学校编码：10384　　分类号_____密级_____

学　　号：200131004　　UDC_____

学　位　论　文

基于 Web 日志的数据挖掘研究及应用

Research and Application: Data Mining Based on Web Log

刘平安

指导教师姓名：罗林开　副教授

申请学位级别：硕　士

专 业　名 称：控制理论与控制工程

论文提交日期：2004 年 6 月

论文答辩时间：2004 年 6 月

学位授予单位：厦 门　大 学

学位授予日期：

答辩委员会主席：_____

评　　阅　　人：_____

2004 年 **6** 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

摘 要

关键词：Web 日志挖掘；数据预处理；聚类分析

由于 Internet 资源的爆炸性增长，进行 Web 数据挖掘，从 Internet 上瀚如烟海的数据中获取有用的知识，已经成为当前一个十分活跃的研究领域。

应用 Web 数据挖掘技术获取用户访问模式的 Web 日志挖掘对于 Web 站点的生存发展是十分有利的。Web 挖掘可以帮助指导站点改进服务、调整结构和实施有针对性的、个性化的商业行为，以便更好的满足访问者的需要。

本文首先介绍了数据挖掘的一些基本概念、方法和技术、工具等。阐述了什么是数据挖掘、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和分类等。随后介绍了 Web 数据挖掘的概念和技术，以及 Web 内容挖掘、Web 结构挖掘、Web 使用挖掘的定义。

同时，由于 XML 标准的出现，对 Web 数据挖掘的异种数据源的集成提供了一个有利的标准，我们介绍了 XML 技术，以及它在 Web 挖掘中的应用。在接下来的章节，着重介绍了 Web 使用挖掘，以及其系统模型结构，详细介绍了数据预处理的 6 个阶段：数据净化、用户识别、会话识别、路径补充、事务识别和数据格式化，同时介绍了事务识别方法 MFP 算法。

结合厦门大学就业指导中心网站 Web 日志对 Web 日志挖掘进行了具体的实施和应用。

本文给出了事务识别的经典算法:最大向前引用路径算法，并对算法的过程进行了详细的分析和具体的实现。描述了一种快速有效的用户访问模式聚类算法 CLOPE，并结合 MFP 算法得到的事务数据进行了实现。

另外，根据得到的频繁访问路径，对经典的关联规则发现算法 Apriori 进行了改造，使它适合于 Web 用户频繁访问路径的发现，提出了一种类 Apriori 算法，并给予了 C++ 代码实现。

对挖掘结果的可视化，作了简单的尝试。

最后，我们设计了一个简单的 Web 日志挖掘系统原型。

Abstract

Key Words:Web Log Mining;Data Preprocessing;Clustering Analysis

The World Wide Web (WWW) continues to grow at an astounding rate in both the sheer volume of traffic and the size and complexity of Web sites.With the huge amount of information available online,the World Wide Web is a fertile area for data mining research.The web mining research is a most vibrant area.

Web log mining which applies web mining techniques to find user access patterns is of benefit to the web sites.

Web mining can guide a web site to better serve the needs of users of the site,adjust its structure and implement effective ,characteristic business behaviors.

First,some basic concepts,methods and techniques,tools of data mining and so on are introduced.What's data mining,why should need data mining and how to data mining,main procedures and categories of data mining are stated in this paper.

Then,concepts and techniques of web mining,and concepts of web content mining,web structure mining and web usage mining are introduced.XML standards coming forth,to integrate all kinds of heterogeneous data sources on the internet has one good chance.So,XML techniques and its application in web mining are presented here.The following chapters focus on the introduction to the web usage mining and its system model,and interpret the 6 phases of data preprocessing in detail:data cleaning,user identification,session identification,path completion,transaction identification and data formatting.In the meantime,an transaction identification method named MFP algorithm is interpreted.

With the log of web site of Career Center,Xiamen University,web log mining is carried into effect.

Classical transaction identification algorithm:maximal forward reference generation is described in this paper,and a detailed implementation for the algorithm is carried out.A fast and effective clustering algorithm for user access pattern is described in detail,and with the transactions generated by MFP,the algorithm is implemented.

In addition,with the frequent access paths,a modification of classical assocication rules algorithm Aprior is made to adapt to mining web user frequent access paths.As a result,an Apriori-Like algorithm is put forward and its implementation by C++ is carried out.A simple attempt is made for visualization of mining results.

