摘 要

数据挖掘是近年来随着数据库技术和人工智能技术的发展而出现的一种新的信息技术，它融合了数据库、人工智能以及统计学等多种学科，试图从数据库中提取出先前未知、有效和实用的知识。Web 数据挖掘是传统数据挖掘技术在 Web 环境下的应用，是从 Web 上的数据（如 Web 日志、页面内容等）中发现用户的浏览模式或寻找相关的 Web 页面等。Web 使用挖掘主要是对于 Web 日志数据进行分析处理。而 Web 日志数据通常是大量的，冗余的，日志中的页面之间的关系也是模糊的，不确定的。粗糙集理论是有效处理不精确、不确定和含糊信息的软计算工具，模糊聚类分析是依据客观事务间的特征、亲疏程度和相似性，通过建立模糊相似关系对客观事务进行分类的方法。Web 使用挖掘可以从网站的日志数据中抽取用户感兴趣的模式，理解用户的浏览兴趣行为，以便进一步改善网站结构，为用户提供个性化服务。所以本文提出的粗糙集理论和聚类算法在 Web 使用上的挖掘研究是具有一定的理论意义和现实意义的。

论文首先介绍了数据挖掘和 Web 数据挖掘的基本理论和方法；其次介绍了粗糙集理论和模糊聚类算法；再结合具体 Web 日志数据提出了 Web 使用挖掘的方法及 Web 日志数据模型，并建立了基于模糊聚类算法的页面用户聚类的一般模型。在第五章中进一步利用上述理论对 Web 日志数据进行预处理，并利用粗集理论对预处理结果中的教务网页面约简，得到在不影响问题分析基础上的有效页面。最后利用模糊等价关系矩阵和图的模糊聚类方法，在上述数据处理的基础上对其进行了进一步的分析研究。论文利用 Java 语言结合教务网数据源实现了算法编程。

关键词：

数据挖掘，Web 使用挖掘，粗糙集，模糊聚类

Abstract

Data Mining is a new information technology that has been developed with the technology of Database and Artificial Intelligence, which integrates of Database, AI and Statistics and etc.It tries to extract the unknown, effective and useful knowledge from database. Web Mining is the traditional Data Mining technology application used in web, which can extract user's browse pattern and find the relative web pages from data (such as web log, web page content) on eb. Web Usage Mining mainly processes and analyses the web log data which is generally redundancy. Moreover the relations among the web pages are fuzzy and uncertain.Rough Sets theory is a soft computing tool dealing with vague, imprecise, uncertain and incomplete data. And Fuzzy Clustering Analysis is an analysis method of object through establishing fuzzy analogical relations based on the character, distance and similarity among objects. Web Usage Mining can get the interesting pattern from the log of websites, and apprehend the user's browse interest behavior, so as to improve the website's structure and provide individual services for the users.So the research into "Rough Sets Theory and Fuzzy Clustering Algorithm" is a research of theoretical significance and realistic value.

Firstly, principle theories and methods of Data Mining, Web Data Mining, Rough Sets and Fuzzy Clustering Algorithm theory are introduced. Then method of Web Usage Mining and model of Web Log data are established through actual Web Log data. The page-user clustering's general model based on Fuzzy Clustering Algorithm is put forward as well. Furthermore, based on the educational administration's website of our university, the primal Web Log data is pretreated through the above theories. And reduction of the web pages are gained, which doesn't affect the analysis. Finally, the result which is got through fuzzy equivalence matrix and fuzzy clustering method of graph is analyzed and research in futher depth. The Algorithm is realized in Java Language.

Keywords:

Data Mining, Web Usage Mining, Rough Sets, Fuzzy Clustering

1. 绪 论

1.1 论文的研究背景及选题意义

在 Internet 浪潮的冲击下，人们面临着数据爆炸的挑战；随着数据挖掘（Data Mining，DM）技术的迅速发展及数据库管理技术的广泛应用，人们积累的数据越来越多，如何从浩如烟海的数据中找到内在的规律，如何更方便地传递、交流、获取有用的信息，挖掘这些激增数据背后隐藏的重要信息已成为当前高科技领域研究的热点。

目前，互联网已经和我们的生活密不可分，它可以说是一个巨大的、分布广泛和全球性的信息服务中心。它涉及新闻、广告、消息信息、金融信息、教育、政府、电子商务和许多其他信息服务。根据有关机构统计，目前互联网的数据以几百兆兆字节来计算，而且增长速度很快，如果将这些庞大的数据用一般的统计分析来处理的话，显然是有心无力的。自从数据挖掘技术成功地应用于传统数据库领域之后，人们对于数据挖掘在像互联网数据这样的一些特殊数据源的应用也寄予了厚望，并且做了许多相应的研究和发展了相应的技术。

Web 挖掘（Web Mining），从广义上解释就是通过数据挖掘技术来分析与网站相关的资料，例如：网站浏览记录（Web Log）、网页内容（Web Content）、网络链接结构（Web Structure）等。随着 Web 技术的发展，各种网站数量每天都在激剧增加，特别是将 Web 转化为关键发展工具的信息网站（电子商务网站），采用各种手段使网站更加具有生命力成为每个经营者的首要工作。因此及时了解用户的需求和特点，为每个用户提供个性化、智能化的服务，以吸引大量的用户，就变得尤其重要。其中网站浏览记录，记录了使用者每次访问网站的一些资料，它最能反映使用者对网站的浏览需求。所以以数据挖掘技术来分析网站浏览记录，已成为解决上述问题的主要工具之一。

如何从数以亿计的页面中发现需要的内容，如何合理有效的组织网站的拓扑结构，如何将用户网页进行聚类，以提供个性化服务成了人们迫切希望解决的问题，尤其是对于电子商务网站来说更是如此。Web 使用挖掘是从 Web 使用数据，即网页被用户使用的记录文件 Web Log 中抽取感兴趣的模式的过程。分析这些数据可以帮助网站管理者理解用户的行为，得到用户群体普遍的访问行为模式和用户个体的访问模式，从而

根据这种模式为用户定制合适的推荐页面。

Rough Sets（粗糙集简称 RS）理论是由波兰华沙理工大学 Pawlak 教授于 20 世纪 80 年代初提出的一种研究不完整、不确定知识和数据的表达、学习、归纳的理论方法。随着知识发现的兴起，粗糙集理论也受到众多研究者的重视进而受到研究界的广泛注意，它为知识发现、数据挖掘提供了一种新的方法和工具，能很好解决其中的数据多样、数据冗余、噪声数据和不确定性、大规模数据等问题。属性约简是粗糙集的核心内容之一。在处理二维表形式的信息决策表中，它可以在不影响其分类能力的前提下进行属性约简，进而简化数据表的分析处理，提高知识发现的能力和效率。近年来，粗糙集理论广泛应用于数据挖掘中，极大地提高了数据挖掘的数据预处理能力和效率。

同传统的数据挖掘一样，Web 挖掘主要包括聚类、关联规则分析和序列分析。聚类分析已经广泛应用于市场分析，通过聚类可以从客户基本数据库中发现不同的客户群，刻画不同客户群的特征。然而 Web 日志中存在着许多的噪音数据和信息的不完整性，这导致聚集只是一个模糊的边界，聚集中的对象成员没有一个精确的定义。这样对象成员就有可能不只属于一个聚类。为解决这一问题，我们采用模糊理论，它主要是研究模糊现象、不精确性现象的数学工具。将模糊集理论中的模糊聚类应用到 Web 挖掘，分析用户访问 Web 的模式，将特性相同页面聚在一起，为网站设计者提供一个参考的理论平台，改进网站的设计，从而更好的吸引用户，增加企业的竞争力。在数据预处理中，对于噪音数据，我们采用粗集理论来对其进行处理。

1.2 国内外研究现状

目前，国内外有关 Web 用户模式挖掘的研究主要集中在对用户浏览路径进行挖掘[1]。

文献[1]提出了一个基于代理的 Syskill&Webert 软件，该软件通过分析用户访问日志预测哪些页面是用户感兴趣的页面。文献[2]设计了一个个性化的新闻站点，结合用户的反馈信息提供个性化服务。文献[3]提供了两种个性化网站的方案机器实现，包括了用户定制和计算用户兴趣度的方法。文献[4]介绍的 WebACE 系统，采用了分类算法来划分用户的上网访问的历史记录信息，划分出每一个分类代表用户在这方面的一个兴趣。文献[5]设计了一个 Web 挖掘算法，利用模糊集理论挖掘用户

浏览模式。

Web 使用挖掘的研究多应用于推荐系统，提供个性化网站，动态自适应网站的建造等。对于个性化定制服务，目前常用的方法包括 Web 使用记录挖掘与合（协）作式过滤、矩阵聚类[6]、挖掘间接关联规则[7]、数据立方体[8]、第一马尔可夫传输链[9]、All-mth-Order Markov Model[9]、Prediction by Partial-Match[10]、概念分层[11]、Inter-based Coarsening[11]等等。

1.3 论文的研究内容和组织结构

数据挖掘是数据处理的一个新的热点和前沿领域，它的研究目标是采用有效的算法，从大量现有的数据集合中发现并找出最初未知，但最终可理解的有用知识，并用简明的方式显示出来。Web 使用挖掘是 Web 数据挖掘研究的一个重要方向，也是本文研究的重点。本文的工作主要有以下几个方面：

(1)在阅读大量文献的基础上，分析了数据挖掘技术、Web 挖掘技术、Web 使用挖掘技术及其应用和研究方向。

(2)介绍了粗糙集理论及其在数据挖掘中的应用，粗糙集的几个约简算法。

(3)介绍了模糊聚类概念及其几个模糊聚类算法。

(4)建立了 Web 日志数据的数学模型，并提出了在 Web 使用挖掘中如何进一步应用粗集理论和模糊聚类算法。

(5)结合我校教务网日志数据，利用粗糙集理论首先进行数据预处理处理，再利用模糊聚类算法对页面聚类。

(6)利用 Java 语言实现了数据预处理，并利用粗糙集国内软件 Ridas 和国外软件 Rosetta 对冗余页面约简，并利用 Java 编程实现了模糊聚类算法。

论文的结构如下：

第二章介绍了数据挖掘、Web 数据挖掘基本理论、方法、研究现状和未来研究方向。第三章介绍了粗糙集理论概念及其几个约简算法，模糊聚类基本概念和几个模糊聚类算法。

第五、六、七章是本论文的重点，也是本文的主要工作。提出了基于粗集理论和模糊聚类算法的 Web 使用挖掘，并结合我校教务网的日志数据进行数据预处理，最后利用模糊聚类算法进行用户页面的聚类。

2. 数据挖掘、Web 数据挖掘

2.1 数据挖掘概述

近十几年来，人们利用信息技术生产和搜集数据的能力大幅度提高，成千上万个数据库被用于商业管理、政府办公、科学研究和工程开发等等，并且这一势头仍将持续发展下去。在这被称之为信息爆炸的时代，信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率呢?面对数据爆炸、知识贫乏的挑战，数据挖掘和知识发现(DMKD)技术应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。

2.1.1 数据挖掘的概念

数据挖掘(Data Mining，DM)有多种术语[12]，如“知识发现”(Knowledge Discovery in Database, KDD)，“知识抽取”(Information Extraction)，“信息发现”(Information Discovery)，“智能数据分析”(Intelligence Data Analysis)，“信息收获”(Information Harvesting)，“数据考古”(Data Archeology)等。

从技术上定义，数据挖掘(DM)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。这个定义包括好几层的含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

从商业的角度定义，数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。也可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进有效的方法。

数据挖掘的原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的;可以是演绎的，也可以是归纳的。发现了的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。因此，数优化、决策支持、过程控制等，还可以用于数据自身的维护。因此，数

