学校编码：10384　　分类号 _____ 密级 _____

学　　号：200031017　　UDC _____

学　位　论　文

关联规则与用户访问模式挖掘研究

胡慧蓉

指导教师姓名：　王周敬 副教授

自　动　化　系

申请学位级别：　硕　　士

专　业　名　称：　系　统　工　程

论文提交日期：　2003 年 6 月

论文答辩日期：　2003 年 6 月

学位授予单位：　厦　门　大　学

学位授予日期：　2003 年　　月

答辩委员会主席 _____

评　　阅　　人 _____

2003 年 6 月

摘　要

由于 WWW 资源的爆炸性增长，对于已经把 Web 转化为关键发展工具的电子商务来说，运用数据挖掘技术获取用户的访问模式对于电子商务网站的生存发展是十分有益的。把 Web 数据挖掘用于电子商务，可以帮助指导站点改进服务、调整结构和实施有针对性的商业行为，以更好地满足访问者的需求。Web 数据挖掘就在这样的背景下与电子商务结合在一起，它是在 Internet 出现后数据挖掘的一个新的分支，主要研究在 Internet 网络上，对各种数据源，如 Web 日志、用户登记信息、交易数据库、页面内容等，利用数据挖掘技术寻找网络上数据间各种隐含的知识模式和获取一些预测性信息。

本文首先介绍了数据挖掘的一些基本概念、方法和技术、工具。阐明了什么是数据挖掘、为什么要数据挖掘、如何进行数据挖掘、数据挖掘的主要过程以及分类。随后介绍电子商务系统，对两者有了一定认识后，讨论了电子商务网站挖掘的准备工作。接下来从关联规则、用户访问模式两个角度详细介绍了本文的挖掘算法。

关联规则挖掘作为数据挖掘的一个重要研究分支，其主要的研究目的是从大型数据集中发现隐藏的、有趣的、属性间存在的规律。关联规则挖掘首要解决的是效率（Efficiency）与伸缩性（Scalability）问题。由于在电子商务环境中，数据挖掘任务所面对的数据集通常是由数以百万计的记录所构成的大型数据库或数据仓库，因此如何提高从大型数据库中挖掘关联规则的效率和伸缩性，以便有效的降低计算的复杂性、提高算法的运行速度，便成为关联规则挖掘研究中的核心问题，本文的工作都是围绕此核心展开的。

本文首先研究事务数据库的关联规则挖掘问题。首先深入的分析了经典的 Apriori 算法，随后提出了高效算法 SLIG，该算法借鉴离散数学中的关系原理，引入了项目可辩识向量的“与”运算，克服了 Apriori 及其相关算法的缺点，该算法只需对数据库遍历一次，大大提高了算法的效率。

接着本文将算法 SLIG 在多概念层次上进行扩展，提出了基于概念层次树的多层关联规则挖掘算法 MLIG 和交叉层次挖掘算法 CLIG。

另外，电子商务中的交易是一个不断进行的过程，当有新的事务数据集添加到事务数据库中时，关联规则也会发现变化。本文讨论了随着新数据的产生如何增量式地更新频繁项目集，提出了算法 $USLIG_1$。

用户在对事务数据库挖掘的过程中，为了发现事先未知的感兴趣的规则，必然需要不断调整两个阀值：最小支持度和最小可信度，直到发现自己感兴趣的规则为止。这显然是一个动态的交互过程。因此，迫切需要高效的更新算法来满足用户对较快的响应时间的需求。本文在分析了 IUA 算法的基础上，提出了算法 $USLIG_2$，其中详细阐述了该算法的算法思路。

目前学术界对序列模式的研究主要集中在事务数据库中的序列模式的发现上，对 Web 环境下的序列模式发现研究较少，本文在综述 Web 数据挖掘分类、研究内容和目前研究现状、关联规则发现的基础上，明确了 Web 用户访问模式挖掘研究难点在于：如何对原始日志数据进行预处理；如何设计有效的挖掘算法。针对这两个难题，本章研究及总结了预处理技术，介绍如何获取最大向前引用路径 MFP，并利用前一章关联规则挖掘算法的思想，将算法 SLIG 进行扩展，提出算法 FAPG 用于挖掘用户访问频繁路径，算法只需要扫描数据库一遍，因此降低了算法的运行时间，提高了效率。

接着对本文提出的算法进行了实验分析，通过算法分析并且与以前的经典算法对比，实验结果显示了本文提出的算法是十分高效的。

【关键词】： 数据挖掘；电子商务；Web 挖掘；关联规则；频繁访问模式

Abstract

Applying web mining in E-commerce can help a site to improve its service and structure in order to meet the requirement of visitors. Web mining has been combined with E-commerce on this occasion. It is a new branch of data mining and it focuses on the research in the Internet on how to find out all implicit knowledge modes among all kinds of data including web logs, user register information, transaction databases, web page etc, and on how to gain some predictive information.

This paper is organized as follows: First, we describes the data mining techniques in general, concepts, method, techniques and algorithms. It interprets what's data mining, why data mining, how to data mining etc. Then, we introduce E-commerce system. After having known the above two, we introduce the data preparation. Then we analyze mining detailed in two aspects, one is the association rule, and the other is the sequential pattern.