In the last place,a simple web log minging system prototype is designed.

第一章　　引言

1.1 Web 数据挖掘产生的背景和意义

1.1.1. 研究背景

“数据丰富，信息贫乏”的矛盾使得数据挖掘的产生成为必然。

自 20 世纪 70 年代以来，由于计算机硬件日新月异的进步以及数据库技术的不断发展和数据库管理系统的广泛应用，导致了数据库中存储的数据量急剧膨胀。快速增长的海量数据收集、存放在大型和大量数据库中，没有强大的数据分析工具，这些海量的数据无非是一堆让人无法理解二进制码，理解它们已经远远超出了人的能力，人们需要凭借强大的数据分析工具，让这些硬梆梆的二进制码变为人们可以理解的自然语言。以前，因为决策者缺乏从海量数据中提取有价值知识的工具，重要的决策常常不是基于数据库中信息丰富的数据，而是基于决策者的直觉。另外，当前有些专家系统技术用于数据分析过程，这种系统依赖于用户或相当领域的专家人工地将知识输入知识库，数据分析过程常常带有偏差和错误，并且耗时、费用高。海量数据和有用知识之间的矛盾要求系统地开发数据挖掘工具。使用数据挖掘工具进行数据分析，从海量数据中挖掘出有趣数据模式，对商务决策、知识库、科学作出了巨大贡献。

数据挖掘是指从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘有趣知识的过程。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，它的迅速发展，引起了学术界和工业界的广泛关注。许多著名的工业研究实验室，如 IBM Almaden 和 GTE，及众多的学术单位，如 uC Berkeley，都在这个领域开展了各种各样的研究。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。

20 世纪 90 年代以来，Internet 得到了的飞速发展，WWW 作为全球最大、最方便的信息集散地，更是积聚了海量的信息，成为人们工作与学习的最大支持平台。在全球 Web 站点服务器数目迅速增长的同时，各个 Web 站点的信息量和复杂度也在迅速上升，形成一个数以亿计的超文本的载体。如何从大量的页面中发现需要的内容，如何从大量的访问中发现固有的模式和关联，成了人们迫切希望解决的问题。Web 挖掘作为数据挖掘和 Internet 技术的结合，研究网上内容自动分类，智能 Agent，用户访问模式发现，成了计算机工作者研究的新热点。

1.1.2. 研究意义

Internet 时代，如何从 Web 上海量的数据中发现隐藏在数据背后有用的知识给人类造成了巨大的挑战。Web 数据挖掘为解决这个问题的指出了一条道路。

数据挖掘在传统的结构化的事务数据的挖掘领域，已经取得了比较成功的应用。然而，Web 上的信息不同于结构化的数据库，Web 上包括文本、图片、多媒体等多种信息，它们是半结构化的。因此，Web 上的挖掘需要用到不同于常规的数据库开采的很多技术。现实领域中，存在的多是半结构化的、异源的数据，Web 挖掘的研究也将极大的推动数据挖掘在其它领域的应用。

Web 内容挖掘提供了自动的文档分类与聚类功能，及基于内容挖掘的智能搜索代理可以给人们提供更好的信息服务。而 Web 使用模式的挖掘，能够辅助改进分布式网络的设计性能，如在有高度相关的站点间提供快速有效的访问通道；能帮助更好的组织 Web 页面，帮助改善市场营销策略，如把广告放在适当的 Web 页面上或更好的理解用户的兴趣。

1.2 发展现状

数据挖掘作为一个大的研究领域，目前在很多生产领域都有了一些成功应用.然而，互联网上的数据挖掘作为一个具有广阔应用前景的新兴应用领域，国外的研究已经取得了初步的成果，国内的研究则刚刚起步。新加坡南洋大学，美国明尼苏达州立大学，澳大利亚的Simon Fraser大学都展开了这方面的研究，并推出一些原型系统，包括WIND，WebMinner，WebWatcher，WHOWEDA等等。它们主要为分两种:站点文档自动分类，用户导航系统和用户访问模式挖掘系统。

1.3 本文的主要工作和创新点

本文的工作内容主要是研究了国外与国内Web数据挖掘方面的学术和应用成果，并运用这些方法进行了应用。阐述了基于用户访问模式和关联规则发现的Web用户访问日志挖掘系统的过程与框架。尤其在日志的预处理方面进行的较为详尽研究。