据挖掘是一门很广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

数据挖掘原理图,如下图所示：

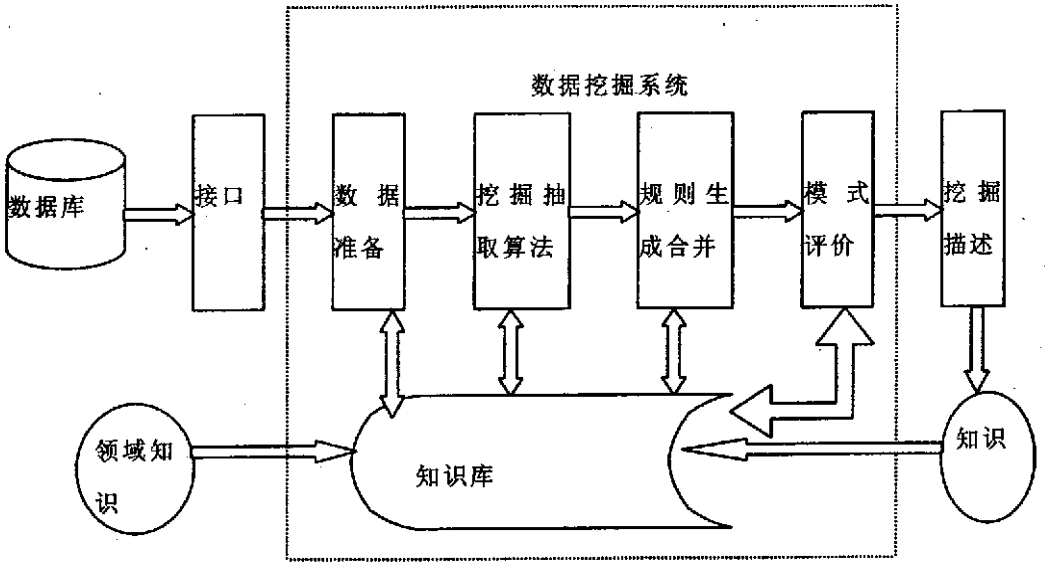

图 2-1 数据挖掘系统逻辑原理图

2.1.2 数据挖掘的研究现状

数据挖掘即从数据库中发现知识(KDD),最早是 1989 年 8 月在美国底特律市召开的第十一届国际联合人工智能学术会议上正式形成的[13]。刚开始每两年召开一次国际 KDD 学术会议，93 年以后每年举行一次 KDD 国际学术会议，把对数据挖掘和知识发现的研究推入高潮。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。由于把数据库中的“数据”形象地比喻成矿床，“数据挖掘”一词很快流传开来。1995 年以来，国外在数据挖掘方面论文非常多，已形成了热门研究方向。还有一些其它国际或地区性数据挖掘会议，如“知识发现和数据挖掘太平洋亚洲会议”(PAKDD)，“数据库中知识发现原理与实践欧洲会议”(PKDD)，“数据仓库与知识发现国际会议”(DaWaK)等。涉及数据挖掘和数据仓库的研究结果已在许多数据库国际学术会议论文集发表，包括“ACM-SIGMOD 数据管理国际会议”(SIGMOD)，“超大型数据库国际会议”

(VLDB)，“ACM-SIGMOD-SIGART 数据库原理研讨会”(PODS)，“数据工程国际会议”(ICDE)，“扩展数据库技术国际会议”(EDBT)，“数据库理论国际会议”(ICDT)，“信息与知识管理国际会议”(CIKM)，“数据库与专家系统应用国际会议”(DEXA)，“数据库系统高级应用国际会议”(DASFAA)等。这些国际研讨会规模由原来的专题讨论会发展到国际学术大会，研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。

Robert Grossman 提出了数据挖掘系统划分为四代的观点。归纳成下表可以看出四代是基于技术的划分[14]。KDnuggets主席 Gregory Piatetsky-Shapiro 的观点认为数据挖掘软件发展的三个阶段为：独立的数据挖掘软件、横向的数据挖掘工具集、纵向的数据挖掘解决方案[14]。

数据挖掘软件对比如下表：

表 2-1 数据挖掘软件发展对比表

	特征	数据挖掘算法	集成	分布计算模型	数据模型	软件代表
第一代	作为一个独立的应用	支持一个或者多个算法	独立的系统	单个机器	向量数据	Salford System 公司的 CART(http://www.slford-systrems.com/)
第二代	和数据库、数据库管理系统(DBMS)以及数据仓库集成	多个算法：能够挖掘更复杂的数据集	数据管理系统，包括数据库和数据仓库	同质、局部区域的计算机群集	有些系统支持对象，文本和连续的媒体数据	上海复旦德门软件公司 DBMine(http://www.dbminer.com.cn/)，SAS Enterprise Miner)
第三代	和语言模型系统集成	多个算法	数据管理和语言模型系统	Intranet/Extranet 网络计算	支持半结构化数据和 Web 数据	SPSS Clementine(以 PMML 的格式提供与预言模型系统的接口)
第四代	和移动数据/各种计算设备的数据联合	多个算法	数据管理、预言模型、移动系统	移动和各种计算设备	普遍存在的计算模型	尚为出现

目前，随着新的挖掘算法的研究和开发，第一代数据挖掘系统仍然会出现，第二代系统是商业软件的主流，部分第二代系统开发商开始研制相应的第三代数据挖掘系统，比如 IBM Intelligent Score Service。第四代数据挖掘原型或商业系统尚未见报导，PKDD2001 上 Kargupta 发表了一篇在移动环境下挖掘决策树的论文，Kargupta 是马里兰巴尔的摩州立大学(University of Maryland Baltimore County)正在研制的 CAREER 数据挖掘项目的负责人，该项目研究期限是 2001 年 4 月到 2006 年 4 月，目的是开发挖掘分布式和异质数据(Ubiquitous 设备)的第四代数据挖掘系统。

另外不仅如此，在 Internet 上还有不少 KDD 电子出版物，其中以半月刊 Knowledge Discovery Nuggets 最为权威，如要免费订阅，只需向 http://www.kdnuggets.com/subscribe.html发送一份电子邮件即可，还可以下载各种各样的数据挖掘工具软件和典型的样本数据仓库 ，供人们测试和评价。在http://www.kdnuggets.com/上还能发现有许多关于数据挖掘的书和软件，国内数据挖掘讨论组 http://www.dmgroup.org.cn/上也有丰富的资源。

2.2 数据挖掘的特点、方法及过程

2.2.1 数据挖掘的特点

数据挖掘与传统的数据库查询区别表现在：前者是主动的、不生成严格的结果集和不同层次的挖掘，而后者则是被动的、只对字段进行严格的查询。归纳起来，数据挖掘有如下特点：

1）处理的数据规模十分庞大；

2）由于用户不能形成精确的查询要求，因此需要依靠数据挖掘技术来寻找其可能感兴趣的东西；

3）数据挖掘对数据的迅速变化做出快速响应，以提供决策支持信息；

4）数据挖掘既要发现潜在规则，还要管理和维护规则，随着新数据的不断加入，规则需要随着新数据更新；

5）数据挖掘中规则的发现基于统计规律，发现的规则不必适合所有数据，而且当达到某一阈值时，便认为有此规则。

2.2.2 数据挖掘的方法

数据挖掘的方法可粗分为：

1）统计方法

统计方法细分为：回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类等)、探索性分析(主元分析法、相关分析法等)、以及模糊集、粗糙集、支持向量机等。

2）机器学习方法

机器学习可细分为：归纳学习方法(决策树、规则归纳等)、基于范例的推理 CBR、遗传算法、贝叶斯信念网络等。神经网络方法，可细分为：前向神经网络(BP 算法等)、自组织神经网络(自组织特征映射、竞争学习等)等。

3）数据库方法

数据库方法主要是基于可视化的多维数据分析或 OLAP 方法，另外还有面向属性的归纳方法。

2.2.3 数据挖掘的过程

一个数据挖掘系统不是多项技术的简单组合，而是一个完整的整体，它除了核心技术以外，还需要其他辅助技术的支持，才能完成数据挖掘的过程，最后将分析结果呈现在用户面前。数据挖掘的数据分析过程如下：

1）数据准备(Data Preparation)

数据准备又可分为数据集成(integration)、数据选择和预分析(Data selection and pre-analysis)。数据集成将从操作型环境中提取并集成数据，解决语义二义性问题，消除脏数据等。数据选择和预分析将负责缩小数据范围，提高数据挖掘的质量。

2)挖掘(Mining)

数据挖掘处理器(data mining processor)综合利用前面提到的各种数据挖掘方法分析数据。

3）表述(Presentation)

与检验型工具一样，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。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。评价数据挖掘工具的主要指标有：数据准备、数据访问、算法与建模、模型评价和解释、用户界面。

2.3 数据挖掘的发展及应用

2.3.1 数据挖掘未来研究方向

数据挖掘未来研究方向：与数据库数据仓库系统集成；与语言模型系统集成；挖掘各种复杂类型的数据；与应用相结合研制和开发数据挖掘标准；支持移动环境。

2.3.2 数据挖掘的应用

数据挖掘已广泛的应用于银行金融市场、零售业、医疗业等各行业。其应用行业表如下：

表 2-2 数据挖掘应用行业表

行 业	数 据 挖 掘 应 用
银行金融保险业	信用评估、客户定制化金融服务、授信利用率额度、客户资产管理、坏帐分析、道德危机分析、逆向选择风险分析、潜在客户名单分析、客户区域分隔、交叉销售、连续销售、设点区域分析等
零售业	即时辅助购买决策、会员客户营销、连续销售、促销商品组合、库存分析、货品、商品排架、物流整合及配置辅助决策广告业、客户反馈率提升、市场区隔、促销商品组合等
医疗业	成本分析、研究分析、预防医学分析、院内感染分析、临床病因分析等
生物技术业	基因图谱比对、基因序列分析、演化分析等
连锁店业	设点区位分析、库存分析、成本分析等
制造业	生产质量分析、原物料库存管理、半成品（再制品）库存管理、销售分析、成本分析、生产调度等
电信业	系统最优化、客户区分、客户反馈率提升、费率制定、客户定制化行销等
证券投资业	股票市场预测、客户反馈率提升、客户定制化行销等
航空业	客户区隔、客户反馈率提升、航段促销组合、成本分析、客户定制化行销等
教育业	学生招募、市场区分、学生来源分析、建议课程顺序、学习评价度量、学生生涯规划等

体育竞技类	队员替换策略、战术组合优化策略等
海关部门	提高查获率、打击价格瞒骗行为等
科学研究	公式推导与发现、知识发现与信息处理、知识管理等

2.4 Web 数据挖掘

近年来，随着 Internet 在全球范围的迅速普及和应用，网络日益成为人们生活、工作的重要组成部分。据估计，网络已经发展成为拥有 3 亿页面的分布式信息空间，而且这些信息仍以每 4 到 6 个月增长一倍的速度增加。在这些大量、不同的各类 WEB 信息数据中，蕴含着巨大潜在价值的信息，如何快速、有效地发现这些信息资源已成为急待解决的问题。

2.4.1 概述

Web 数据挖掘是指使用数据挖掘技术从 Web 文档及 Web 服务中自动发现并提取潜在的、有用的模式或信息，其原理图如下所示。与传统的数据挖掘相比，Web 数据挖掘有自身的特点：挖掘对象是海量的、异构的、分布的文档；Web 在逻辑上是一个由文档节点和超链接构成的图；Web 文档是半结构化或无结构的，且缺乏机器理解的语义。因此，传统数据挖掘并不能直接应用于 Web 数据挖掘，需要对 Web 文档进行一定的处理。Web 挖掘研究覆盖了多个研究领域，包括国际互联网、计算机语言学、数据库、信息获取、统计学、人工智能中的机器学习和神经网络等领域。

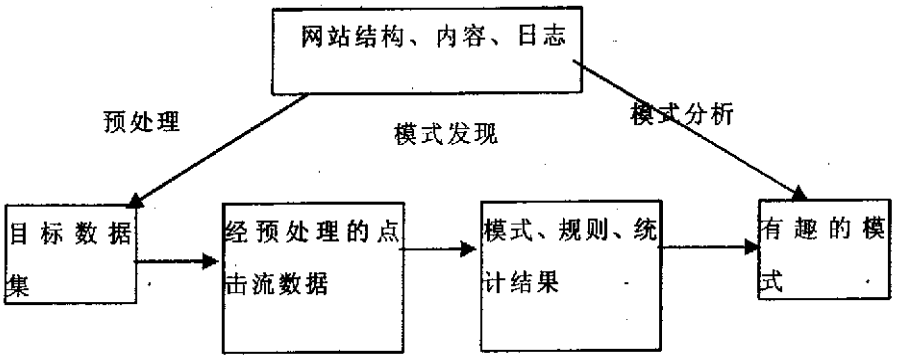

图 2-2 Web 数据挖掘原理图

2.4.2 Web 数据挖掘的难点

由于 Web 具有开放性、动态性与异构性等固有特点，所以如何从这些分散的、异构的、没有统一管理的海量数据中快速准确地获取信息成

为 Web 挖掘所要解决的一个难点，也使得用于 Web 的挖掘技术不能照搬数据库的挖掘技术。Web 数据挖掘的难点表现在如下几个方面：

(1) 数据来源分析

在对网站进行数据挖掘时，所需要的数据主要来自三个方面：Web 服务器中的日志文件、Web 服务器中的其他信息以及客户的背景信息。

(2) 异构数据环境

从数据库研究的角度出发，Web 网站上的信息也可以看作是一个更大、更复杂的数据库。Web 上的每一站点就是一个数据源，每个数据源都是异构的，因而每一站点之间的信息和信息的组织不一样，这就构成了一个巨大的异构数据库环境。

(3) 半结构化的数据结构

Web 上的数据和传统的数据库中的数据不同：传统的数据库都有一定的数据模型，可以根据模型来具体描述特定的数据；而 Web 上的数据非常复杂，没有特定的模型描述。

(4) 解决半结构化的数据源问题

Web 数据挖掘技术首先要解决半结构化数据源模型和半结构化数据模型的查询与集成问题。针对 Web 上的数据半结构化的特点，寻找一个半结构化的数据模型是解决问题的关键所在。