Association rule mining is a form of data mining to discover previously unknown, interesting relationships among the sets of items from large databases. To mining association rule, the primary problem need to solve is the efficiency and scalability of the algorithm. Because in the E-commerce environment, the dataset for mining usually is the large database or data warehouse consists of millions of records, how to improve the efficiency and scalability of the algorithm, so as to efficiently descend the calculation complexity and decrease the execution time then becomes the key problem of mining association rule. In this paper, we focus on this key problem.

In mining association rules, we considers of exploring transaction database. We first introduce famous Apriori algorithm, then propose an efficient algorithm SLIG learning from relationship theory and "AND" operations on recognizable vectors for overcoming the disadvantage of algorithm Apriori and its relative algorithms. Algorithm SLIG scans the database only once, which improve the performance of the algorithm enormously.

Then, we extend algorithm SLIG based on concept hierarchy to form algorithm MLIG and CLIG respectively to mining large itemsets in multiple level and cross level.

Next, we respectively study on how to incremental update large itemsets according to new data added to the databases or when the minimal support threshold changes, furthermore, we propose algorithm to $USLIG_1$ and $USLIG_2$ to solve these two problems and interprets the algorithm processes in detail.

To some degree, sequential patterns are a kind of extension of association rules, however, on many occasion, the effect of sequential pattern can't be replaced with association rules.

So far the study on sequential patterns mainly focuses on transaction database, but the study on sequential patterns in the web environment is relative less. After describing the web mining techniques in general, concepts, method, techniques and association rule algorithms, we affirm the difficulty in mining web access logs is as follows: one is the data preparation of original web access logs; another is how to create an effective algorithm. To solve these two problems, in this paper, we study and synthesize the data preparation technique, introduce how to get MFP. Then we extend SLIG to form FAPG to find frequent user access paths.

Then, we do experiments with the algorithms by use of synthetic data, through the performance evaluation and comparison with the classical algorithms, the experiment results show the algorithms are very effective.

【Key words】: Data Mining; E-commerce; Web Mining; Association Rule; Frequent Access Pattern

目 录

第一章　引　言

1.1. 数据挖掘产生的背景

随着数据库技术的不断发展和数据库管理系统的广泛应用，数据库中存贮的数据量急剧增大。快速增长的海量数据收集、存放在大型和大量数据库中，没有强有力数据分析的工具，理解它们已经远远超出了人的能力。目前人们仅通过数据库系统所能获得的只是整个数据库包含的信息的一部分，隐藏在这些数据背后的更重要的信息是关于这些数据的整体特征的描述及其对其发展趋势的预测，这些重要信息可以很好地支持人们的决策。缺乏数据挖掘背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。

在数据库技术飞速发展的同时，人工智能领域的一个分支——机器学习的研究也取得了很大进展。用数据库管理系统来存储数据，用机器学习的方法来分析数据，挖掘大量数据背后的知识，这两者的结合促成了数据库中的知识发现 KDD(Knowledge Discovery in Database)的产生。

1989 年 8 月在美国底特律召开的第十一届国际人工智能联合会议的专题讨论会上首次出现 KDD 这个术语。随后在 1991 年、1993 年和 1994 年都举行了 KDD 专题讨论会，汇集了来自各个领域的研究人员和应用开发者的成果，集中讨论了数据统计、海量数据分析算法、知识表示、知识运用等问题。随着参与人员的不断增多，KDD 国际会议发展成为年会。

实际上，KDD 是一门交叉性学科，涉及到机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、高性能计算、专家系统等多个领域。从数据库中发现出来的知识可以用在信息管理、过程控制、科学研究、决策支持等许多方面。数据挖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，以及众多的学术单位，例如 UC Berkekey，都在这个领域开展了各种各样的研究计划。

1.2. 数据挖掘在 Web 上的应用

Internet 的迅猛发展，尤其是 Web 的全球普及，使得 Web 上信息量无比丰富。在全球 Web 站点数目迅速增长的同时，各个 Web 站点的信息量及其复杂度也在迅速上升，包含成千上万个网页与超链接是很平常的。在竞争日益激烈的网络经济中，只有赢得用户才能最终赢得竞争的优势。Web 上有海量的数据信息，怎样对这些数据进行复杂的应用成了现今数据库技术的研究热点。电子商务模式下激烈的竞争趋势要求对这些信息进行实时和深层的分析。

把 Web 数据挖掘用于电子商务，可以帮助指导站点改进服务、调整结构和实施针对性的商业行为，以更好地满足访问者的需求。Web 数据挖掘就在这样的背景下与电子商务结合在一起，它是在 Internet 出现后数据挖掘一个新的分支，主要研究在 Internet 网络上，对各种数据源，如 Web 日志、用户登记信息、页面内容、交易数据库等，利用数据挖掘技术寻找网络上数据间各种隐含的知识模式和获取一些预测性信息。当访问者访问某网站时，有关访问者的数据便会被逐渐积累起来。由于以下的因素，数据密集型 Web 站点的挖掘也变得越来越困难。

1） 各个用户具有不同的身份、性别、年龄、背景、语言、文化、习惯、爱好等等，不同的用户访问 Web 站点时带有各自不同的目的，关注的内容也就各不相同；

2） 在不同的时期，同一用户对 Web 信息的需求也可能不同；
3） Web 站点随着时间的推移不断发展，内容逐渐增加，导致其初始的设计不再适合；
4） Web 站点实际提供的信息服务往往超出了其设计的范围，甚至完全改变了定位。

