计算机科学 2006Vol. 33№. 12

商务网站数据挖掘方法

黄 斐
（苏州大学计算机科学与技术学院 苏州 215006）

摘 要 本文介绍了商务网站数据挖掘基本原理和方法，结合数据仓库研究，论述了多维数据模型，关联规则挖掘算法，商务网站数据挖掘的基本过程，以及联机分析处理。

关键词 数据仓库，挖掘算法，联机分析处理

The Commercial Website Data Mining Method

HUANG Fei
(Computer Science&Technology School of Suzhou Univ., Suzhou 215006)

Abstract This paper introduces the commercial website data mining basic principle and the method. Union data warehouse studies, details account on the multi-dimensional data model, connection rule mining algorithm, commercial website data mining unit process, as well as online analytical processing.

Keywords Data warehouse, Mining algorithm, Online analytical processing

商务网站数据挖掘，就是采集各种商务数据，通过对商务数据的分析、监控商务网站的运行状况，充分了解客户的喜好、购买模式，设计出满足不同客户群体需要的个性化网站，进而增强电子商务企业的竞争力。

利用商务数据挖掘技术，对电子商务的海量数据进行分析，并依据分析结果做出正确的决策，随时调整经营策略，适应市场的需求，已经成为众多商务网站的选择。通过数据挖掘从大量数据中寻找有用的信息，采用切片旋转、钻取等方式，对同一层面的数据从不同的角度进行分析，转换分析问题的角度及对数据的细节从不同的层次进行观察，从而为企业的竞争构筑信息与决策的优势，为企业带来显著的经济效益。

1 数据仓库模型

数据挖掘的数据源可以是任何信息源，包括关系数据库、数据仓库、事务数据库、对象-关系数据库、空间数据库、文本、多媒体数据、时间序列数据和 Web 数据等。商务网站数据源主要来自两个方面，一是客户的背景信息，此部分信息主要来自于客户的登记表，另一部分数据主要来自浏览者的点击流，此部分数据主要用于考察客户的行为表现。构建基于数据仓库的控制系统，就可以进行商务网站数据挖掘。

数据仓库是一个面向主题的、集成的、非易失的且随时间变化的数据集合，用来支持管理人员的决策。数据仓库的本质是一个建立在数据库系统之上的数据管理系统，其数据来自若干分布的、异构的数据源，数据仓库不仅具有数据库系统的特点，还具有数据面向主题、数据集成、数据不可更新、数据随时间不断变化的特点。

数据仓库模型通常采用 3 层结构，即数据准备层、数据管理层和数据表现层。数据准备层实现多种数据源数据的提取、清理、转换和集成，数据管理层通常采用关系模型、多维模型、混合模型等，数据表现层通常利用分析、查询、报告、挖掘、图形接口等前端工具向决策者提供分析的结果。

设计数据仓库通常采用多维数据模型，在多维数据模型中，数据集合被视为多维空间中的点集，数据集合的属性分为维和度两类。维是人们观察事物的角度，如日期、地区等。度是分析目标的数值指标，如销售额、利润等。通常认为度是维的函数，即维决定度，且二者之间的关系是静态的。根据实际应用中出现的问题，可以在维和度之间进行转化，实现维和度的对称处理，即维降格为度或度升级为维。多维数据模型的数据组织形式通常是数据立方体，每个数据立方体都由多个维组成，根据数据粒度选择的不同，每个维有一个或多个层。计算了 n 维的数据立方体可能产生的方体总数是：

$$T=\prod_{i=1}^{n}(L_i+1)$$

其中，L_i 是维的层次数。维的存储将会影响数据仓库的性能，在数据仓库的数据模型中，数据的逻辑组织方式则以星型模式、雪花模式和星型-雪花模式存在。设计时通过构建维表和事实表来完成。实际上，仍然可以用表格来表示二维数据库，其情形就如表 1 所示。这个表中不但记录了每个交点的内容，也记录了对每个纬度的统计值以及总值，可以认为是一个物理的存储机制。