2.4.3 Web 数据挖掘的分类[13]

Web 数据主要来自于三个方面：Web 服务器中的日志文件、Web 服务器中的其他信息以及客户的背景信息。归纳起来，Web 数据有三种类型：HTML 标记的 Web 文档数据、Web 文档内的连接的结构数据和用户访问记录数据如服务器的 log 日志信息。按照对应的数据类型，Web 挖掘可分为：Web 内容挖掘、Web 结构挖掘、Web 使用挖掘(即用户访问模式挖掘)(如图所示)，而 Web 内容挖掘和用户访问模式挖掘是 Web 挖掘的两个主要方面。

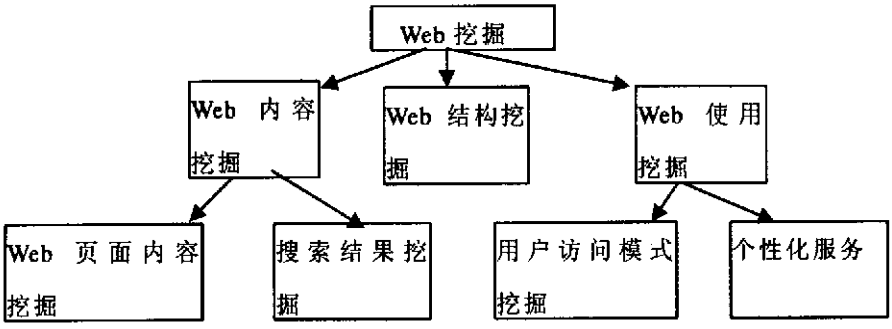

图 2-3 Web 挖掘分类

1、Web 内容挖掘

Web 内容挖掘是从文档内容或描述中抽取知识的过程。Web 上的内容挖掘多为基于文本信息的挖掘和基于多媒体文档(包括 image、audio、video)的挖掘。基于文本信息的挖掘是利用 Web 文档中部分标记，如 Title、Head 等包含的额外信息，可以提高 Web 文本挖掘的性能。多媒体挖掘主要是针对 Web 中音频、视频、图形、图像数据进行相应的处理，采用改进的数据挖掘方法发现蕴含在里面的潜在的、有意义的信息和模式的过程。

许多基于数据仓库的挖掘算法经过相应的改进处理就可以用于文本的挖掘。比如数据归纳、分类、聚类、关联规则的挖掘等。Web 文本的挖掘对象可以是结构化的、也可以是半结构化的和非结构化的。挖掘的结果是对某个文本文件的概括和总结，也可以是对整个文本集的分类或聚集的结果。

Web 上的内容挖掘实现技术主要有：文本总结、文本聚类和关联规则[14]。

2、Web 结构挖掘

Web 结构挖掘是对 Web 页面之间的结构进行挖掘，从 WWW 上的组织结构和链接关系中推导知识。Web 结构挖掘主要针对的就是页面的超链接结构，如果有较多的超链接指向它，那么该页面就是重要的，发现的这种知识可用来改进搜索路径等。

挖掘 Web 结构主要是通过对 Web 站点的结构进行分析、变形和归纳，将 Web 页面进行分类，以利于信息的检索。其目的是发现页面的结构和 Web 结构，在此基础上对页面进行分类和聚类从而找到权威页面。Page-Rank 方法(Brine and Page 1998)就是利用文档之间链接信息来查找相关的 Web 页。

Page-Rank 的基本思想是：一个页面被多次引用，则这个页面很有可能是重要的；一个页面尽管没有被多次引用，但被一个重要的页面引用，该页面也可能是很重要的；一个页面的重要性被均分并被传递到它所引用的页面中。如对于一个查询 q，搜索引擎首先利用相似度函数找到 K 个页面，然后利用公式计算每个页面的重要性，然后进行排序。

3、Web 使用挖掘

（1）概述

Web 使用挖掘即 Web 使用记录挖掘，通过挖掘相关的 Web 日志记录，来发现用户访问 Web 页面的模式，通过分析日志记录中的规律，可以识别用户的忠实度、喜好、满意度，可以发现潜在用户，增强站点的服务竞争力。Web 使用记录挖掘是通过挖掘 Web 日志记录，来发现用户访问 Web 页面的模式。还可以通过分析和探究 Web 日志记录中的规律，来识别电子商务的潜在客户，增强对最终用户的互联网信息服务的质量，并改进 Web 服务器系统的性能。

Web 使用记录数据除了服务器的日志记录外还包括代理服务器日志、浏览器日志、注册信息、用户会话信息、交易信息、cookie 中的信息、用户查询、鼠标点击等一切用户与站点之间可能的交互记录。

（2）Web 日志数据格式

目前市面上比较流行的 Web 服务器，例如 IIS，Apache 等，通常都保存了对 Web 页面的每一次访问的日志项，这些记录项即 Weblog 项。它忠实地记录了访问该 Web 服务器的数据流的信息。日志格式如表所示：

表 2-3 服务器日志格式

域（field）	描述（description）
日期（date）	请求页面的时间、日期和时区(date, time and time zone of request) 例：[26/Apr/2003:03:04:41-0500]
客户端 IP（client IP）	远程主机的 IP 或者 DNS 入口(remote host IP and/or DNS entry)
用户名（user name）	远程登录的用户名(remote log name of the user)
字节（bytes）	发送和接收的字节（bytes transferred and received）
服务器(server)	服务器、IP 地址和端口(server name , IP address and Port)
请求(request)	URL 查询和枝节(URL query and stem)
状态(status)	返回给 HTTP 状态标识（http status code returned to the client）
服务名(service name)	用户请求的服务名称（request and service name）

耗用时间(time taken)	完成浏览的时间(time taken for transaction to complete)
协议版本(protocol version)	传输用的协议版本（version of used transfer protocol）例："Get C.html HTTP/1.0"
用户代理(user agent)	服务提供者(service provider)例：Moz（Win98）
Cookies	标识号(cookies ID)
参照页(referrer)	本页的上一页

日志文件记录的内容还可以根据客户的不同需要，来调整记录的信息。例如 IIS5.0 中 W3C 扩展日志文件格式中，除了时间这些日志文件肯定有的元素外，还有多达 19 项可以选择记录的扩展属性，比较常用的属性是所请求的 URI 资源，客户端 IP 地址和时间戳。在 W3C 扩展日志文件格式中，缺省的属性有：时间戳，客户端 IP 地址，访问方法，URI 资源，协议状态。

（3）Web 使用挖掘分类

Web 使用挖掘可以分为两类：一类是将 web 使用记录的数据转换并传递进传统的关系数据库里，再使用数据挖掘算法对关系表中的数据进行常规挖掘；另一类是将 web 使用记录的数据直接预处理再进行挖掘。Web 使用挖掘中的一个有趣的问题是在多个用户使用同一个代理服务器的环境下如何识别某个用户，如何识别属于该用户的会话和使用记录，这个问题在很大程度上影响着挖掘质量。

2.4.4 Web 数据挖掘的研究方向

Web 挖掘是把 Internet、WWW 和数据挖掘结合起来的一种新兴技术，Web 挖掘的应用非常广阔，不但涉及页面信息的提取、站点的分析和设计，而且在蓬勃发展的基于 Internet 的电子商务方面也有很好的应用前景。目前，在国内 Web 挖掘的研究仍处于起步阶段，是前沿性的研究领域。今后几年 Web 挖掘研究的主要方向有[16]：

1）在数据预处理方面，多种 Web 数据的收集、结构转换等处理技术的研究；
2）Web 挖掘方法和模式识别技术在构造自适应站点以及智能站点服务的个性化和性能优化方面的研究；
3）Web 知识库的动态维护、更新，各种知识和模式的评价综合方法的研究；
4）基于 Web 挖掘和信息检索的，高效的、具有自动导航功能的智能搜索引擎相关技术的研究；

5）半结构、结构的文本数据、图形图像数据、多媒体数据的高效挖掘算法；

6）研究专门用于知识发现的数据挖掘语言及其标准化；

7）研究和开发基于 Web 的多层数据体系结构和智能集成系统，提供相应的查询语言，优化和维护机制；

8）现有的数据挖掘方法与技术的改进及其向 Web 数据的扩展，挖掘算法的适应性和时效性的研究；

9）Web 文档内的模式发现及其在信息提取、文本分析中的应用研究等；

10）Web 挖掘的相关技术在电子商务领域的应用研究等。

2.5 Web 使用挖掘

2.5.1 Web 使用挖掘的分类

若根据数据来源、数据类型、数据集合中的用户数量、数据集合中的服务器数量等又可将 Web 使用挖掘分为五类：

个性挖掘：针对单个用户的使用记录对该用户进行建模，结合该用户基本信息分析他的使用习惯、个人喜好，目的是在电子商务环境下为该用户提供与众不同的个性化服务。

系统改进：Web 服务的性能和其他服务质量是衡量用户满意度的关键指标，Web 用法挖掘可以通过用户的拥塞记录发现站点的性能瓶颈，以提示站点管理者改进 Web 缓存策略、网络传输策略、流量负载平衡机制和数据的分布策略。此外，可以通过分析网络的非法入侵数据找到系统弱点，提高站点安全性，这在电子商务环境下尤为重要。

站点修改：站点的结构和内容是吸引用户的关键。Web 用法挖掘通过挖掘用户的行为记录和反馈情况为站点设计者提供改进依据，比如页面连接情况应如何组织、那些页面应能够直接访问等。

智能商务：用户怎样使用 Web 站点的信息无疑是电子商务销售商关心的重点，用户一次访问的周期可分为被吸引、驻留、购买和离开四个步骤，Web 使用挖掘可以通过分析用户点击流等 Web 日志信息挖掘用户行为的动机，以帮助销售商合理安排销售策略。

Web 特征描述：通过分析用户对站点的访问情况，统计各个用户在页面上的交互情况，对用户访问情况进行特征描述。

2.5.2 Web 使用挖掘技术及预处理

Web 使用挖掘即使用模式挖掘，其常用的技术有 Web 使用的特有的路径分析技术和数据挖掘领域常用的关联规则、序列模式、分类聚类技术。

数据的预处理就是把数据源转化为适合进行用户模式挖掘的、准确可靠的规范数据，预处理结果的好坏会直接影响到数据挖掘的质量。一般来说，直接用于挖掘的数据是用户的会话数据库，比较精细的挖掘则使用事务数据库，也可使用会话子序列数据进行挖掘。

Web 访问模式挖掘的预处理包括内容预处理、结构预处理和数据文件预处理。其中内容预处理包含把文件、图像、脚本及其他文件（如多媒体）等转换成 Web 使用挖掘处理所需的数据格式。

数据预处理是在将日志文件转换成数据库文件以后进行的，其目的是把 Web 日志转化为适合进行数据挖掘的可靠精确的数据。这个过程主要包括五个阶段：数据清洗、用户识别、会话识别和路径补充（完善）、事务识别等[16] [17]。

1、数据清洗

数据清洗是指根据需求，对日志文件进行处理，包括删除无关紧要的数据，合并某些记录，并对用户请求页面时发生错误的记录进行适当的处理等等。

当用户请求一个网页时，与这个网页有关的图片、音频等信息会自动下载，并记录在日志文件中；而如果我们挖掘的目的是用户访问模式，这些信息对我们来说显然用处不大(除非图片、音频等是用户请求的，即用户所需要的内容正是这些图片和音频等文件)，所以可以把日志中文件的后缀为.gif、.jpg、.jpeg 等的记录删除。但是，当挖掘的目的是为了进行网络流量分析或为页面缓冲与预取提供依据时，这些信息又会显得格外重要，所以在删除这些记录的时候一定要把相关信息记录下来。我们选择将其中的“发送字节数”和“接收字节数”这两个域的内容记录下来。此外，后缀名为 cgi、js 和 JS 的脚本文件因对后面的分析处理不造成任何影响，所以应该删除。

2、用户识别

用户识别是将用户和请求的页面相关联的过程，主要处理多个用户通过代理服务器和防火墙访问站点的情况。如果进行用户访问模式的挖掘或对用户进行聚类分析，用户识别问题则显得至关重要，因为群体是由个体组成的，只有对个体有清楚的了解，才能识别群体的特征。

由于缓存、代理服务器和防火墙的使用，使得用户识别这一步变得很复杂：例如：不同的用户可以在同一时间通过一个简单的代理访问 Web 服务器；同一个用户可能在不同的机器上访问 Web 服务器；一个用户可能在同一台机器上使用不同的浏览器访问 Web 服务器，而当不同的用户使用同一台机器浏览某一站点时也会造成混淆。

因此可以设想：不同的 IP 地址代表着不同的用户；当 IP 地址相同时，默认不同的操作系统或浏览器代表不同的用户；在 IP 地址相同，用户使用的操作系统和浏览器也相同的情况下，则判断每个请求访问的页面与访问过的页面之间是否有链接，如果一个请求访问的页面与上一个已经访问过的所有的页面之间并没有直接的链接，则假设在访问 Web 站点的机器上同时存在多个用户。