1.3. 论文的组织结构

本文首先对数据挖掘的技术基础、体系结构等方面做了归类和分析。

为了对所需解决问题深入了解，本文在第三章中详细介绍了电子商务基本概念，在分别对数据挖掘技术和电子商务系统有了一定的了解后，第三章对 Web 挖掘作了基本介绍。

接下来在第四章和第五章从关联规则和序列模式两个角度详细阐述挖掘算法。

在第四章关联规则的挖掘中，本文研究的是事务数据库的关联规则挖掘问题。首先深入地分析了经典的 Apriori 算法，Apriori 和相关改进算法都需要产生大量的候选集，而且需要多次扫描数据库。所以本文提出了高效算法 SLIG，该算法借鉴离散数学中的关系原理，克服了 Apriori 及其相关算法的上述缺点。

接着将算法 SLIG 在多概念层次上进行扩展，提出了基于概念层次树的多层关联规则挖掘算法 MLIG 和交叉层次挖掘算法 CLIG。

接下来，本文讨论了随着新数据的产生或者当最小支持度阀值发生变化时，如何增量式地更新频繁项目集，分别提出了算法 $USLIG_1$ 和 $USLIG_2$，其中详细阐述了各算法的算法思路。

随后在第五章中，本文首先对用户访问模式挖掘作了基本介绍，然后在用户的访问路径分析中，我们首先详细叙述了 Web 日志挖掘的预处理过程，接着介绍如何获取最大前向引用路径 MFP，进而提出算法 FAPG 获得频繁访问路径。

第六章对本文提出的算法进行了实验分析，通过算法分析并且与以前的经典算法对比，实验结果显示了本文提出的算法是十分高效的。

最后，对本文进行了总结，并指出了进一步研究的方向。

1.4. 本文的主要工作和创新点

1. 根据关联规则挖掘的数据集的不同，本文研究的是事务数据库的关联规则挖掘问题。关联规则挖掘首要解决的是效率（Efficiency）与伸缩性（Scalability）问题。如何提高从大型数据库中挖掘关联规则的效率和伸缩性，以便有效的降低计算的复杂性、提高算法的运行速度，便成为关联规则挖掘研究中的核心问题，本文的工作都是围绕此核心展开的。

(1) 对国内外在关联规则挖掘方面的研究进行了详细的介绍，对关联规则挖掘问题进行了描述，给出了关联规则挖掘问题的定义、基本过程、关联规则挖掘问题的分解，指出了关联规则挖掘研究现状、研究重点。

(2) 研究了布尔关联规则挖掘问题中各种典型算法，并深入的分析了经典的 Apriori 算法，随后提出了高效算法 SLIG，克服了 Apriori 及其相关算法的缺点。首先，算法只对数据库遍历一次，削减了 I/O 操作时间；其次，该算法采用了关系理论的思想，引入了项目可辩识向量的“与”运算，对于m个项目，只需对m个项目的向量进行 C_m^2 次“与”运算，即可得到满足最小支持度的频繁 2-项集, 因而缩短了产生频繁 2-项集所耗费时间，提高了算法的效率；最后，只需要对 k 个项目的向量进行“与”运算，就可以容易地求得频繁 k-项集的支持度，不需要扫描整个

事务数据库，只需扫描小得多的向量。

(3) 将算法 SLIG 在多概念层次上进行扩展，提出了基于概念层次树的多层关联规则挖掘算法 MLIG 和交叉层次挖掘算法 CLIG。

(4) 另外，电子商务中的交易是一个不断进行的过程，当有新的事务数据集添加到事务数据库中时，关联规则也会发现变化。本文讨论了随着新数据的产生如何增量式地更新频繁项目集，提出了算法 $USLIG_1$。

(5) 用户在对事务数据库挖掘的过程中，为了发现事先未知的感兴趣的规则，必然需要不断调整两个阀值：最小支持度和最小可信度，直到发现自己感兴趣的规则为止，这显然是一个动态的交互过程。因此，迫切需要高效的更新算法来满足用户对较快的响应时间的需求。本文在分析了 IUA 算法的基础上，提出了算法 $USLIG_2$。

2. 目前学术界对序列模式的研究主要集中在事务数据库中的序列模式的发现上，对 Web 环境下的序列模式发现研究较少，本文研究其中一种重要的 Web 序列模式－访问路径模式挖掘。在综述 Web 数据挖掘分类、研究内容和目前研究现状、关联规则发现的基础上，介绍了在电子商务中可以被用来进行数据挖掘的数据源，明确了 Web 用户访问模式挖掘研究难点在于：如何对原始日志数据进行预处理；如何设计有效的挖掘算法。针对这两个难题，本文研究及总结了预处理技术，介绍了如何获取最大前向引用路径 MFP，另外，传统的挖掘频繁访问路径的算法基本都是基于 Apriori 算法或类 Apriori 算法，本文将算法 SLIG 进行扩展，提出算法 FAPG 用于挖掘用户访问频繁路径，算法只需要扫描数据库一遍，因此降低了算法的运行时间，提高了效率。

3. 对本文提出的算法进行了实验分析，通过算法分析并且与以前的经典算法对比，实验结果显示了本文提出的算法是十分高效的。

4. 最后总结全文，并指出了进一步研究的方向。

第二章　数据挖掘技术

一切新事物的产生都是由需求驱动的。希望能让计算机自动智能地分析数据库中的大量数据以获取信息是推动挖掘技术产生并发展的强大动力。例如，超市的经理人员希望能从过去几年的销售记录中分析出顾客的消费习惯和行为，以便及时变换营销策略等。运用数据挖掘技术，为决策者提供重要的、极有价值的信息或知识，从而产生不可估量的效益，成为数据挖掘技术不断发展的推动力。