表 1 二维数据库的表格表示

时间:y	客户:x	销售额
1 月份	E	200
2 月份	A	300
…	…	…

理解了二维数据库，就可以引伸至三维、四维或更多维数据库。上面的例子中，如果在时间和客户维的基础上再加上产品维，就形成了如图 1 所示的三维数据库的视图。

数据仓库建库目的不同，在设计上是有区别的。在性能上，数据仓库要求快速查询和统计计算。为了减少表与表之间的复杂关系和提高查询运算速度，商务网站的数据仓库可以采用星型关系模型。以一个核心的主题数据表（称事实表）为中心，其他关系表格（称为维表）只同核心表发生关系，维表之间没有直接的关系。星型数据关系模型可以大大提高查询

速度，这相对于单纯从一个很大的数据表中利用单个 SQL 语句查询来说显然要有效得多。

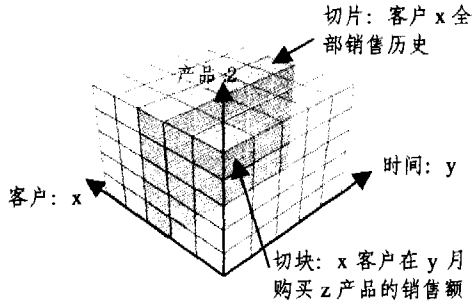

图 1　三维数据库视图

典型商务网站数据仓库是基于关系型数据库管理系统的设计，精简并整合了企业多个数据源的原始数据，其主要用途是为数据访问和企业分析决策之用。数据库仓库元数据、交易型数据分开存放，它们是工具集的数据源。典型商务网站数据仓库结构如图 2 所示。

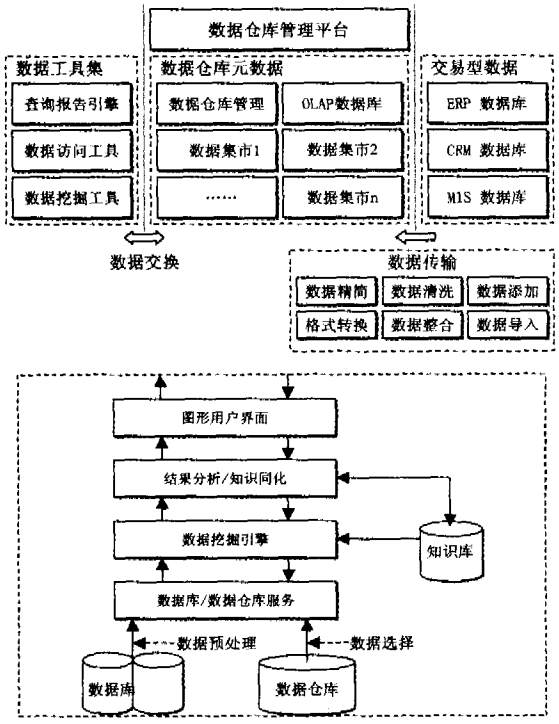

图 2　典型商务网站数据仓库结构

数据仓库管理主要体现在，设计目的、实体关系特征、信息冗余、表格数量、历史记录、用户数量、表格字段数量等方面，在性能上，数据仓库要求快速查询和统计计算。为了减少表与表之间的复杂关系和提高查询运算速度，数据仓库可以采用星型关系模型，事实表包含了所有纬度表的外键，这些外键指向各纬度表的首键，每个纬度都有深度索引；查询先用体积小得多的维度去过滤很大的事实表，从而首先获得较小的相关数据集。