3、会话识别

用户会话是指用户对服务器的一次有效访问，通过其连续请求的页面，可以获得他在网站中的访问行为和浏览兴趣。

日志文件中不同用户访问的页面当然属于不同的会话。当某个用户的页面请求在同一个网站，我们可以将用户的访问记录分成多个会话来处理。最简单的方法就是设置一个 timeout 值，如果用户访问页面的时间差超过了这个值，则认为用户开始了一个新的会话。许多商业产品都采用 30 分钟作为缺省的 timeout，但是 L.Catledge 和 J.Pikow[19] 由实验得出 timeout 值设为 25.5 分钟更好。会话识别的目的就是要创建每一个用户的有意义的页面聚类。

4、路径补充

路径补充(完整)就是将由于本地缓存或代理服务器缓存所造成的遗漏的请求页面补充完整。解决这一问题的方法类似于用户识别，如果一个页面请求信息与该用户上次请求的页面没有直接的链接关系，可以查看参考日志文件来决定这个页面来自哪个页面的链接，这样，一般可以修正由于本地高速缓存的使用而引起的当用户单击“Back”键时所产生的路径信息不完整的描述。同样也可以借助网络拓扑结构的信息，将服务器访问日志文件中一些未扫描的信息补充完整。

5、事务识别

事务识别是建立在对用户会话识别的基础上，依据数据挖掘任务的请求将事务做分割或合并的处理，以利于知识的发现。用户会话对数据挖掘而言仍不够精确，需要把会话进一步分解为具有一定语义的事务。

事务识别就是对用户会话进行语义分组。

Web 页可简单的分为两类，内容页和导航页。内容页一般是用户关心的信息，浏览时间较长；导航页可以看作是用户快速找到所需信息而设置的路标。划分方法可以采用页面所含超链接的多少为标准。页面分成两类以后，就有两种定义事务的方法：由多个导航页和一个内容页组成事务，导航页是到达内容页所经过的路径；事务仅由多个内容页组成。也可以采用 Chen[20]等人提出的最大前向引用路径(MFP)来定义事务，对每个用户会话，从开始页面为起点，每个最大前向引用路径既为一个事务。

3. Rough Sets 理论和模糊聚类算法

粗糙集理论在数据库知识发现(Knowledge Discovery in Databases，KDD)中显示了其特有的优势。KDD 所面对的实际数据可能包含各种噪声，存在许多不确定因素和不完备信息。与传统的不确定信息处理方法，如模糊论、证据理论和概率统计相比，粗糙集无需提供问题所需处理的数据集合之外的先验信息，对问题的不确定性的描述或处理更加客观，可以克服它们在面对大规模数据时的不足。粗糙集处理的对象是类似二维关系表的信息表(决策表)，目前成熟的关系数据库系统(KDD 研究实施的主要对象)以及新发展起来的数据仓库系统，为基于粗糙集方法的数据挖掘奠定了坚实的基础。运用粗糙集方法得到的知识发现算法有利于并行执行，可极大地提高挖掘效率。

3.1 Rough Sets 理论产生和发展

Rough Sets(粗糙集简称 RS)理论是由波兰华沙理工大学 Pawlak 教授于 20 世纪 80 年代初提出的一种研究不完整、不确定知识和数据的表达、学习、归纳的理论方法。1982 年， Pawlak 发表了经典论文 Rough Sets[21]，宣告了 RS 理论的诞生。1991 年，Pawlak 的第一本关于 RS 理论的专著“Rough Sets: Theoretical Aspects of Reasoning about Data”[22]和 1992 年，Slowinski R 主编的“Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory”[23]的出版，奠定了 RS 理论的基础，并推动了国际 RS 理论与应用的深入研究。

从 1992 年至今，每年召开以 RS 为主题的国际会议。为了促进 RS 理论在中国的发展，中国计算机协会人工智能与模式识别专业委员会于 2001 年 5 月在重庆邮电学院召开了第一届粗糙集与软计算(CRSSC)学术研讨会，以后每年召开一次 CRSSC 会议，共同理解和探讨粗糙集理论及其应用研究的新内容和新方法。目前 RS 理论已经成为人工智能领域中一个较新的学术热点，引起了越来越多的科研人员关注。

粗糙集理论已经广泛应用于医疗分析诊断、经济、金融、商业、环保、工程设计、信息科学、决策分析、社会、分子生物学和材料科学等诸多领域。

目前已经开发的基于粗糙集理论的知识发现系统有：挪威 Troll

Data Inc.在 4GL DBMS Paradox for Windows 下开发的一个基于 Rough 集理论的数据挖掘工具 Rough Enough；波兰 Poznan 工业大学计算科学研究所智能决策支持系统实验室开发的一个模块化软件系统 ROSE(Rough set data explorer)，它实现了 Rough 集理论的基本理论和规则获取技术；挪威科技大学数学研究所合作开发的一个基于 Rough 集理论框架的表格逻辑数据分析工具包 Rosetta；加拿大 Regina 大学研制的 KDD-R 系统是基于可变精度粗糙集模型，采用知识发现的决策矩阵方法；美国 Kansas 大学开发的基于粗糙集的实例学习系统 LERS(Learning from examples based on RS)。此外，国内还有中国科学院计算技术研究所开发的 KDT 和南京大学研制的 Knight，重庆邮电学院研发的 RIDAS 等。

3.2 粗糙集基本概念和理论

3.2.1 基本概念

粗糙集(RS)理论是一种刻化不完整和不确定的数学工具，其主要思想就是在保持分类能力不变的前提下，通过知识约简，有效地分析和处理不精确，不一致，不完整等各种不完备信息，从而导出问题的决策或分类规则，并从中发现隐含的知识，揭示潜在的规律。

粗糙集把客观世界或对象抽象为一个信息系统(Information System)，也称属性一值系统，信息系统的数据以关系表的形式表示，关系表的行对应要研究的对象，列对应对象的属性，对象的信息是通过指定对象的各属性值来表达。

令决策系统为 $S=(U,R,V,f)$，$R=C\cup D$，其中 $C=\{a_i \mid i=1,2,...,m\}$ 是条件属性集合，$D=\{d\}$ 是决策属性。$U=\{x_1,x_2,...,x_n\}$ 为论域，V 为属性的值域，$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值，如果 $D=\varnothing$ 则 S 为一个信息系统。

概念 3.1 知识。给定一对象的论域 U，对于任何子集 $X\subseteq U$ 可称为 U 中的一个概念或范畴，并且 U 中的任何概念族称为关于 U 的抽象知识，简称为知识(Knowledge)。

概念 3.2 近似空间。一个近似空间(Approximate Space)(或知识库)定义为一个关系系统(或二元组) $K=(U,R)$，其中 U 非空且为一个被称为全域或论域的所有要讨论的个体的集合，R 是 U 上等价关系的一个族集。

概念 3.3 不可分辨关系。设任一属性集合$B \subset R$，且B不空，B中所有等价关系的交集称为B上的一种不可分辨关系(Indiscernibility relation)(或不可区分关系、不分明关系)，记作$IND(B)$，即$IND(B)=\{(x,y)\in U\times U \mid \forall b\in B, f(x,b)=f(y,b)\}$。

定义 3.1 给定知识表达系统$S=(U,R,V,f)$，对于每个子集(对象)$X\in U$和不分明关系B(即$R=C\cup D$)，X的上近似集和下近似集分别可以由B的基本集定义如下(其中$IND(B)$是属性B的一个等价关系)[23]。

下近似集(lower approximation)：

$$B_{-}(X)=\cup\{Y_i \mid Y_i\in(U \mid IND(B)) \wedge Y_i\in X\}$$

上近似集(upper approximation)：

$$B^{-}(X)=\cup\{Y_i \mid (Y_i\in(U \mid IND(B)) \wedge Y_i\cap X\neq\varnothing)\}$$

$(B_{-}(X),B^{-}(X))$称为粗糙集。

定义 3.2 集合$pos_B(X)=B_{-}(X)$称为X的B正域，是U中所有根据分类$IND(B)$的信息可以准确划分到集合X的元素构成的集合，集合$BN_B(X)=B^{-}(X)-B_{-}(X)$称为$X$的$B$边界，是指那些根据$B$既不能判断肯定属于$X$又不能判断肯定属于$(U-X)$的$U$中元素组成的集合；$NEG_B(X)=U-B_{-}(X)$称为$X$的$B$负域，是指那些根据$B$判断肯定不属于$X$的$U$中元素组成的集合。如果$BN_B(X)=\varnothing$，则$X$是$B$可定义集，否则$X$是$B$粗糙集。

当且仅当$BN_B(X)=\varnothing$即$B^{-}(X)=B_{-}(X)$时，X为可定义集。当且仅当$BN_B(X)\neq\varnothing$即$B^{-}(X)\neq B_{-}(X)$时，X为粗糙集。

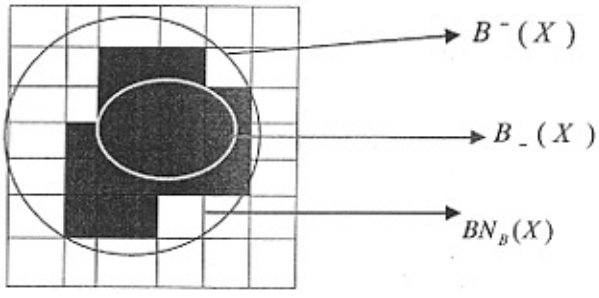

图 3-1

3.2.2 基本理论[24,25]

知识约简是粗糙集理论的核心内容之一。我们都知道知识库中的知识（信息决策表里面对象所具有的属性）并不是同等重要的，甚至其中某些知识是冗余的。知识约简就是在保持分类能力不变的条件下，删除其中不相关或不重要的知识。

定义 3.3 设U是一个论域，P是定义在U上的一个等价关系簇，$R\in P$。如果$IND(P-\{R\})=IND(P)$，则称关系R在P中是绝对不必要的(冗余的)；否则，称R在P中是绝对必要的。

绝对不必要的知识在知识库中是冗余的，除去这些知识不会改变该知识库的分类能力。相反，若知识库中去掉一个绝对必要的关系，则一定会改变知识库的分类能力。

定义 3.4 设U是一个论域，P是定义在U上的一个等价关系簇，P中所有绝对必要关系组成的集合Q（Q是独立的）称为关系簇P的绝对核，记作$CORE(P)$。

如果知识Q是知识P的绝对约简，则U中通过知识P可区分的对象，同样可以用知识Q来区分。

定理 3.5 $CORE(P)=\cap RED(P)$，其中$RED(P)$表示P的所有约简。

对于信息决策表而言，所有的条件属性形成条件属性集合C对论域U的划分U/C，同时，决策属性集$D=\{d\}$也对论域形成一个划分U/D。这两个划分形成了条件属性和决策属性集合在对论域样本分类上的知识。属性约简的目标就是要从条件属性集合中发现部分必要的条件属性，使得根据这部分条件属性形成的相对于决策属性的分类和所有条件属性所形成的相对于决策属性的分类一致，即和所有条件属性相对于决策属性D有相同的分类能力。

3.2.3 粗集约简算法

(1) 一般属性约简算法

令决策系统为$S=(U,R,V,f)$，$R=C\cup D$，其中$C=\{a_i\mid i=1,2,...,m\}$是条件属性集合，$D=\{d\}$是决策属性。算法描述如下：

```
BEGIN
FOR i=1 TO m
DO
{
    IF (C-{a_i}, POS_{(C-{a_i})}(Q)=POS_C(Q))
     THEN
          {C'=C-{a_i};
           删除a_i所在的列且合并U中重复的行；}
     ELSE C'=C
}
```

(2) 基于信息熵的约简算法[23]

U是一个论域，P和Q为U上的两个等价关系簇（属性集），可以认为U上任一等价关系簇是定义在U上的子集组成的σ代数上的一个随机变量，其概率分布可通过如下方法来确定。

定义 3.5 设P和Q在U上导出的划分分别为X和Y，

$X=\{X_1,X_2,\cdots,X_n\}$，$Y=\{Y_1,Y_2,\cdots,Y_m\}$，则P和Q的子集组成的σ代数上的概率分布为

$$(X:p)=\begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{bmatrix},$$

$$(Y:p)=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{bmatrix},$$

其中：$p(X_i)=\frac{|X_i|}{|U|}$，$i=1,2,\ldots,n;p(Y_j)=\frac{|Y_j|}{|U|}$，$j=1,2,\ldots,m$。

定义 3.6 知识（属性集合）P的熵$H(P)$定义为$H(P)=-\sum_{i=1}^{n}p(X_i)\log(p(X_i))$。

定义 3.7 知识（属性集合）$Q(U|IND(Q)=\{Y_1,Y_2,\cdots,Y_m\})$相对于知识（属性集合）$P(U|IND(P)=\{X_1,X_2,\cdots,X_n\})$的条件熵$H(Q|P)$定义为：