2.1. 数据挖掘的技术基础

数据挖掘是 KDD 最核心的部分，是采用机器学习、统计等方法进行知识学习的阶段。数据挖掘算法的好坏将直接影响到所发现知识的好坏。目前大多数的研究都集中在数据挖掘算法和应用上，人们往往不严格区分数据挖掘和数据库中的知识发现，把两者混淆使用。一般在科研领域中称为 KDD，而在工程领域则称为数据挖掘。

2.1.1. 数据挖掘的定义和特点

2.1.1.1. 定义

数据挖掘是从大量的数据中，抽取出潜在的、有价值的知识（模式或规则）的过程[1]。

2.1.1.2. 特点

KDD 就是利用机器学习的方法从数据库中提取有价值知识的过程，是数据库技术和机器学习两个学科的交叉学科。数据库技术侧重于对数据存储处理的高效率方法的研究，而机器学习则侧重于设计新的方法从数据中提取知识。由于 KDD 使用的数据来自实际的数据库，所要处理的数据量可能很大，因此 KDD 中的学习算法的效率和可扩充性就尤为重要；此外，KDD 所处理的数据由于来自现实世界，数据的完整性、一致性和正确性都很难保证。如何将这些数据加工成学习算法可以接收的数据也需要进行深入的研究；再者，KDD 可以利用目前数据库技术所取得的研究成果来加快学习过程，提高学习的效率。最后，由于 KDD 处理的数据来自于实际的数据库，而与这些数据库数据有关的还有其他的一些背景知识，这些背景知识的合理运用也会提高学习算法的效率。

2.1.1.3. 数据挖掘的任务

从数据中发现模式。模式是一个用语言 L 来表示的一个表达式 E，它可用来描述数据集 F 中数据的特性，E 所描述的数据是集合 F 的一个子集 FE。E 作为一个模式要求它比列举数据子集 FE 中所有元素的描述方法简单。例如，“如果成绩在 81～90 之间，则成绩优良”可称为一个模式，而“如果成绩为 81、82、83、84、85、86、87、88、89 或 90，则成绩优良”就不能称之为一个模式。

模式有很多种，按功能可分为两大类：预测型（Predictive）模式和描述型（Descriptive）模式。

预测型模式是可以根据数据项的值精确确定某种结果的模式。挖掘预测型模式所使用的数据也都是可以明确知道结果的。例如，根据各种动物的资料，可以建立这样的模式：凡是胎生的动物都

是哺乳类动物。当有新的动物资料时，就可以根据这个模式判别此动物是否是哺乳动物。

描述型模式是对数据中存在的规则的一种描述，或者根据数据的相似性把数据分组。描述型模式不能直接用于预测。例如，在地球上，70%的表面被水覆盖，30%是陆地。

在实际应用中，往往根据模式的实际应用细分为以下 6 种：

（1） **分类模式：**分类模式是一个分类函数（分类器），能够把数据集中的数据项映射到某个给定的类上。分类模式往往表现为一棵分类树，根据数据的值从树根开始搜索，沿着数据满足的分支往下走，走到树叶就能确定类别。

（2） **回归模式：**回归模式的函数定义与分类模式相似，它们的差别在于分类模式的预测值是离散的，回归模式的预测值是连续的。如给出某种动物的特征，可以用分类模式判定这种动物是哺乳动物还是鸟类；给出某个人的教育情况、工作经验，可以用回归模式判定这个人的年工资在哪个范围内，5000 元以下，还是在 5000 元到 1 万元之间，抑或是在 1 万元以上。

（3） **时序模式：**时序模式根据数据随时间变化的趋势预测将来的值。这里要考虑到时间的特殊性质，（像一些周期性的时间定义如星期、月、季、年等）、不同的日子可能造成的影响、日期本身的计算方法，还有一些需要特殊考虑的地方如时间前后的相关性（过去的事情对将来有多大的影响力）等。只有充分考虑时间因素，利用现有数据随时间变化的一系列的值，才能更好地预测将来的值。

（4） **聚类模式：**聚类模式把数据划分到不同的组中，组之间的差别尽可能大，组内的差别尽可能小。与分类模式不同，进行聚类前并不知道将要划分成几个组和什么样的组，也不知道根据哪一（几）个数据项来定义组。

（5） **关联模式：**关联模式是数据项之间的关联规则。关联规则是如下形式的一种规则：“在无力偿还贷款的人当中，60%的人的月收入在 3000 元以下。”

（6） **序列模式：**序列模式与关联模式相仿，而把数据之间的关联性与时间联系起来。为了发现序列模式，不仅需要知道事件是否发生，而且需要确定事件发生的时间。例如，在购买彩电的人们当中，60%的人会在 3 个月内购买影碟机。

在解决实际问题时，经常要同时使用多种模式。分类模式、回归模式、时序模式也被认为受监督知识，因为在建立模式前数据的结果是已知的，可以直接用来检测模式的准确性，模式的产生是在受监督的情况下进行的，一般在建立这些模式时，使用一部分数据作为样本，而用另一部分数据来检验模式。聚类模式、关联模式、序列模式则是非监督知识，因为在模式建立前结果是为未知的，模式的产生不受任何监督。