本文的主要工作和创新点:

1. 研究了国内外在Web挖掘方面所取得的研究成果，探讨了Web日志挖掘过程中的数据预处理问题。

2. 研究了聚类算法，介绍了一种适合大型事务数据库的Web日志用户访问模式快速有效的聚类算法，把其应用于就业指导网站的日志挖掘。

3. 研究关联规则发现算法，提出了适用于用户访问模式挖掘的类Apriori算法进行页面关联规则挖掘。

4. 初步探讨了XML技术与Web挖掘的结合。

5. 开发了一个Web日志挖掘原型系统。

1.4 论文的组织结构

第一章为引言，主要介绍了数据挖掘与Web数据挖掘的研究背景、意义以及发展现状和趋势。

第二章讨论数据挖掘的一些基本概念，数据挖掘分类与任务，方法与技术，并介绍了几种数据挖掘工具。

第三章介绍了什么是Web数据挖掘，它的内容，它的一些方法与技术以及它与数据挖掘的关系。

第四章介绍Web使用挖掘的基本概念和技术。

第五章是Web日志挖掘系统的应用，介绍了本文设计的系统，并提出Web日志挖掘的具体实施过程和算法。

第六章总结和展望了Web挖掘技术的发展和应用前景。

第二章　　数据挖掘技术

数据挖掘是信息技术自然演化的结果。由于存在大量数据，迫切需要能有智能地分析这些数据以转换成有用的信息和知识的技术，数据挖掘技术便应运而生。数据挖掘技术可以为商务管理、决策支持、生产控制、市场分析、工程设计和科学探索等提供极其有价值的信息或知识。

2.1. KDD 和数据挖掘

KDD 即数据库中的知识发现（Knowledge Discovery in Database）这一术语首先出现在 1989 年在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上，1991、1993 和 1994 年又接着继续举行 KDD 专题讨论会。1995 年在加拿大召开了第一届知识发现和数据挖掘国际学术会议。

KDD 是一门交叉性学科，涉及了机器学习、模式识别、统计学、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。

许多人把数据挖掘视为和 KDD 等价的概念。而另一些认则把 KDD 看作是发现知识的完整过程，而将数据挖掘视为其中的一个基本步骤。 这里我们采用数据挖掘作为知识发现过程的一个重要步骤的观点。