$H(Q|P)=\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))$，其中

$p(Y_j|X_i)=|Y_j\cap X_i|/|X_i|$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。

由上所述，基于信息熵的属性约简算法描述如下：

步骤 1 计算决策表T中决策属性D相对于条件属性C的条件熵；

步骤 2 计算决策属性相对于每个条件属性的条件熵$H(D|\{a_i\})(a_i\in C)$，将a_i按$H(D|\{a_i\})$的大小降序排列；

步骤 3 令$B=C$，按$H(D|\{a_i\})$递减的顺序对每个a_i重复下列运算：

（1）计算决策属性集相对于条件属性集B在去掉a_i后的条件熵$H(D|B-\{a_i\})$；

（2）判断条件熵的变化：如果$H(D|C)=H(D|B-\{a_i\})$，则属性a_i应约简掉，记$B=B-\{a_i\}$；如果$H(D|C)\neq H(D|B-\{a_i\})$，则属性$a_i$不被约简，$B$不变。

(3) Johnson’s 约简算法

定义 3.5[25] 设$S=(U,R,V,f)$为一决策表，称

$$M_d(i,j)=\begin{cases} a^*(x_i,x_j)=\{a_k\,|\,a_k\in A\wedge a_k(x_i)\neq a_k(x_j)\wedge w(x_i,x_j)\}, & \text{满足条件}T \\ 0, & \text{其他} \end{cases}$$

，其中条件 T 为：对于$x_i,x_j\in U$，$w(x_i,x_j)$满足：$x_i\in pos_B(D)$且$x_j\notin pos_B(D)$或者$x_i\notin pos_B(D)$且$x_j\in pos_B(D)$或者$x_i,x_j\in pos_B(D)$且$(x_i,x_j)\notin IND_B(D)$。$M_d(i,j)$为决策系统的可辨识属性矩阵，其中$i,j=1,\cdots,n$。显然，可辨识属性矩阵是一个对称矩阵，为了讨论方便，我们可以将其看为三角阵。

区分函数定义为：$DisF(x_i,x_j)=\prod_{(x_i,x_j)\in U\times U}\sum a^*(x_i,x_j)$。区分函数$DisF(x_i,x_j)$有如下性质：函数$DisF(x_i,x_j)$的极小析取范式中的所有合取式是$C$的所有$D$约简。

设B为约简结果集；将区分函数看成一系列集合的集合，则设S为整个区分函数集；s为区分函数内某元素，也为一集合；$w(s)$为S中s的权重，其中$w(s)$是根据决策表中数据属性而得到的。Johnsons 约简算法[26]如下：

BEGIN

Let $B=\varnothing$;

While $\{S\neq\varnothing\}$

{ { For i=1 to n　(n 为决策表中条件属性个数)

$W(a_i)=\sum w(s)$ where $a_i\in s\wedge s\subseteq S$;}

$S=S-\{s\}$;

$B'=\{a_i\}$ where $\max\{W(a_i)\}$;

$B=B+B'$;}

END

3.3 模糊聚类算法

3.3.1 聚类的基本概念

聚类是一种观察式学习，它不依赖预先定义的对象和带符号的训练实例，而是通过对象之间的相似性，将数据对象分为多个类或簇，同一个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。对象之间的相似度是根据描述对象的属性值来计算的。通常采用距离作为一种度量方式来描述对象间的相似度。

几种常用的距离函数[27]：

(1)Minkowsky 距离：$d(X,Y)=[\sum_{i=1}^{n}|x_i-y_i|^{\lambda}]^{\frac{1}{\lambda}}$

(2)Manhattan 距离：当 $\lambda=1$ 时，(1)为 Manhattan 距离：

$$d(X,Y)=\sum_{i=1}^{n}|x_i-y_i|$$

(3)Euclidean 距离：当$\lambda=2$时，(1)为 Euclidean 距离，又称欧式

距离：

$$d(X,Y)=[\sum_{i=1}^{n}|x_i-y_i|^2]^{\frac{1}{2}}$$

(4)Mahalanobis 距离：Mahalanobis 距离又称马氏距离：$D^2=(X-M)^tC^{-1}(X-M)$，C为样本的协方差。在假定各特征分量互相独立的情况下，协方差C为对角矩阵，在各分量方向上的密度分布均匀的情况下，C为单位矩阵，此时马氏距离实际上就是欧式距离。

(5)Hamming 距离：Hamming 距离最初用在编码中，两个码字的对应比特取值不同的比特数为这两个码字的海明距离。在粗集中使用 Hamming 距离，主要应用在粗集的定性属性上。其定义为：对于粗集中的定性属性a，如果这两个对象在属性a上完全相同，其 Hamming 距离为 0；否则为 1。对于决策表中的n个定性属性，两个对象x_i和x_j之间的 Hamming 距离公式为：$d(x_i,x_j)=\sum_{k=1}^{n}|a_k(x_i)-a_k(x_j)|$。

在数据挖掘中，聚类算法是一种比较常用的方法，现有的常用的聚类方法包括划分方法、层次方法、基于密度的方法和基于网格的方法。基于划分的方法包括常用的 K-means 算法和 k-medoids 算法；基于层次的方法包括 BIRCH 和 CURE 算法；基于密度的方法包括 DBSCAN 算法；基于网格的方法包括 STING 算法、CLIQUE 算法和 Wave cluster 算法。

聚类分析是按照一定的要求和规律将事物进行分类的一种数学方法。随着近年来数据挖掘技术的发展，聚类分析越来越多地用于大量的未知数据类别的分类。但是聚类是一个无监督的学习过程，分类是有监督的学习过程，两者的根本区别在于：分类时需要事先知道分类所依据的属性值，而聚类是要找到这个分类属性值。传统的聚类分析是一种硬划分，它把每个待辨识的对象严格地划分到某个类中，这种分类的界限是分明的。然而在实际的应用中，许多对象的类与类之间并无清晰的划分，边界具有模糊性，它们之间的关系是模糊关系。Zadeh[27]提出的模糊集理论为这种划分提供了有利的分析工具，用模糊数学的方法来进行聚类分析称为模糊聚类分析。模糊聚类分析是依据客观事务间的特征、亲疏程度和相似性，通过建立模糊相似关系对客观事务进行分类的方法。

3.3.2 模糊聚类方法

常用的模糊聚类方法有模糊 C 均值聚类法(Fuzzy C-means, FCM)、

基于模糊等价关系的传递闭包法、动态直接聚类法、基于模糊图论的最大树聚类法、基于摄动的模糊聚类 FCMBP 法和系统聚类法。下面介绍几个算法。

(1) 模糊 C 均值聚类法（FCM） [29]

模糊 C 均值聚类法，即模糊 ISODATA 法，是由 Bezdek J C 和 Dunn J C 提出的，用隶属度确定每个数据点属于某个聚类程度的一种聚类算法。算法如下：

把特征空间 $X=(x_1,x_2,\cdots,x_n)$ 划分为 c 个模糊组，可用模糊隶属矩阵 $U=[u_{ij}]\in R^{cn}$ 表示，U 中的元素 u_{ij} 表示第 $j(j=1,2,\cdots,n)$ 个数据点属于第 $i(i=1,2,\cdots,c)$ 类的隶属度，u_{ij} 应满足下列条件：

$$\sum_{i=1}^{c}u_{ij}=1,\forall j=1,2,\cdots,n;\quad u_{ij}\in[0,1],\forall i,j \tag{1}$$

$$\sum_{i=1}^{c}u_{ij}>0,\forall j=1,2,\cdots,c \tag{2}$$

Bezdek 将 Dunn 定义的目标函数（或价值函数）$J(U,C_1,\cdots,C_c)$ 算法推广到更一般的情况：

$$J_m(U,c_1,\cdots,c_i)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m d_{ij}^2 \tag{3}$$

式中 $C_i\in R^n$ 为模糊组 i 的聚类中心，$d_{ij}=|x_j-c_i|$ 为第 i 个聚类中心与第 j 个数据点的欧几里德距离，$d_{ij}^2(x_j,c_i)=(x_j-c_i)^T A(x_j-c_i)$，目标函数 $J_m(U,c_1,\cdots,c_i)$ 为每个数据点到相应的聚类中心的加权距离平均和；且 $m\in[1,\infty]$ 是一个模糊加权指数，用来控制隶属矩阵的模糊程度。目前 m 的选择大都来自实验或经验，一般取 $1.1\le m\le 5$。FCM 算法是一个使目标函数 $J_m(U,c_1,\cdots,c_i)$ 最小化的迭代收敛过程。

在 u_{ij} 满足约束条件下，用 Lagrange 乘法求解，使式（3）最小化的必要条件为：

$$c_i=\sum_{j=1}^{n}u_{ij}^m x_j/\sum_{j=1}^{n}u_{ij}^m,$$

$$u_{ij}=[\sum_{k=1}^{c}(\frac{d_{ij}}{d_{kj}})^{2/(m-1)}]^{-1} \tag{4}$$

由上述两个必要条件，模糊 C 均值聚类算法是一个简单的迭代过程。

(2) 基于模糊等价关系的传递闭包法

基于模糊相似关系的模糊聚类，首先要建立模糊相似矩阵，建立模糊相似矩阵的关键是标定相似系数。相似系数反映了样本之间相对于某些属性的相似程度。设 $O=\{x_1,x_2,\cdots,x_n\}$ 为被分类对象的全体，以 $\{x_{i1},x_{i2},\cdots,x_{im}\}$ 表示一个对象 x_i 的特征数据，可用数量积法、夹角余弦法和统计相关系数法等方法确定相似系数，建立模糊相似矩阵。

模糊相似矩阵 R 的传递闭包 R^* 是包含 R 的最小模糊等价矩阵，传递闭包法是根据 R 的传递闭包进行聚类的。

(3) 系统聚类法

系统聚类法先将 n 个样本各自看成一类，然后规定样本之间的距离和类与类之间的距离。开始时各个样本自成一类，类与类之间的距离和样本与样本之间的距离是相等的，选择距离最近的一对合并为一个新类。计算新类和其他类的距离，再将距离最近的两类合并，这样每次减少一类，直到所有的样品都归为一类为止。类与类之间的距离常用的有最小离差平方和法、最短距离、最长距离和中间距离等。虽然距离定义方式不同，但每种系统聚类法的并类原则和步骤是完全相同的。下面列举最小离差平方和法。

第 i 个类 C_i 样品间离差平方和为：

$$W_i=\sum_{l=1}^{n_i}(x_l-\bar{x}_l)'(x_l-\bar{x}_l) \tag{5}$$

k 个类的类内离差平方和为：

$$P_k=\sum_{i=1}^{k}W_i=\sum_{i=1}^{k}\sum_{l=1}^{n_i}(x_l-\bar{x}_l)'(x_l-\bar{x}_l) \tag{6}$$

n 个样品总的离差平方和为：

$$T^2=\sum_{l=1}^{n}(x_l-\bar{x})'(x_l-\bar{x}) \tag{7}$$

$\bar{x}$ 为所有样品的总重心：

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{x}_i \tag{8}$$

当 k 固定时，应选择使 P_k 达到最小的分类。两类合并后增加的离差平方和看成类间的平方距离，即：

$$D_{ij} = W_k - (W_i + W_j) \tag{9}$$

当C_i和C_j合并为新类C_k后，按离差平方和法计算类C_k和其他类C_l之间的距离的递推公式为：

$$D_k = \frac{N_i + N_l}{N_k + N_l} * D_{il} + \frac{N_j + N_l}{N_k + N_l} * D_{jl} - \frac{N_l}{N_k + N_l} * D_{ij} \tag{10}$$

4. 基于粗集理论和模糊聚类的 Web 使用挖掘

4.1 Web 使用挖掘

用 Web 使用挖掘技术能够从服务器、浏览器端的日志记录和用户的个人信息中自动发现隐藏在数据中的模式信息，了解系统的访问模式以及用户的行为模式，从而作出预测性分析。