2.1.2. 数据挖掘的方法和技术

目前数据挖掘研究和开发表明数据挖掘需要覆盖各种各样不同的应用任务，从数据的预处理到关联规则、聚类分析、数据分类、偏差检查、序列模式等等特定的模式。因此，这一技术应用是一个极富挑战性的任务。

数据挖掘是从大量的数据中抽取以前未知并具有潜在可用的模式。然而数据挖掘领域还缺乏独立性，数据挖掘是人工智能 (AI)技术与数据库技术的结合。它的核心概念是 AI 领域中的机器学习。数据挖掘系统所采用的主要算法是 AI 中知识发现技术的应用。下面介绍数据挖掘和知识发现的几种常用方法。

1. 人工神经网络（Artificial Neural Networks）

神经网络方法是模拟人脑神经元结构，以 MP 模型和 Hebb 学习规则为基础。它主要有三种神经网络模型：

（1） 前馈式网络。它以感知机、反向传播模型、函数型网络为代表，可用于预测、模式识

别等方面。

（2） 反馈式网络。它以 Hopfield 的离散模型和连续模型为代表，分别用于联想记忆和优化计算。

（3） 自组织网络。它以 ART 模型、Koholon 模型为代表，用于聚类分析等方面。

神经网络的知识体现在网络连接的权值上，是一个分布式矩阵结构；神经网络的学习体现在神经网络权值的逐步计算上，包括反复迭代和累加计算。

2. 遗传算法（Genetic Algorithms）

遗传算法是模拟生物进化过程的算法，由三个基本算子（或过程）组成：

（1）繁殖（选择）。即从一个旧种群（父代）选出生命力强的个体，产生新的种群（后代）的过程。

（2）交叉（重组）。即对选择两个不同的个体（染色体）的部分基因进行交换，形成新的个体的过程。

（3）变异（突变）。即对某些个体的某些基因进行变异（0 变 1，或 1 变 0），形成新的个体的过程。

遗传算法可起到产生优良后代的作用。这些后代需满足适应值，经过若干代的遗传，将得到满足要求的后代（即问题的解）。遗传算法已在优化计算和分类机器学习方面发挥了显著作用。

3. 决策树方法（Decision Trees）

决策树方法是利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的属性字段，建立决策树的一个结点，再根据该属性字段的不同取值建立树的分支；在每个分支集中重复建立树的下层结点和分支的过程。国际上最早的、也是最有影响的决策树方法是 Quiulan 研究的 ID3 方法[2]。

决策树方法在现有的数据挖掘产品中有较为广泛的应用，如 Business Object 公司在它的 OLAP 产品新增加的一个数据挖掘的模块 Business Miner，其中就采用了一种称为 GINI 的决策树方法。

在数据挖掘和知识发现中应用的人工智能技术还有邻近搜索方法、集合论的粗糙集方法、规则推理（Rule Induction）、模糊逻辑（Fuzzy Logic）、公式发现等等。

2.1.3. 数据挖掘的分析方法

无论采用哪几种技术来完成任务，从功能上可以将数据挖掘的分析方法划分为以下四种（根据 IBM 的划分方法）：关联分析（Association Analysis）；序列模式分析（Sequential Patterns）；分类分析（Classifiers）；聚类分析（Clustering）。

1. 关联分析（Association Analysis）

顾名思义，关联分析的目的就是为了挖掘出隐藏在数据间的相互关系。关联分析发现关联规则，关联规则是形如 $A \Rightarrow B$，即“$A_1 \wedge \cdots \wedge A_m \Rightarrow B_1 \wedge \cdots \wedge B_n$” 的规则，其中，$A_i$（$i \in \{1,\ldots,m\}$），$B_j$（$j \in \{1,\ldots,n\}$）是属性－值对，与之相关的是支持度（Support）和可信度（Confidence）[1]。

例如， 给定某公司交易数据库，一个数据挖掘系统可能发现如下形式的关联规则

age (X, “20～29”) ∧ income (X, “20K～29K”) ⇒ buys (X, “CD_player”)

[Support=2%, Confidence=60%]

其中 X 是变量，代表顾客。该规则是说，所研究的公司交易数据库中有 2%（支持度）的顾客在 20～29 岁，年收入 20K～29K，并且在该公司购买 CD 机。这个年龄和收入组的顾客购买 CD 机的可能性有 60%（可信度）。

2. 序列模式分析（Sequential Patterns）

序列模式分析和关联分析法相似，其目的也是为了挖掘出数据之间的联系，但序列模式分析的侧重点在于分析数据间的前后（因果）关系。运用序列模式分析销售记录，零售商可以发现客户潜在的购物模式，例如 9 个月以前购买奔腾 PC 的客户很可能在一个月内购买新的 CPU 芯片。

3.　分类分析（Classifiers）

假定记录集合和一组标记（TAG），所谓标记是指一组具有不同特征的类别。分类分析时首先为每一个记录赋予一个标记，即按标记分类记录，然后检查这些标定的记录，描述出这些记录的特征。这种描述可以是显式的，例如一组规则定义；或者是隐式的，例如一个数学模型或公式。利用它可以分类新纪录，实际上它就是一种模式。目前已有很多种分类分析模型得到应用，其中的几种典型模型是线性回归模型、决策树模型、基于规则模型和神经网络模型。