数据集市是一个针对某个主题的经过预统计处理的部门级分析数据库，如销售数据集市、营销数据集市、库存集市和财务集市等。一般理解为企业级数据仓库里的主题数据库，是数据仓库管理系统下的一部分。数据仓库要从各种数据源中获得数据，必须具备有效的输入工具，对这些原始数据进行必要的处理工作。由于数据仓库有自己的独立数据库系统，字段长度、字段类型、索引定义等与交易数据库有很大的不同，数据在导入之前，各种转换工作是必然的。数据仓库的管理平台是系统管理员的日常维护和管理环境，其主要管理任务包括批处理作业管理，数据安全管理，数据冲突解决，数据质量核查，管理数据仓库元数据的更新，数据删除与复制，备份与恢复等，从而为保证数据仓库正常运转提供了基本的管理环境。

2　数据挖掘算法

商务网站数据挖掘表现在大型数据库中搜索有价值的商务信息，需要对巨量的材料进行详细的过滤，需要智能且精确地定位潜在价值。数据挖掘的范围主要有趋势预测和探测以前未发现的模式。在大型数据库里寻找潜在的预测信息。传统上需要很多专家花费大量时间进行分析和研究的问题，现在可以快速而直接地从数据中间找到答案。数据挖掘工具扫描整个数据库并辨认出那些隐藏着的模式。

商务网站数据挖掘的基本过程包括，界定取数对象，明确取数对象的范围，选择目标数据集。数据预处理，去除无关数据，确认数据的品质。数据转换，找到数据的特征表示，提出假设、选取数据挖掘算法（如汇总、分类、回归和聚类等）、提取规则。数据挖掘，根据确认的、合适的挖掘算法自动对数据进行挖掘。分析和评价，解释并评估结果，去掉无实用价值的信息。最后，将分析所得到的知识集成到业务信息系统的组织结构中去。

在传统的单维关联规则挖掘中，通常基于支持度-置信度框架，即类似于 90% 的客户在购买面包的同时也会购买牛奶的规则。这种支持-置信式关联规则的挖掘，在商务网站的应用中却存在着不足，现有的传统关联规则挖掘的对象通常基于交易型事务数据库，寻求交易过程中不同的项目之间潜在的联系，而电子商务网站的一个事务中，采样数据只是表现为属性值的大小的变化。此时，采用传统的单维关联规则挖掘所获得的结果，就没有实际意义。为了弥补现有关联规则挖掘的这种不足，可以采用修改支持度-置信度框架，采用灰关联度框架，构造新的挖掘框架，寻求新的挖掘算法。

灰色系统理论是一种研究信息不完全、不确定问题的新方法。灰关联分析对系统行为特征和因素之间的相互关系进行了量化分析，其基本思想是从几何关系出发，对曲线间相似程度进行量化的比较分析，如果曲线的形状越接近，则其关系越密切。

这种曲线间的相似性度量，获得了系统行为特征和因素之间的潜在关系，而数据挖掘中的关联分析则是获得项目集之间的潜在联系，二者之间具有一定的相似性。根据大量的历史数据对各因素之间相互作用的定量研究，对更好地分析商务网站的运行状况，调整相关的控制策略，具有积极的作用。

结合灰色系统理论和商务网站的实际应用，引入数据挖掘的关联分析中，可以采用适合于商务网站的关联规则挖掘算法，即基于灰关联度框架的灰关联规则挖掘算法，由于拓扑空间具有良好的几何性质，将灰关联度框架建立在拓扑空间的基础之上。该挖掘算法通过灰关联分析，进而获得基于时间属性的灰关联规则，以调整策略，获得更好的控制效果。

由于灰色系统理论是一门新兴的学科，这种新理论仍存在有不完善的地方。例如，在灰关联分析中，对灰关联度的计算，要求序列具有极性一致性以及满足灰色四公理，这种条件

比较苛刻。此外，计算方法本身还存在两点缺陷，首先是局部点关联趋向，即由关联系数大的点决定整体关联度，其次是关联系数的平均值生成关联度，容易造成信息损失。在实际应用中采用的改进方法，如趋势关联度、型关联度、绝对关联度和灰色关联熵等。然而，这些方法通常只是针对其中的某一个缺陷进行调整，仍然有不满意的地方。灰色斜率熵关联度计算方法在上述基础上作出了改进。根据曲线间相似程度，灰色关联度的基本思想如下：