Web 使用挖掘有几个主要阶段：数据预处理、模式发现、模式分析，在不同阶段的挖掘方法不同。Web 使用挖掘的框架模型如下图：

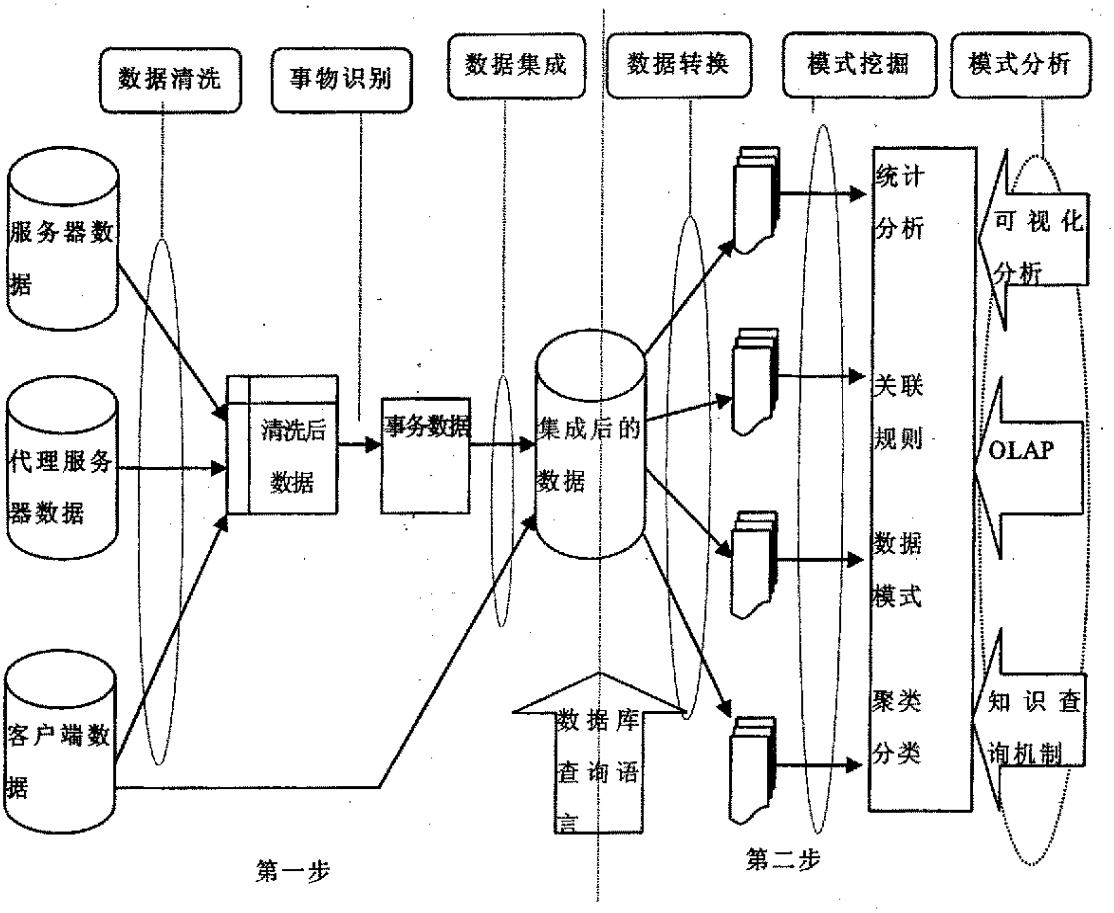

图 4-1 Web 使用框架模型

从上图可知，Web 使用挖掘分为两个部分。第一部分是把 Web 数据转换成适于挖掘的事务形式，根据不同的需求对日志数据进行预处理；第二部分包括通常的数据挖掘部分和具体的需求无关的一些挖掘算法。

Web 使用挖掘的第一部分在本文后面的实例中具体介绍。

4.1.1 模式挖掘

在 Web 使用挖掘研究中，模式挖掘是挖掘的算法实现部分。在模式挖掘中，常用统计法、机器学习和模式识别等方法。实现算法可以是：统计分析、关联规则、序列模式、分类、聚类等。

(1) 统计分析

统计方法是从 Web 中提取有用信息最常用的一种技术，通过对 Session 文件的分析，可以对感兴趣的信息进行统计，一般的包括各种统计数据，如最频繁访问的 N 个页面、每页平均浏览时间、网址路径平均访问长度等，也可能涉及一些关于限制的错误分析，如统计非法 IP、无效 URL 和未授权访问等。这些信息对于提高系统性能，加强网站安全起到辅助决策作用。

(2) 关联规则

关联规则是由 Agrawal[30]于 1993 年首先提出的。在 Web 使用挖掘中，关联规则主要用于发现用户之间、页面之间以及用户浏览页面和网上行为之间存在的潜在关系。基于用户访问页面的关联规则挖掘应用于页面推荐，是采用精确的访问模式匹配，推荐准确率高。

目前基于关联规则的最著名的挖掘方法是 Agrawal 提出的 Apriori 算法。现在已有一些改进算法，如无须产生候选挖掘频繁项目集的 FP 一树算法，提出闭合项目集概念的 A-Close 算法和 CLOSET 算法，采用频繁项目差集方式减少数据存储量的 C-HARM 算法等[31]。基于关联规则的推荐系统有：Lin 等提出的基于关联规则的推荐系统[32]，Mobasher 等提出基于关联规则[33]的推荐算法，Fu 等提出应用关联规则挖掘导航历史进行推荐的方法[34]。

无论哪种关联规则算法，都遵循两个步骤：第一步是迭代识别所有的频繁项目集，要求频繁项目集的支持率不低于用户设定的最小支持度；第二步是从频繁项目集中构造可信度不低于用户设定的最小置信度。

(3) 序列模式

序列模式挖掘目的是发现含有时间戳的事务间的关联关系。在 Web 服务器事务日志中记录的是一段时间内的用户访问行为，那么在数据预处理阶段，每个事务都会附带一个时间片，称为时间戳。Web 使用数据的序列挖掘，可以帮助研究人员预测用户访问行为。

(4) 分类

在 Web 使用挖掘中，分类技术可以发现如下关系：“从教育网发出

的用户请求 50%会访问/companv/products/book.html”。分类算法首先建立一个模型，通过对训练数据的分析，给出预定数据类集或概念集的特征描述，然后抽取未知数据对象的自身特性，根据模型中的定义，将其划分到相对应的类别中。最为典型的分类方法是基于决策树的分类方法，如著名的 ID3，它采用自顶向下不回溯策略，能保证找到一个简单的树。算法 C4.5 和 C5.0 都是 ID3 的扩展，它们将分类领域从类别属性扩展到数值型属性。

(5) 聚类

聚类(clustering)是一个将数据集划分为若干组(class)或类(cluster)的过程，并使得同一个组内的数据对象具有较高的相似度；而不同组中的数据对象是不相似的。相似或不相似的描述是基于数据描述属性的取值来确定的。通常是利用(各对象间)距离来表示的。

在 Web 使用挖掘领域有两种聚类：使用聚类和页面聚类。使用聚类就是将那些经常访问相同页面的用户群划分出来，他们具有相同的使用习惯和网上行为，可以对他们开展特定的广告策略或是个性化定制。页面聚类则发现内容相关的页面组，为搜索引擎和 Web 服务商提供有用信息。在 Web 使用挖掘中，聚类算法将用户浏览页面的综合视为数据空间，构造一个稀疏图。首先，根据每个页面的内容相似性和路径互联性，将数据对象分割为若干个 K-最近邻居子图(簇)，图中的每个点都代表一个页，子图的密度作为边的权重被记录下来。如果发现两个子图间的互联性和相似性与子图内部页面的互联性和相似性高度相关的话，则将二者合并为一个簇。

4.1.2 模式分析

在 Web 使用挖掘中，模式分析是结果的可视化部分。

如果没有合适的工具和机制来辅助分析人员的理解，采用各种技术挖掘出来的模式，数目庞大，表达晦涩，不能得到很好的利用。模式分析技术包括：统计、图形可视化、可用性分析和智能查询等。

4.1.3 Web 使用挖掘的系统产品

从 1996 年以来已出现了许多针对不同目标的分析 Web 使用数据的研究项目和商业软件，下作一个简单的分类和比较。

虽然可用于对 Web 使用挖掘的项目、产品进行分类的维度(标准)有不少，这里只采用了其中主要的五种[30]：1)获取输入的数据来源(服务器/代理服务器/客户机)；2) 输入数据的类型结构/内容/使用数据/用

（户注册信息）；3）各数据集中包含的用户个数(单个/多个)；4）该项目/产品所针对的应用领域类型(通用型、个性化服务型、商用型、网站改进型、使用特性型和系统改善型)。

Web 使用挖掘项目/产品分类比较如下表：

表 4-1 数据挖掘软件产品表

项目、产品名称	应用领域类型	数据来源			数据类型				用户		网站	
		服务器	代理服务器	客户机	结构	内容	使用数据	用户信息	单	多	单	多
WebSIFT	通 用	Y			Y	Y	Y			Y	Y	
SpeedTracer	通用	Y					Y			Y	Y	
WUM	通用	Y			Y		Y			Y	Y	
Shahabi	通用			Y	Y		Y			Y	Y	
Site Helper	个性化	Y				Y	Y		Y		Y	
Letizia	个性化			Y		Y	Y		Y			Y
Web Watcher	个性化		Y			Y	Y	Y		Y		Y
Krishnapuram	个性化	Y					Y			Y	Y	
Analog	个性化	Y					Y			Y	Y	
Mobasher	个性化	Y			Y		Y			Y	Y	
Tuzhilin	商用	Y					Y			Y	Y	
SurfAid	商用	Y				Y	Y			Y	Y	
Buchner	商用	Y					Y	Y		Y	Y	
WebTrends,Hitlist	商用	Y					Y			Y	Y	
WebLogMiner	商用	Y					Y			Y	Y	
PagerGather, SCML	网站修改	Y			Y	Y	Y			Y	Y	
Manley	特性化	Y				Y	Y			Y		Y
Arlitt	特性化	Y				Y	Y			Y		Y

Pitkow	特性化	Y		Y		Y	Y			Y		Y
Almeida	特性化	Y					Y			Y		Y
Rexford	系统改进	Y	Y				Y			Y	Y	
Schechter	系统改进	Y					Y			Y	Y	
Aggarwal	系统改进		Y				Y			Y	Y	
Wusage	系统改进	Y			Y		Y			Y	Y	
FastStats	系统改进	Y			Y		Y			Y	Y	

4.2 Web 日志数据模型

4.2.1 Web 浏览行为的模型描述

Web 服务器日志包括访问(服务器)日志、引用日志和代理日志。对这些日志预处理之后，我们以 $L=<ip,uid,url,time>$ 的形式表示 Web 服务器日志。其中，*uid*、*ip*、*url*、*time* 分别表示 Web 用户 ID，用户 IP 地址、用户请求的 URL 和相应的浏览时间。然后，对其进一步处理，以反映用户在某一段时间内的浏览行为。

定义 4.2(用户浏览行为)用户的浏览行为是具有如下形式的 $2\times n$ 元组：$B=<ip_B,uid_B,\{(l_B.url,hits)\}^n>$。其中 $l_B\in L, l_B.ip=ip_B, l_B.uid=uid_B, n\geq 1$，*hits* 表示到目前为止用户 uid_B 浏览页面 $l_B.url$ 的频度。

一个 Web 站点的拓扑结构就是一幅有向图，而用户一段时间内的访问模式则为其子图。具有相似访问子图的用户显然为需求相似客户，此即用户群体聚类。

定义 4.3(Web 站点)一个 Web 站点就是一幅具有如下形式的有向图：$G=(N,Np,E,Ep)$。其中 N 为结点集；$Np=\{Node\in N,\{<UserId,hits>\}^n\},n\geq 1$，记录用户 *UserId* 及其访问结点 *Node* 的频度，为结点属性集；E 为有向边集；$Ep=\{(e\in E,\{Number\ of\ path\}^p)\},p\geq 1$记录有向边及该有向边所在路径的编号，为有向边属性集。

假设数据预处理之后，得到大小为 n 的页面浏览的集合 $NP=\{Np_1,Np_2,...,Np_n\}$ 和大小为 m 的用户事务集合 $T=\{t_1,t_2,...,t_m\}$，其中 $t_i\in T$ 是 NP 的一个子集。那么每个事务 t 可以表示为一个长度为 l 的序列：$t=<(Np_1^t,w(Np_1^t)),(Np_2^t,w(Np_2^t)),...,Np_l^t,w(Np_l^t))>$。这里

$Np_i^t = Np_j(j \in \{1,...,n\})$，$w(Np_i^t)$是$Np_i^t$的权值。在 KNN 聚类算法中，将每个用户事务表示为n维页面空间中的一个向量：$\vec{t} = < w_{Np_1}^t, w_{Np_2}^t, ..., w_{Np_n}^t >$。当$Np_j$出现在$t$中时，$w_{Np_i}^t = w(Np_i^t)(i \in \{1,...,n\})$否则$w_{Np_i}^t = 0$。这样，所有的事务就组成了一个$m \times n$的矩阵$TP$。同时为了使内容信息能指导挖掘过程，提取每个浏览页面中的语义特征，表示为特征向量：$Np = \{< fw(Np, f_1), fw(Np, f_2), ..., fw(Np, f_k) >\}$，其中$Np \in NP$，$fw(Np, f_j)$,是第$j$个特征的权值（$j \in \{1,...,k\}$）。那么，浏览页一特征矩阵则可以表示为$PF = \{p_1, p_2, ..., p_n\}$。

将数据整理为上述矩阵后，在语义信息和用户信息结合时，将两个矩阵相乘：$TP \times PF = TF = \{t_1', t_2', ..., t_m'\}$。其中，$t_i'(i = \{1,...,n\})$是语义特征空间中的一个$k$维向量。这样，用户事务就用内容特征向量来表示。我们分别在事务矩阵TP和特征矩阵PF上做聚类，分别得到具有相似用户行为的事务集T和页面聚类P。