举一个简单的例子，信用卡公司的数据库中保存着各持卡人的记录，并根据信誉程度（标记），将持卡人分作三类：良好，普通，较差。这一过程实际就是将持卡人记录标定为三类。分类分析法检查这些记录，然后给出一个对信誉等级的显示描述：

“信誉良好的用户是指那些年收入在 25000 美元以上，年龄在 45 到 55 岁之间，居住在 XYZ 地区附近的人士”。

4.　聚类分析（Clustering）

与分类分析法不同，聚类分析法的输入集是一组未标定的记录，也就是说此时输入的记录还没有进行任何分类。其目的是根据一定的规则，合理地划分记录集合，并用显式或隐式的方法描述不同的类别。而所依据的这些规则是由聚类分析工具定义的。由于聚类分析可以采用不同的算法，所以对于相同的记录集合可能有不同的划分。

2.2.　数据挖掘的结构和步骤

2.2.1.　数据挖掘的过程

知识发现和数据挖掘过程是交互式的(interactive)和循环往复的(iterative)。Fayyad, Piatetsky-Shapiro, &Smyth 在 “Knowledge Discovery and Data Mining Towards a Unifying Framework”这篇论文提出了数据挖掘过程图，如图 2.1 所示。

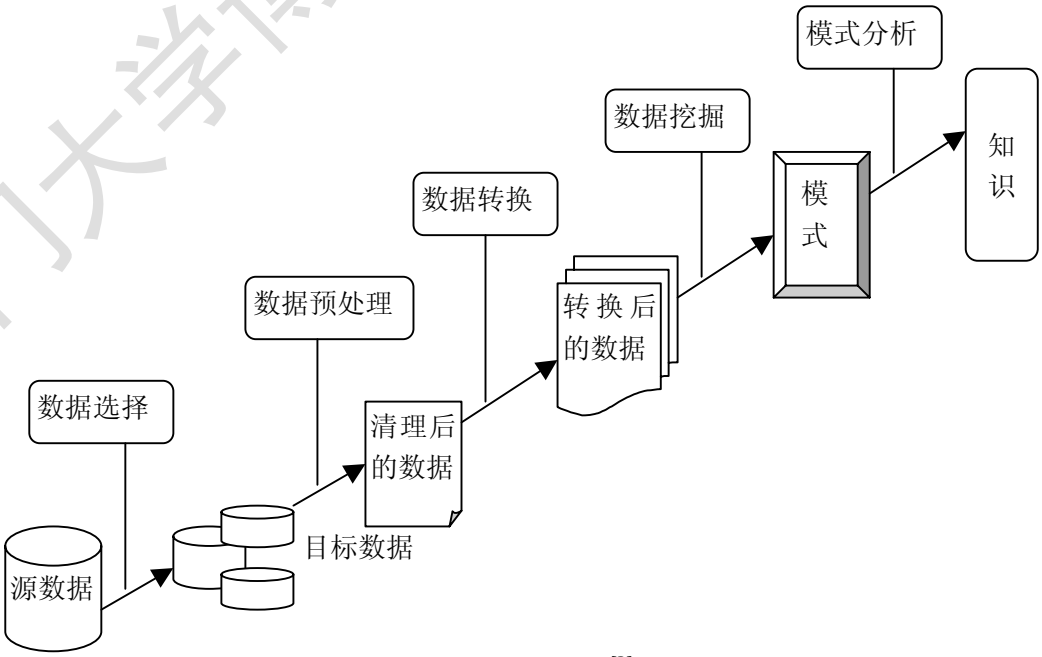

图 2.1　数据挖掘过程图[3]

2.2.2. 数据挖掘系统的结构

如上所述，数据挖掘的核心技术是人工智能、机器学习、统计等，但一个 DM 系统绝不是多项技术的简单组合，而是一个完整的整体，它还需要其它辅助技术的支持，才能完成数据采集、预处理、数据分析、结果表述这一系列任务，最后将分析结果呈现在用户面前。根据功能，整个 DM 系统可以大致划分为三级结构（如图 2.2）。