设 $X(t)$ 为母函数，$Y_i(t)(i=1,2,\cdots,m)$ 为子函数，则称

$$\xi(t)=\frac{1+0.5*|\frac{\Delta x(t)}{\Delta t}|}{1+0.5*|\frac{\Delta x(t)}{\Delta t}|+|\frac{\Delta x(t)}{\Delta t}-\frac{\Delta y_i(t)}{\Delta t}|}$$

为函数 $X(t)$ 与 $Y_i(t)$ 在 t 时刻的斜率关联系数。其中

$\frac{\Delta x(t)}{\Delta t}$ 为母函数 $X(t)$ 在 t 到 $t+\Delta t$ 的斜率；

$\frac{\Delta y_i(t)}{\Delta t}$ 为子函数 $Y_i(t)$ 在 t 到 $t+\Delta t$ 的斜率；

且 $\Delta x(t)=x(t+\Delta t)-x(t)$、$\Delta y_i(t)=y_i(t+\Delta t)-y_i(t)$，由于在事务拓扑空间中，所有的序列都为 1 时距的离散序列，故有 $\Delta t=1$。根据上述定义，有灰色斜率关联系数，设 $X_0=\{\langle x_0(k),\oplus\rangle|k=1,2,\cdots,n\}$ 是 Ω 中关于时间$\oplus$的主属性序列，$X_i=\{\langle x_i(k),\oplus\rangle|k=1,2,\cdots,n\}$，$(i=1,2,\cdots,m)$ 是非主属性序列，其中，时间$\oplus$为属性集中需考察某一给定的时间区间，序列中 n 的取值由$\oplus$确定。

令 $\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle)=\frac{1+0.5*a(k)}{1+b(k)+0.5*a(k)}$

其中，$a(k)=|\langle x_0(k+1),\oplus\rangle-\langle x_0(k),\oplus\rangle|$，$b(k)=|(\langle x_0(k+1),\oplus\rangle-\langle x_0(k),\oplus\rangle)-(\langle x_i(k+1),\oplus\rangle-\langle x_i(k),\oplus\rangle)|$，且 $k=1,2,\cdots,n-1$，则称 $\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle)$ 为 X_0 与 X_i 在 k 点的灰色斜率关联系数。

根据 X_i 相对于 X_0 在时间$\oplus$的灰色斜率熵关联度，可以获得相应的灰关联规则。由该条规则，便能得知在时间$\oplus$时不同非主属性对主属性的影响程度，管理人员根据这种影响程度的排序，可以根据实际需要适时地调整控制参数或相关条件，以适应商务网站的实际运作。在采用典型灰关联度计算方法时，对时间区间进行改变，可以获得具有时间属性的灰关联规则集；这些关联规则在不同时间区间的变化，从定性的角度说明了各影响因素的发展变化历史，有利于商务网站控制。

而引入灰色斜率关联熵的概念后，可以构造一个时间窗口，根据商务网站的特性，采用一个时间跨度。利用该时间窗口的滑动，便可以获得同一个因素在不同状态时熵的变化。利用这些熵，构造一条熵值曲线，由熵的涨落便能得知系统的稳定程度，以及与外界的协调性。或者构造一个系统进化的模型，可以表示为，$\Delta S=E_n(X_i)_{k+1}-E_n(X_i)_k$，显然 ΔS 的取值只有 3 种可能：$\Delta S<0$，$\Delta S>0$ 或者 $\Delta S=0$，ΔS 的含义是系统与外界进行能量、信息、物质等的交换引起的熵变。这种熵变，说明了系统状态的变化，为商务网站的分析提供了新的思考方法。