从有向图G的结点集N中可以得到该站点的所有URL，从相应的结点属性集Np中可以访问每一个结点$UserId$及相应的访问频度。据此可以建立如下所示的$UserId - URL$关联矩阵。

$$M_{mn} = \overbrace{\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1j} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2j} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{i1} & v_{i2} & \cdots & v_{ij} & \cdots & v_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mj} & \cdots & v_{mn} \end{bmatrix}}^{URL} \Big\} UserId \qquad (4.1)$$

其中v_{ij}表示第i个用户在一段时间内访问第j个页面(URL)的访问度，m为用户数，n为页面数，即 Web 事务数。每一个行向量$\alpha_i = (v_{i1}, v_{i2}, ..., v_{in})(i = 1..m)$表示用户对页面的访问情况，即相当于一个 Web 事务，它反应了用户类型，也勾勒出了用户的个性化访问子图。每一个列向量$\beta_j = (v_{1j}, v_{2j}, \cdots, v_{ij}, \cdots, v_{mj})(i \in \{1..m\}, j = 1..n)$即代表了站点的结构，又蕴含有用户共同的访问模式。

4.2.2 用户浏览时间的离散化表示方法

在聚类的相似性度量上，不仅要考虑在 Web 事务中对某页面的访问次数，而且要考虑在该页面上的浏览时间，故用离散化技术应用到用户浏览时间的表示上，本文将时间属性域划分为区间，用区间的标号来代替实际的时间值，可按照用户在网页上的浏览时间，将 Web 访问时间离散化如下：

表 4-2 访问时间离散化

离散化值 （根据数据值离散化）	访问情况(单位秒 s)
-3	访问该页面以后彻底离开
-2	访问页面后暂时离开
-1	访问一次离开
1	0s<=页面访问时间<20s
2	20s<=页面访问时间<40s
3	40s<=页面访问时间<60s
4	60s<=页面访问时间<80s
5	80s<=页面访问时间<100s
6	100s<=页面访问时间<200s
7	200s<=页面访问时间<300s
8	300s<=页面访问时间<400s
9	400s<=页面访问时间<500s
10	500s<=页面访问时间<600s
11	600s<=页面访问时间<700s
12	700s<=页面访问时间<800s
13	800s<=页面访问时间<900s
14	900s<=页面访问时间<1000s
15	1000s<=页面访问时间<1500s
16	1500s<=页面访问时间<2000s
17	2000s<=页面访问时间<2500s
18	2500s<=页面访问时间<3000s
19	3000s<=页面访问时间

采用这种浏览时间离散化的表示方法，用户只要访问了页面，即使时间再短也有离散化时间（离散化值为-1）；用户在页面上的浏览时间

即使很长，也有离散化时间（离散化值为 19）。这样就避免了在进行 Web 事务相似度量时，在采用连续时间情况下忽略用户浏览次数的情况，也避免了单纯地考虑访问次数而不考虑访问时间的问题。

在（4.1）中矩阵 M_{mn} 中的元素 v_{ij} 表示用户 i 访问页面 j 的访问度，数值上用该用户访问此页面的离散化时间刻画。

4.3 用户和页面聚类的方法-矩阵模糊聚类法[35,36]

4.3.1 原始数据标准化

要构造模糊相似矩阵，必须对数据进行标准化处理，使数据压缩到[0,1]闭区间内。设有 m 个对象 $o_1,o_2,\cdots,o_m$，每个对象具有 n 个对象指标 $y_1,y_2,\cdots,y_n$；x_{ij} 表示第 i 个对象的第 j 个指标。

m 个对象第 j 个指标的平均值和标准差分别为

$$\bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij},\quad s_j=\left[\frac{1}{m}\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2\right]^{1/2} \tag{4.2}$$

原始数据标准化为：$x'_{ij}=(x_{ij}-\bar{x}_j)/s_j$。运用极值标准化公式，将标准化数据压缩到[0,1]内，即 $x_{ij}=\dfrac{x'_{ij}-x'_{\min j}}{x'_{\max j}-x'_{\min j}}$，式中：$x'_{\min j}$ 和 $x'_{\max j}$ 分别是 $x'_{1j},x'_{2j},\cdots,x'_{mj}$ 中的最小值和最大值。

4.3.2 构造模糊相似矩阵

模糊相似矩阵：$R^F=\begin{bmatrix}r_{11}&r_{12}&\cdots&r_{1m}\\r_{21}&r_{22}&\cdots&r_{2m}\\\vdots&\vdots&\vdots&\vdots\\r_{m1}&r_{m2}&\cdots&r_{mm}\end{bmatrix}$，其中式中 r_{ij} 为两个对象 o_i 与 o_j 之间相似程度的变量，r_{ij} 越接近于 1，表明这两个对象越相似。r_{ij} 可以用距离法、夹角余弦法、相关系数法、主观评分法等来确定。常见的距离函数有 Hamming 函数，Minkowski 函数、Euclidean 函数和 Maximum 函数等。其中最常使用的是 Euclidean 的 $d_{ij}=\left[\sum_{k=1}^{n}|x_{ik}-x_{jk}|^2\right]^{1/2}$，$r_{ij}=1-pd_{ij}$，在这里 p 是一个适当选取的正数，使 r_{ij} 在[0,1]区间内。这里我们发现实质上 p 的选取应该是使得 $0\le r_{ij}\le 1$ 即 $0\le 1-pd_{ij}\le 1$。也就是 $0\le p\le\dfrac{1}{d_{ij}}$，即也可以表示为 $0\le p\le\dfrac{1}{\max\{d_{ij}\}}$。

由于对象之间是相似关系，即该关系只满足自反性、对称性，建立

的只是模糊相似矩阵，但是模糊聚类分析是利用模糊等价关系所对应的矩阵进行分类的，所以要对这个模糊相似矩阵求其传递闭包进一步得到模糊等价矩阵。本文运用图论的知识对最大树法的改进，进一步得到模糊等价矩阵。

4.3.3 求模糊相似矩阵的模糊等价矩阵

因模糊相似矩阵为一对称矩阵，将模糊相似矩阵 R 中的 r_{ij} 从大到小排列，并记录下相应的下标值 i,j，以用户对象为顶点，连接这些顶点为一个连通图，其中 i,j 分别对应第 i,j 个用户，若出现回路时，不连此线，直至全部样本对象连通为止，从而得到一棵最大树。在连接顶点的同时标注上对象之间的相似度。求等价矩阵的矩阵元素 r_{ij} 时，若 $i=j$：r_{ij} 为 1，否则取从 i 顶点到 j 的连通图中最小的权值[36,37]。基本算法如下：

模糊等价矩阵 R^*，其中 r_{ij}^*（$i,j=1,2,...,n$）是矩阵 R^* 的元素。

BEGIN

FOR i=1 to n

{IF（$i=j$）

THEN {$r_{ij}=1$}；

ELSE　//r_{ij} 为对象 u_i 与 u_j 之间的相似度

{

对模糊相似矩阵 R 中对称相似度的一半依次从大到小排列，并记下行数和列数；

生成只有结点没有边的无向图 G，每一个结点对应论域中的每一个个体对象；

WHILE(图 G 不连通)

{

IF（一条边加到图 G 后，图 G 出现回路）

THEN {丢弃此边}

ELSE

{按生成的相似序列依次在图 G 上连接相应的结点，并标上权重，即两个个体对象之间的相似度。}

}；}

FOR i=1 to n

{$r_{ii}^*=1$；

r_{ij}^{*}=min{树 G 上结点 i,j 通路上的最小权值}；

}

END

4.3.4 聚类

用上述图论法在求模糊相似矩阵的传递闭包后得到的模糊等价矩阵。最后得到的是一个连通图。取阈值 $\lambda\in(0,1)$，去掉线段上值小于 λ 的连线，剩下互相连接的样本对象则在水平 λ 下归为一类。λ 取值越大，分类精度越高，一个元素属于多个子类的可能性就越小，有可能是域内元素各成一类；反之 λ 取得越小，则分类越粗糙，一个元素同时属于多个子类的可能性就越大，有可能使域内全部元素都聚成一类。所以实际应用中，要根据聚类结果数目反馈动态的调节，选择适当的 λ。

5. 教务网日志数据实例分析

学校教务网（http://jiaowu.swjtu.edu.cn）的用户都是各年级学生，教务网服务器上每天都能收集到大量的用户访问记录。通过这些记录的处理，即 Web 使用挖掘，将教务网页面聚类、用户聚类，由此得到页面的内在关系和用户的访问模式，为教务网的网站设计提供理论依据，可以帮助教务处了解学生的网络利用情况，对于学生关心和感兴趣的问题有个清晰的把握。该方法还可以进一步应用于电子商务网站，以发现用户的访问模式，提高电子商务网站的客户访问量。

5.1 数据预处理

数据预处理是数据挖掘过程中关键的一步 ，因为现实世界的数据一般是脏的、不完整和不一致的。数据预处理可以改进数据的质量，从而提高其后续的数据挖掘过程的精度和性能。它是数据挖掘的重要步骤。

对于 Web 使用挖掘的研究也就是对 Web 日志数据的挖掘进一步发现有用的信息。在 Web 原始日志数据中，有很多的“垃圾”页面，如框架页面，也有很多“垃圾”记录，如记录中包含.jpg,.gif,.css 的记录。正因为 Web 日志的这些鲜明的特性，我们应该采用适当的方法将原始的“脏”数据组织成使用于挖掘的形式。

本文参考 Web 使用挖掘日志分析软件 WUMPREP 功能（可在 http://hypknowsys.sourceforge.net/wiki/Web-Log_Preparation_with_WUMprep中下载），使用 Java 编程算法对日志数据进行预处理，最后导入到 ACCESS 数据库中，利用 SQL 查询语句得出预处理以后的数据。

5.1.1 数据收集

原始日志数据为我校教务网 2004-9-12 16：00 到 2004-9-13 16：00 一天的日志数据，大小 392M，总计 2788549 条日志记录。日志格式在文 3.1.2.1 中已有描述，这里结合我校教务网日志记录格式描述如下

表 5-2 日志数据格式表

域	描述
date	日期
time	时间
s-ip	服务器 IP 地址
cs-method	客户端请求方法

cs-uri-stem	URI 资源
cs-uri-query	URI 查询
s-port	服务器端端口号
cs-username	客户端用户名
c-ip	客户端 IP
cs(User-Agent)	客户端代理
sc-status	协议状态
sc_substatus	协议子状态
sc_win32_status	Win32 状态

具体一条记录如：2004-09-12 16:00:00 202.115.66.198 GET /Default.asp - 80 - 218.194.14.53 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) 200 0 0，缺省值用“-”表示。

5.1.2 数据清洗

(1) 垃圾数据清理

现实世界中的数据一般是脏的、不完整的和不一致的。数据清洗过程试图填充空缺值，识别孤立点、消除噪声，并纠正数据中的不一致。

1）对于空缺值，通常有忽略元组、人工填写、使用全局量填充，使用平均值填充或使用最有可能的值填充。由于校园网日志数据中很少记录出现空缺值，所以对于空缺值，采用忽略记录的方法。

2）对于噪声数据，有分箱、聚类、计算机或人工检查以及回归等方法可以用来平滑数据，去掉噪声。在教务网日志中，我们利用计算机检查。一个用户请求的页面中包括几个框架、图片和脚本。由于服务器记录的是下传到客户端的一个个文件流，但是，用户请求的不是图片、框架和脚本，而是一个整体的页面。即在大多数情况下，只有日志中的 HTML 文件和用户会话有关。

因此，在噪声处理过程中，结合网站的拓扑结构，通过检查 URL 的后缀名利用 Java 编程(程序名为 Clean.jar)删除认为是不相关的文件，包括的后缀名有 gif,jpeg,jpg,cgi,css,js 和一些框架模板如：left.html,top.asp,.css 等。清洗数据以后利用 ImportToDb.jar 将数据导入 ACCESS 数据库中，如下所示部分(weblogs 数据表的转置部分视图)共有 738149 条记录：

表 5.3 日志数据部分视图表

date	04-9-12	04-9-12	04-9-13	04-9-13
time	16:00	23:32	00:03	11:53
s_ip	202.115.66.198	202.115.66.198	202.115.66.198	202.115.66.198
cs_method	GET	GET	POST	POST
cs_uri_stem	/Default.asp	/Course/MyCet.asp	/Course/CourseList.asp	/MyJiaowu/main.asp
cs-uri-query	----	----	----	----
s-port	80	80	80	80
cs-username	----	----	----	----
c-ip	218.194.14.53	222.18.43.151	202.115.66.198	202.115.66.198
cs(User-Agent)	Mozilla/4.0+...	Mozilla/4.0+...	Mozilla/4.0+...	Mozilla/4.0+...
sc-status	200	200	200	200
sc-substatus	0	0	0	0
sc-win32-status	0	0	0	0