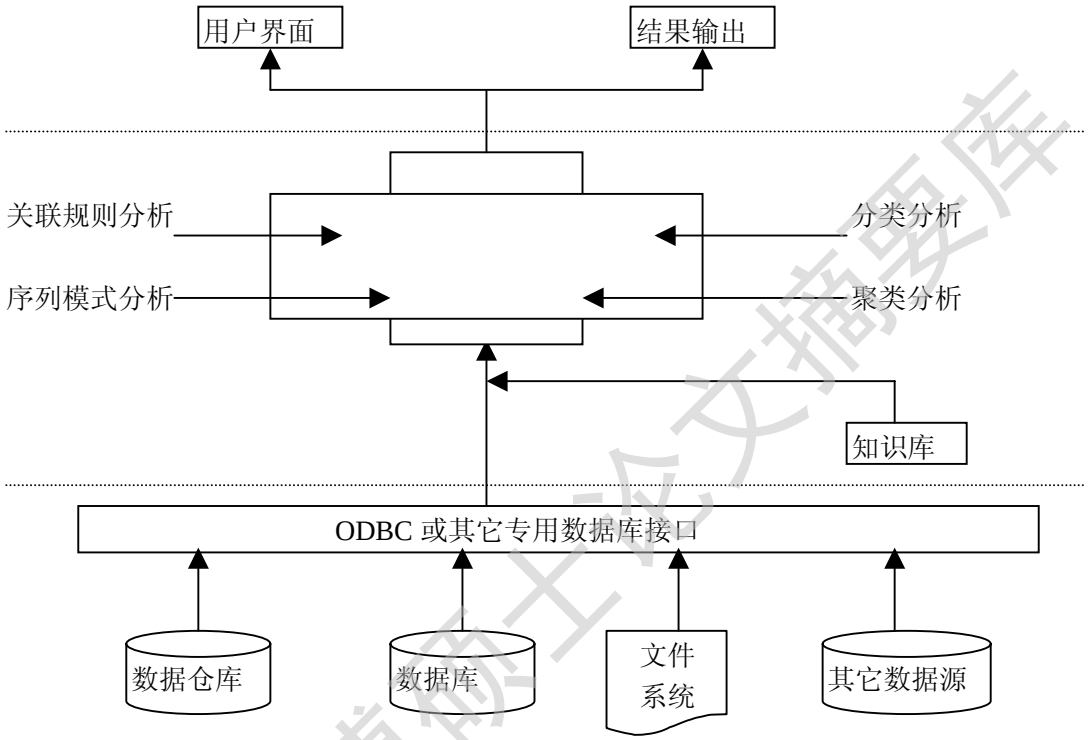

图 2.2 DM 系统的三层结构[4]

2.2.3. 数据挖掘的步骤

1. 数据准备（Data Preparation）

本阶段主要完成数据集成，数据选择和预分析。

2. 挖掘 (Mining)

综合利用前面提到的四种挖掘方法分析数据库中的数据。

3. 表述 （Presentation）

数据挖掘将获取的信息以便于用户理解和观察的方式反映给用户，这时可以利用可视化工具。由于用户要求的不同，DM 分析的数据的范围也会有所不同，例如分析一年内或三个月内的销售情况，再例如分析东部地区或西部地区的销售情况，这样 DM 系统会得出不同的结论。这些基于不同数据集合的分析结果除了通过可视化工具提供给用户外，还可以存储在知识库中，供日后进一步分析和比较。

4. 评价 (Assess)

如果分析人员对分析结果不满意，可以递归地执行上述三个过程，直到满意为止。

2.3. 数据挖掘的应用

2.3.1. 数据挖掘应用设计原则

数据挖掘应用设计原则如下：

1. 面向商业应用

数据挖掘技术从一开始就是面向应用的。它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导实际问题的求解，发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。把人们对数据的应用，从低层次的末端查询操作，提高到为各级经营决策提供决策支持。同时需要指出的是，所有发现的知识都是相对的，是有特定前提和约束条件、面向特定领域的，同时还要能够易于被用户理解，最好能用自然语言表达发现结果。

2. 数据挖掘算法的有效性和可测性

海量数据库通常有上百个属性和表及数百万个元组。GB 量级数据库已不鲜见，TB 量级数据库已经出现，高维大型数据库不仅增大了搜索空间，也增加了错误模式的可能性。因此必须利用领域知识降低维数，除去无关数据，从而提高算法效率。从一个大型数据库中抽取知识的算法必须高效、可测量，即数据挖掘算法的运行时间必须可预测，且可接受，指数和多项式复杂性的算法不具有实用价值。

2.3.2. 数据挖掘的应用

数据的迅速增加与数据分析方法的滞后之间的矛盾越来越突出，人们也希望能够在对已有的大量数据分析的基础上进行科学研究、商业决策或者企业管理，但是目前所拥有的数据分析工具很难对数据进行深层次的处理，使得人们只能望 “数” 兴叹。

数据挖掘从大量数据中提取出隐藏在数据之后的有用的信息，它被越来越多的领域所采用，并取得了较好的效果，为人们的正确决策提供了很大的帮助。

把数据挖掘技术应用到电子商务网站，从大量的用户访问数据中获取有用的信息，将为网站的发展提供很大的帮助。那么如何在电子商务网站中应用数据挖掘技术呢？为了更好的把两者结合起来，首先我们要对电子商务系统进行深入的了解。

第三章　电子商务与 Web 挖掘

随着计算机技术的发展和 Internet 的普及，在各级网站的服务器中的 WWW 数据也飞速膨胀。尽管传统的数据库技术和数据挖掘技术已经取得了飞速的发展并且日趋完善，但由于 Web 数据应用的特殊性，使得传统的技术不能直接应用在 Web 的信息挖掘中。Web 日志数据是记录用户对 Web 站点访问信息的数据，保存有大量的路径信息，对这些信息的分析有利于设计人员掌握用户的喜好和访问习惯，并可以用来对网站的结构进行优化和页面重组。

由于 WWW 资源的爆炸性增长，对于已经把 Web 转化为关键发展工具的电子商务来说，运用数据挖掘技术获取用户的访问模式对于电子商务网站的生存发展是十分有益的。本章将主要对电子商务与 Web 挖掘作基本介绍。

3.1. 概述

3.1.1. 电子商务的概念

电子商务是于九十年代初，在欧美兴起的一种全新的商业交易模式，它实现了交易的无纸化、效率化、自动化，表现了网络最具魅力的地方。快速的交换信息、地理界限的模糊，这所有的一切也必将推动传统商业行为在网络时代的变革。早在网络盛行的时代，通过网络的电子邮件、视频交换、文件交换以及目前还很热门的 EDI（电子数据交换）所进行的商业行为，都可以说是在电子商务的某种形式的表现，也就是说，电子商务是它们崭新的应用集合。