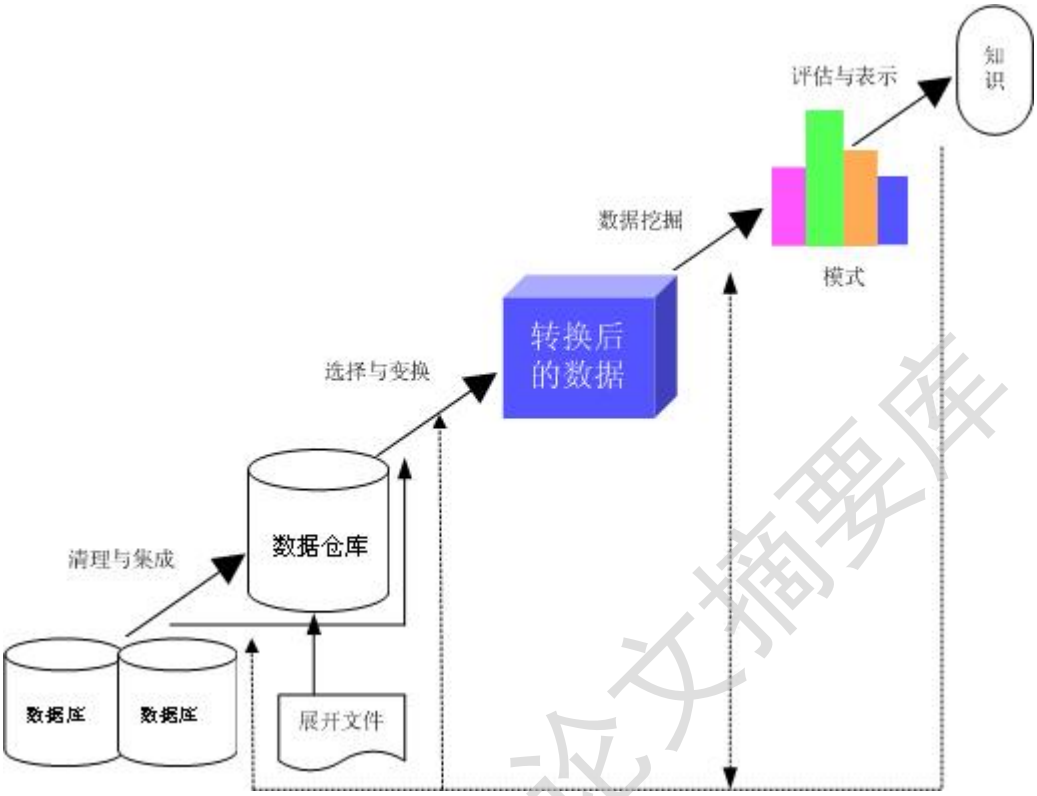

图 2.1 数据库中的知识发现过程（数据挖掘视为知识发现过程的一个步骤）[1]

知识发现基本步骤[1]：

1) **数据清理**（消除噪声或不一致数据）
2) **数据集成**（多种数据源组合在一起）
3) **数据选择**（从数据库中检索与分析任务相关的数据）
4) **数据变换**（数据变换或统一成适合挖掘的形式，如通过汇总或聚集操作）
5) **数据挖掘**（基本步骤，使用智能方法提取数据模式）
6) **模式评估**（根据某种兴趣度度量，识别表示知识的真正有趣的模式）
7) **知识表示**（使用可视化和知识表示技术，向用户提供挖掘的知识）

我们将前 4 个步骤统称为数据预处理过程(Data Preprocession)。

数据挖掘质量的好坏主要受两个因素的影响:数据挖掘算法的有效性、可伸缩性和用于挖掘的数据的质量和数量(数据量的大小)。

如果选择了错误的数据或属性，或对数据进行了错误的转换，即使挖掘算法非常有效，但由于基于质量不高的数据源，则有可能得到不正确的挖掘结果。所以，数据预处理对于数据挖掘来讲也是非常重要，同时，数据预处理也是数据挖掘中的一个重要研究课题。

整个挖掘过程是一个不断重复的过程。

假如用户在挖掘过程中发现选择的属性或数据有偏差，或者使用的挖掘技术产

生不了预期的结果，这时就需要根据反馈结果，不断重复先前的过程，甚至从头重新开始，最终得到令人满意的挖掘结果。

可视化在数据挖掘的各个阶段都扮演着重要的作用。

在数据准备阶段，用户可能要使用数据立方体、直方图等可视化统计技术来显示有关数据，以便对数据有一个初步的理解，从而为更好地选取数据打下基础。在挖掘阶段，用户有可能要使用与领域问题有关的可视化工具，来选择挖掘算法或者调整挖掘算法的参数。在结果表示阶段，则可能又要用到其它的可视化技术，以利于用户对挖掘结果的理解。

2.2. 数据挖掘概念

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。

2.3. 数据挖掘功能

数据挖掘功能即是从指定数据中发现模式类型。

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”就不能成为一个模式[10]。

数据挖掘任务一般分为两类：描述型模式或预测型模式。

描述型挖掘任务描述数据库中数据的一般特性。描述型数据挖掘以简洁概要的方式描述数据，并提供数据的有趣的一般性质，或将它与对比类相区别。

预测型挖掘任务在当前数据上进行推断，以进行预测。预测型数据挖掘分析数据，建立一个或一组模型，并试图预测新数据集的行为[1]。

根据数据挖掘的任务不同，数据挖掘发现的模式可以分为以下几类：

- ✓ **分类模式**

分类的目的是通过机器学习，产生一个分类函数或分类模型(分类器)，该模型把数据集中的数据项映射到给定类别中的某一个。分类和回归都可用于预测。预测的目的是从利用历史数据纪录中自动推导出对给定数据的推广描述，从而能对未来数据进行预测。和回归方法不同的是，分类的输出是离散的类别值，而回归的输出则是连续数值。分类模式通常表现为一棵分类树。

- ✓ **回归模式**

回归模式的函数定义与分类模式相似，差别在于分类模式的预测值是离散的，而回

归模式的预测值是连续的。如给出一个顾客信用信息的数据库，可以通过分类模式判定他的信誉度是优良还是一般；给出一个人的教育背景、工作经验，可以使用回归模式判定此人的年工资在哪个范围内，5000 以下还是 5000 到 1 万元之间，或者在 1 万以上。