这种时间窗口的滑动引起的熵变和灰关联规则的改变，即为灰关联规则的增量更新。在灰色系统理论中，根据解的非唯一性原理，可能存在有不同的灰关联度。无论是采用典型的灰关联度计算方法，还是利用改进的灰关联度计算方法，在实际的应用中，需要结合商务网站的实际需求，综合比较各种算法，以获得满意的效果。

3 联机分析处理

商务网站联机分析处理工具，采用多维数据库，利用灰关联规则进行数据挖掘，可以发现隐藏在商务数据间的相互关系，它能发现商务数据库中所分析对象之间的关系，把分析所需的数据从数据仓库中抽取出来，物理地组织成多维数据库。联机分析处理可以对商务网站的联机数据访问和分析，通过对信息进行快速、稳定、一致和交互式的存取，对数据进行多层次、多阶段的分析处理，以获得高度归纳的分析结果。电子商务是数据挖掘的重要应用领域，从 WallMart 到 Amazon.com，都有着成功的应用，如销售、顾客、产品、时间和地区的多维分析，网站购物的推荐服务等等。在银行和金融业中，信用欺诈的建模与预测、风险评估、收益分析、客户关系优化以及股票价格、商品价格和金融危机的预测等方面，有着较好的应用。

商务网站联机分析处理是一种自上而下、不断深入的分析工具，在用户提出问题或假设之后，它负责提取出关于此问题的详细信息，并以直观的方式呈现给用户。在进行联机分析处理分析时，从用户角度和查询性能考虑，可以采用多视图模式，即除了使用数据报表格式向用户表现查询结果以外，还采用曲线图、饼图、柱状图等多种格式显示。

参 考 文 献

1 黄斐. 电子商务网站学习网站建设. 计算机应用，2002(7)
2 黄斐. 电子商务网站数据处理. 微机发展，2002(3)
3 刘业翔. 铝电解控制中灰关联规则挖掘算法的应用. 中国有色金属学报，2004(3)
4 邓聚龙. 灰理论基础. 武汉：华中科技大学出版社，2002
5 黄斐. 基于网络环境的项目协调管理. 计算机科学，2004(10)
6 黄斐. 基于 MS Project 的多课程管理系统. 成都：西南交大出版社，2004
7 黄斐. MS Project 2002 项目管理与应用. 科学出版社，2004

（上接第 173 页）

参 考 文 献

1 Haas L M, Kossmann D, Wimmers E L, et al. Optimizing queries across diverse data sources. In: Proc of the 23th VLDB Conf. Athens, 1997. 276～285
2 Levy A, Rajaraman A, Ordille J J. Querying heterogeneous information sources using source descriptions. In: Proc. of the 22th VLDB Conf. Bombay, 1996. 251～262
3 Kushmerik N. Wrapper induction: Efficiency and expressiveness. Journal of Artificial Intelligence, 2000, 118(1-2): 15～68
4 Soderland S. Learning information extraction rules for semi- structured and free text. Journal of Machine learning, 1999, 34(1-3): 233～272
5 Embley D W, Campbell D M. A conceptual-modeling approach to extracting data from the web. In: Proc. of the 17th Intl. Conf on Conceptual Modeling. Singapore, 1998. 78～91
6 Chang C, Lui S. IEPAD: Information extraction based on pattern discovery. In: Proc of 10th WWW Conf. Hong Kong, 2001. 681～688
7 Crescenzi V, Mecca G, Merialdo P. ROADRUNNER: Towards automatic data extraction from large web sites. In: Proc of the 27th VLDB Conf. Roma, 2001. 109～118
8 Crescenzi V, Mecca G. Automatic Information Extraction from Large Websites. Journal of the ACM, 2004, 51(5): 731～779
9 Sarawagi S. Automation in InformationExtraction and Data Integration (tutorial). VLDB, 2002
10 Myllymaki J. Effective Web data extraction with standard XML technologies. In: Proc of 10th WWW Conf. Hong Kong, 2001. 689～696