那么什么是电子商务呢？电子商务[5]的英文名称为 Electronic Business，简称 EB，也有的将其称为 Electronic Commerce，前者指广义的商务，而后者则更确切表现出企业的商务运作。对于确切的电子商务的概念，目前并没有比较统一的定义，只是在实践应用的基础上加以总结形成的。经合组织 OECD 是较早对电子商务进行系统研究的机构，它将电子商务定义为是关于利用电子化手段从事的商业活动，它基于电子处理和信息技术，如文本、声音和图像等数据传输，主要是遵循 TCP/IP 协议、通讯传输标准，遵循 WEB 信息交换标准，提供安全保密技术。OECD 的定义特别强调了 Internet 基础上的电子商务发展，但是是不全面的。我们可以把电子商务理解为以信息技术为基础的商务活动，它包括生产、流通、分配、交换和消费等环节中连接生产和消费的所有活动的电子信息化处理。

3.1.2. 电子商务的特性

普遍性：电子商务作为一种新型的交易方式，将生产企业、流通企业以及消费者和政府带入了一个网络经济、数字化生存的新天地；

方便性：在电子商务环境中，人们不再受地域的限制，客户能以非常简捷的方式完成过去较为繁杂的商务活动，如通过网络银行能够全天候地存取资金帐户、查询信息等，同时使得企业对客户的服务质量可以大大提高；

整体性：电子商务能够规范事务处理的工作流程，将人工操作和电子信息处理集成为一个不可分割的整体，这样不仅能提高人力和物力的利用，也可以提高系统运行的严密性；

安全性：在电子商务中，安全性是一个至关重要的核心问题，它要求网络能提供一种端到端的安全解决方案，如加密机制、签名机制、安全管理、存取控制、防火墙、防病毒保护等等，这与传统的商务活动有着很大的不同；

协调性：商务活动本身是一种协调过程，它需要客户与公司内部、生产商、批发商、零售商间

的协调，在电子商务环境中，它更要求银行、配送中心、通讯部门、技术服务等多个部门的通力协作，往往电子商务的全过程是一气呵成的。

3.1.3. 电子商务的现状和发展

以微电子、计算机、通信和网络技术为代表的信息技术，是迄今为止人类社会技术进步过程中发展最快、渗透性最强、应用最广泛的关键技术，代表着先进生产力的发展方向。信息技术的广泛应用，使信息成为重要的生产要素和战略资源，使社会资源能获得高效配置，大幅度提高社会劳动生产率，推动经济结构革新和产业结构升级，并将对全球范围的经济、政治、军事、文化以及意识形态产生越来越广泛和深刻的影响，最终导致经济增长方式，经济管理体制的重大变革，推动工业经济走向新经济——网络经济的新阶段。

最新研究表明，网络经济蓬勃发展，美国、欧洲互联网经济持续增长。电子商务是网络经济的驱动和主体，电子商务从产生开始就以不可阻挡之势和惊人的速度发展着。全世界电子商务的发展，根据 IDG 电子商务研究中心 19991 年 11 月预测：2003 年全球电子商务营业额将达到 28000 亿美元。

作为一种商务活动过程，电子商务将带来一场史无前例的革命。而其影响将远远超出商务本身，它将会对社会的生产和管理、人们的生活和就业、政府职能、法律制度以及教育文化都会带来巨大的影响。电子商务将人类真正带入信息社会。

3.2. 电子商务挖掘分析

当前经济模式的变化，从传统的实体的商店到 Internet 上的电子交易，同时也改变了销售商和顾客的关系。现在，网上顾客的流动性很大，他们关注的主要因素是商品的价值，而不象以前注意品牌和地理因素。因此，电子销售商一个主要的挑战是需要了解到顾客尽可能多的爱好，价值取向，以保证在电子商务时代的竞争力。

随着 Web 技术的发展，各类电子商务网站风起云涌，建立起一个电子商务网站并不困难，困难的是如何让您的电子商务网站有效益。要想有效益就必须吸引客户，增加能带来效益的客户忠诚度。所有客户行为都可以存储在 Web Log 文件中，使得大量收集每个客户的每一个行为数据、深入研究客户行为成为可能。如何利用这个机会，从这些“无意义”的繁琐数据中得到大家都能看懂的、有价值的信息和知识是我们面临的问题。

电子商务业务的竞争比传统的业务竞争更加激烈，原因有很多方面，其中一个因素就是客户从一个电子商务网站转换到竞争对手那边，只需点击几下鼠标即可。网站的内容和层次、用词、标题、奖励方案、服务等任何一个地方都有可能成为吸引客户、同时也可能成为失去客户的因素。而同时电子商务网站每天都可能有上百万次的在线交易，生成大量的记录文件（Log Files）和登记表，如何对这些数据进行分析和挖掘，充分了解客户的喜好、购买模式，甚至是客户一时的冲动，设计出满足不同客户群体需要的个性化网站，进而增加其竞争力，几乎变得势在必行。这样一句话恰好体现电子商务网站的生存发展之道：若想在竞争中生存进而获胜，就要比您的竞争对手更了解客户。

3.2.1. 电子商务挖掘分析的重点

1. 交易数据库中项目间的关联

在某一电子商务 Web 交易数据库中，我们可以发现“购买了 A 产品的用户同时购买 B 产品的可能性是 81%”，这样的模式可以用来帮助安排网站商品的陈列方式以及是否降价出售等等。