- ✓ **聚类模式**

聚类分析对一个数据对象的集合进行分析，与分类不同的是它要划分的类是未知的。聚类将数据对象分组成多个类或簇(Cluster)，在同一个簇中对象之间具有较高的相似度，而不同簇中的对象差别较大。

- ✓ **关联模式**

关联模式是数据项之间的关联规则。关联规则挖掘发现大量数据中项集之间的有趣的关联或相关联系。关联规则是如下形式的一直规则：“在 20－29 岁，年收入 20K－29K 的顾客中在 AllElectronics 购买 CD 机的可能性为 60％”。

- ✓ **时序模式**

时序模式根据数据随时间变化的趋势来预测未来的值。

- ✓ **序列模式**

序列模式挖掘相对时间或其他模式出现频率高的模式，序列数据由有序的事件序列组成，它可以有时间标记，也可以与时间无关。例如，Web 页面遍历序列就是一种序列数据，但可能不是时序数据。一个序列模式的例子是“9 个月以前购买奔腾 PC 的客户很可能在一个月内订购新的 CPU 芯片”。

在解决实际问题时，可能要综合多种挖掘模式。分类模式、回归模式、时序模式在建立模式之前数据的结果时已知的，可直接用来检测模式的准确性。一般在建立这些模式时，使用一部分数据作为样本，用另一些数据来检验模式。聚类模式、关联模式、序列模式在模式建立前结果是未知的，模式的产生不受任何监督。

2.4. 数据挖掘的方法和技术

- ✓ **人工神经网络(Artificial Neural Network)**

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、特征挖掘等多种数据挖掘任务。

- ✓ **遗传算法(Genetic Algorithm)**

基于生物进化的概念设计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，需要把数据挖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。

- ✓ **决策树方法(Decision Tree)**

用树形结构表示决策集合，这些决策集合通过对数据集的分类产生规则。典型的决策方法有分类回归树(CART)，一般用于分类规则挖掘。

✓ **最近邻技术(Nearest Neighbor)**

通过 k 个与之最相近的历史记录的组合来辨别新的记录。这种技术可以用作聚类、偏差分析等挖掘任务。

✓ **规则归纳(Rule Induction)**

通过统计方法归纳、提取有价值的 if… then…规则。规则归纳技术在数据挖掘中广泛使用，例如关联规则的挖掘。

✓ **可视化(Visualization)**

采用直观的图形方式将信息模式、数据关联或趋势呈现给决策者，决策者可以通过可视化技术直观地分析数据关系。

在数据挖掘和知识发现中应用的人工智能技术还有模糊逻辑、公式发现、统计分析方法、粗糙集方法等。

2.5. 数据挖掘的分析方法

✓ **关联分析**

关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系[1]。关联分析发现关联规则。关联规则是形如 A=>B 的规则，$A \subset I$，$B \subset I$，$A \cap B = \varnothing$，$I = \{i_1, i_2, \cdots, i_m\}$是项的集合。

例如,购买计算机也趋向于同时购买财务管理软件可以用以下关联规则表示：

computer=>financial_management_software[support=2%,confidence=60%]

关联规则中的 2%表示分析中的全部事务的 2%（支持度）同时购买计算机和财务管理软件。置信度 60%表示购买计算机的顾客 60%（置信度）也购买财务管理软件。

✓ **分类分析**

分类是这样的一个过程，它找出描述并区分数据类或概念的模型（或函数），以便能够使用模型预测类标记未知的对象类。举例来说，信用卡公司的数据库中保存用户的记录，并根据信誉程度（类标记），将用户分为三类：良好，普通，较差。这一过程实际将用户记录标记为三类，分类分析法检查这些记录，然后给出一个对信誉等级的显式描述：

“信誉良好的用户是那些年收入在 20000 美元以上，年龄在 45 到 55 岁之间，居住在 XX 地区附近的人士”。

✓ **聚类分析**

聚类将物理或抽象对象的集合分组成由类似的对象组成的多个类或簇，在同一各簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。与分类不同的是，它要划分的类是未知的。聚类的目的是根据一定规则，合理划分记录集合，并用显示或隐式的方法描述不同的类别。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由不同的划分。