May 1995. 175～186

20 Ashok Savasere, Omiecinski E, Navathe S. An efficient algorithm for mining association rules in large databases. In: Proceeding of the 21th VLDB Conference. Zürich, Switzerland, 1995. 432～444

21 Hannu Toivonen. Sampling large databases for association rules. In: Proceedings of the 22th International Conference on Very Large Databases (VLDB'96). Bombay, India. Morgan Kaufmann, September 1996. 134～145

22 Agrawal R, Shafer J C. Parallel mining of association rules. IEEE Transacations on Knowledge and Data Engineering, 1996, 8(6):962～969

23 Park J S, Chen M S, Yu P S. Efficient parallel data mining for association rules. In: Proceedings of the 4th International Conference on Information and Knowledge Management, Nov. 29～Dec. 3, 1995

24 Cheung D, Han J, Ng V T *et al*. A fast distributed algorithm for mining association rules. In: Proceedings of the International Conference'96 on Parallel and Distributed Information Systems (PDIS'96). Miami Beach, Florida, USA, Dec. 1996

25 Grupe F H, Owrang M M. Data base mining: discovering new knowledge and competitive advantage. Information Systems Management, Fall 1995,12(3):26～31

26 Han J. Mining knowledge at multiple concept levels. In: Proceedings of the 4th International Conference on Information and Knowledge Management (CIKM'95). Baltimore, Maryland, Nov. 1995. 19～24

27 Han J, Fu Y. Discovery of multiple-level association rules from large databases. In: Proceedings of 1995 International Conference on Very Large Data Bases (VLDB'95). Züich. Switzerland, Sept. 1995. 420～431

28 Cheung D, Han J, Ng V *et al*. Maintenance of discovered association rules in large databases: an incremental updating technique. In: Proceedings of the International Conference'96 on Data Engineering (ICDE'96). New Orleans, Louisiana, USA, Feb. 1996

29 Jiawei Han, Fu Yongjian, Wang Wei *et al*. DBMiner: a system for mining knowledge in large relational databases. In: Proceedings of the International Conference'96 on Data Mining and Knowledge Discovery (KDD'96). Portland, Oregon, August 1996. 250～255

30 Lisa Lewinson. Data mining: tapping into the mother lode. Database Programming and Design, Feb. 1994,7(2):50～56

31 Inmon W H. Dawn of a new age: why everyone is building a data warehouse. Database Programming and Design, Dec. 1992, 5(12):76～77

32 Colin White. Data warehouse: what's in a name? Database Programming and Design, Mar. 1996,9(3):53～54

33 Jia Hong. The construction and application of a data warehouse. China Computer Users, 1996,20:6～11

34 Wang Shan, Luo Li. From a database to data warehouse. Computer World, 1996,7(28):101～103

35 Widom J. Research problems in data warehousing. In: Proceedings of the 4th International Conference on Information and Knowledge Management. Baltimore, Maryland, Nov. 1995. 25～30

36 Wiener J L, Gupta H, Labio W J *et al*. A system prototype for warehouse view maintenance. In: Proceedings of the ACM Workshop on Materialized Views: Techniques and Applications. Montreal, Canada, June 1996. 26～33

37 Inmon W H. EIS and the data warehouse: a simple approach to building an foundation for EIS. Database Programming and Design, Nov. 1992,5(11):70～73

38 Guo Yi-bin. The basic concepts and current development on data warehousing technique. PC World China, 1996.4(total 114): 26～31

39 Hammer J, Garcia-Molina H, Widom J *et al*. The stanford data warehousing project. IEEE Data Engineering Bulletin (Special Issue on Materialized Views and Data Warehousing), June 1995,18(2):41～48

40 Ma Xiaoqiang. Surv on data warehousing. Technical Report, HongKong University, 1996

Large Data Warehouse-based Data Mining: a Survey

HU Kan　XIA Shao-wei

(Department of Automation　Tsinghua University　Beijing　100084)

Abstract　In this paper, the data mining techniques are introduced broadly including its definition, the relationships with other academic fields, the imperative processes and its classifications. The principal techniques used in the data mining are surveyed also. As an example, the studies on the mining association rules are illustrated. Some data mining prototypes and commercial systems are listed in this paper. Several limitations of the data mining are discussed as well as the research and application challenges for it. Due to the development of the data warehouse up to now, the features and potentialities of the data mining based on the data warehouse are discussed also.

Key words　Data mining, data warehouse, information warehouse, knowledge discovery in databases, decision support.

Class number　TP391