(2) 冗余数据清洗

上述表中每一列表示一个用户的一条记录。如此记录共有 739644 条。此二维表可以看成是一个信息系统 $S=(U,R,V,f)$，其中 $R=C=\{a_1,a_2,\cdots,a_n\}$，$a_i(i=1,\cdots,n)$为条件属性，论域$U$可以看成是用户集合。我们初步认为用户以其 IP 标识。则此二维表中，论域U为用户集，属性为 $C=\{a_1,a_2,\cdots,a_{13}\}$=

{date, time, s_ip, cs_method, cs_uri_stem, cs_uri_query, s_port, cs_username, c_ip, cs(User-Agent), sc_status, sc_substatus, sc_win32_status)。

在 3.2.2 中介绍了对于信息决策表而言，所有的条件属性形成条件属性集合C对论域U的划分U/C，这样的划分形成了条件属性集合在对论域样本分类上的知识。从条件属性集合中发现部分必要的条件属性，使得根据这部分条件属性形成的分类和所有条件属性所形成的分类一致。

在上述 weblogs 整个表中，由于用户访问的是我校教务网，IP 地址为：218.194.14.53,，请求方法为 GET/POST，其中 Get 是用来从服务器上获得数据，而 Post 是用来向服务器上传递数据，HTTP 端口都为 80，服务器子状态为 0，所以上述日志数据中属性 $C_0=\{a_1,a_3,a_4,a_7,a_8,a_{12},a_{13}\}$ 为{04-9-12/04-9-13，218.194.14.53，GET/POST，80，---，0，0}，

论域U对属性集的划分$U/C=U/\{C-C_0\}$。所以在 weblogs 表中我们删除属性$\{a_1,a_3,a_4,a_7,a_8,a_{12},a_{13}\}$

=(date, s_ip, cs_method, s_port, cs_username, sc_substatus, sc_win32_status)保留下$C'=\{a_2,a_5,a_6,a_9,a_{10},a_{11}\}$

=(time, cs_uri_stem, cs_uri_query, c_ip, cs(User-Agent), sc_status)。

5.1.3 日志数据的进一步分析

1、用户的访问时间属性a_2。

如果用户$user_i$访问页面$page_j$时间为t_j，后续访问页面$page_{j+1}$的时间为t_{j+1}，则用户在$page_j$页面的停留时间为$staytime_{ij}=t_{j+1}-t_j$。在计算$staytime_{ij}$时，由本论文 2.5.2 中所述时间阈值设为 25.5 分钟，若$staytime_{ij}>25.5$分钟，则认为$page_{j+1}$是$user_i$访问该网站的下一个 session，此时认为$user_i$暂时离开了该网站，$staytime_{ij}$表示为-2。若$staytime_{ij}>25.5$分钟并且$user_i$在以后的访问日志中不再出现，即无后续$page_{j+1}$，则认为该用户彻底离开了该网站，$staytime_{ij}$表示为-3。若$user_i$在整个日志数据中只出现了一次，访问了一个网页后即离开，将$staytime_{ij}$表示为-1。由此，编写出程序 staytime.jar 计算出日志数据中用户访问页面的停留时间$staytime$。因此数据库中，我们删除时间属性a_2，添加一列$staytime$记录每个用户访问每个页面的停留时间。

2、用户访问网站的页面记录属性a_5。

由于该日志数据是我校教务网的日志数据。用户访问的网页页面都位于教务网主网站下的。所以此页面记录是以“/defaut.ap/…”形式，“/”表示教务网根下的即http://jiaowu.swjtu.edu.cn/或者可以看成是“http:// 218.194.14.53/”。其中，我们结合教务网拓扑结构及实际实践可知“/default.asp”和“/”相同。所以在日志数据中，凡是“/”的都改为统一的“/default.asp”形式。

3、用户的查询属性a_6。

查询条件是用户访问网页的一种操作，由于我们只关心用户访问的页面类型而不是用户的需求内容，所以删除这些查询信息列。即忽略属性a_6。

4、用户信息的属性a_9和a_{10}。

在 2.5.2 用户识别中已介绍了不同的 IP 地址代表着不同的用户；当 IP 地址相同时，默认不同的操作系统或浏览器代表不同的用户；在

IP 地址相同，用户使用的操作系统和浏览器也相同的情况下，则判断每个请求访问的页面与访问过的页面之间是否有链接，如果一个请求访问的页面与上一个已经访问过的所有的页面之间并没有直接的链接，则假设在访问 Web 站点的机器上同时存在多个用户。在此，我们只考虑第一、二个判断方法，首先以用户 IP 判断用户，再根据同一 IP 的若为不同操作系统则认为不同的用户，由此得到日志数据中的用户。

5、服务器状态信息的属性 a_{11}。

sc_status 表示的是用户请求后服务器的响应状态，通常有 5 个不同的值类型[36]。1××：信息响应类，表示接收到请求并且继续处理；2××：处理成功响应类，表示动作被成功接收、理解和接受；3××：重定向响应类，为了完成指定的动作，必须接受进一步处理；4××：客户端错误，客户请求包含语法错误或者是不能正确执行；5××：服务端错误，服务器不能正确执行一个正确的请求。上述日志数据表中，状态信息与其对应的记录数为

表 5.4

sc_status	Log_id 计数
200+206	613646+86=613732
301+302+304	3924+83679+605=88208
400+403+404	285+377+1295=1957
500+501	33818+434=34252

结合日志数据记录发现这些错误信息大部分是由于用户操作失误，或在请求服务器端数据时违反了数据库约定，或超时过期或服务器端（如 ASP_0147500_Server_Error）错误。又由于此日志数据用户的信息量很大，将这些错误信息所在的记录保存在一个 errortable 表里面，待后续处理，在分析原始记录时忽略这些错误信息。

5.2 数据归约

在清洗后的 738149 条日志记录中，同一 IP 的记录数从 1 到 15131，由于学校机房 IP 相同，可见记录次数很大的用户一定是机房用户。我们选择日志数据中 IP 记录数小于 100 的记录进行分析。首先将 IP 记录数小于 100 的记录放在一个数据表 alogs 中，一共有 29877 条记录。其

中第一列 log_id 是表示在原日志数据中的记录的 id。部分试图如下所示：

表 5.5

alogs

log_id	l_time	cs_uri_stem	c_ip	stay_time
2	2004-9-12 下午 04:00:00	/Default.asp	218.194.14.53	1
4	2004-9-12 下午 04:00:00	/Course/manual.asp	222.18.39.4	673
7	2004-9-12 下午 04:00:04	/news/html/pyfa/dq.rar	222.18.61.86	207
9	2004-9-12 下午 04:00:06	/Course/NextTermCourseList.asp	222.18.62.89	4
10	2004-9-12 下午 04:00:07	/Course/TeaClassList.asp	222.18.61.32	57
14	2004-9-12 下午 04:00:10	/Course/CourseList.asp	222.18.62.89	5
21	2004-9-12 下午 04:00:15	/Score/default.asp	222.18.62.89	2

将用户单独放在一个用户 cs_user 表，用户共 1567 个。页面单独放在页面 page 表中，page 页面共 211 个。部分视图如下：

表 5.6

表 5.7

cs_user

user_id	c_ip
1	10.11.17.240
2	152.226.7.203
3	159.226.142.62
4	192.168.50.64
5	194.138.202.2
6	202.108.207.181
7	202.108.249.184
8	202.108.249.185
9	202.108.249.186
10	202.108.249.187
11	202.108.249.188
12	202.108.249.189

page

page_id	page_name
1	/BeadRoll/class.asp
2	/BeadRoll/Default.asp
3	/BeadRoll/main.asp
4	/BeadRoll/print_student_list.asp
5	/BeadRoll/Query_Result_c.asp
6	/BeadRoll/Query_Result_s.asp
7	/BeadRoll/Query_Result_t.asp
8	/BeadRoll/student.asp
9	/BeadRoll/teacher.asp
10	/BookOnline/default.asp
11	/cet/default.asp
12	/Course/ChooseCourse.asp

将用户信息和页面信息都用表 5.6 和 5.7 表示，用户访问页面的停留时间 *staytime*，且 *staytime* 的离散化用 4.2.2 所述技术表示为 *visit*，放在一个 relation 表中。部分视图如下所示：

表 5.8

relation

relation_id	user_id	page_id	staytime	visit
1	824	47	-1	-1
2	1176	27	673	11
3	1517	138	207	7
4	1539	31	4	1
5	1497	41	57	3
6	1539	18	5	1
7	1539	172	2	1
8	472	6	29	2

relation

relation_id	user_id	page_id	staytime	visit
9	1539	180	0	1
10	1539	175	4	1
11	988	210	0	1
12	988	63	0	1

在 relation 表中，其中第一列是表中的序号，记录是以在原始 weblogs 表中的顺序排列的。表中一共有 29877 条记录。上述表相当于稀疏矩阵的压缩存储结构三元组表形式的用户信息表。

5.3 数据变换

由表 5.3-5.7 得出在日志数据中记录次数少于 100 的用户访问页面的信息表。在 ACCESS 数据表中，可以用 SQL 语句统计出每个用户访问页面的总的停留时间 *sumtime*，将 *sumtime* 用 4.2.2 的离散化技术离散化后得到用户访问页面的频度 *SVisit*。得到一个用户访问页面的决策表 S。其中 $S=(U,R,V,f)$，$R=C\cup D$，其中 $C=\{a_i \mid i=1,2,...,m\}$ 是条件属性集合，这里把用户访问的每个页面 $page_j(j=1,2,\cdots,211)$ 看成是条件属性。$D=\{d\}$ 是决策属性，这里把用户访问这些页面的总的频度 *SVisit* 看成是决策属性。$U=\{x_1,x_2,...,x_n\}$ 为论域即用户域 $user_i(i=1,2,\cdots,1567)$，$V$ 为属性的值域即用户访问页面的停留时间的离散化程度 *visit*，$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值即 $user_i$ 在 $page_j$ 上的取值 *visit*。

由上述描述得到一个 ACESS 中转换后的数据表 *user_page* 如下形式：

表 5.9

U	$page_1$	$page_2$	……	$page_{210}$	$page_{211}$	*SVisit*
$user_1$	0	0	……	0	0	-1
$user_2$	0	0	……	0	0	-1
……	……	……	……	……	……	……
$user_{1566}$	0	0	……	0	0	1
$user_{1567}$	0	0	……	0	0	-1

在此部分日志数据中，包含的页面共有 211 个，用户 1567 个，若用户未访问该页面则该用户对应的页面访问值为 0。

5.4 基于粗集的数据处理

在 ACESS 数据库中将 *user_page* 表导出为 *us_pg.txt*，*us_pg.xls* 形式。利用重庆邮电学院软件 RIDAS 处理原决策表。由于此决策表已经离散化且无缺省值，故在 *us_pg.txt* 中加入头信息：

Style:train

Stage:2

Condition attributes number:211

Records number:1567

导入到 RIDAS 中，分别用一般属性约简算法和基于信息熵的约简算法。此利用 RIDAS 时，page 计数是从 0 开始，而数据库中是从 1 开始计数，故在约简结果中，page(i+1)为实际保留后页面。

5.4.1 一般属性约简算法结果

Style :train

Stage: 3

Condition attributes number: 60

The Number of Condition attributes deleted: 151

The position of Condition attributes deleted: 0 2 3 4 8 10 12 13 14 19 20 21 22 23 25 29 30 31 32 34 35 36 37 39 41 42 43 44 47 49 51 53 54 56 57 58 61 64 65 67 68 69 70 71 72 73 74 75 76 77 78 79 81 82 83 84 85 87 88 89 90 91 92 93 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 132 133 136 137 138 140 141 144 145 146 147 149 150 151 154 155 156 157 162 163 165 166 167 168 170 172 173 174 175 177 180 181 182 183 184 186 187 188 189 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207

Records number: 1567

page1 page5 page6 page7 page9 page11 page15 page16 page17 page18 page24 page26 page27 page28 page33 page38 page40 page45 page46 page48 page50 page52 page55 page59 page60 page62 page63 page66 page80 page86 page94 page112 page128 page129 page130 page131 page134 page135 page139 page142 page143 page148 page152 page153 page158 page159 page160 page161 page164 page169 page171 page176 page178 page179 page185 page190 page191

page208 page209 page210

时间花费 3 分 56 秒。此结果说明，211 个页面中有 151 个页面可以删掉且不影响原决策表的知识分类能力。在数据库中找到这些删除的页面如下：

表 5.10

page_id	page_name
0	/BeadRoll/class.asp
2	/BeadRoll/main.asp
3	/BeadRoll/print_student_list.asp
...	.....
115	/News/html/20040406.htm
116	/News/html/20040526.htm
117	/News/html/20040527.htm
118	/News/html/20040913.htm
118	/News/Html/2004a/hycydb2004.htm

结合教务网站，这些删除的网页都是一些对某用户的特例网页。所以删除这些页面的确符合实际情况。

5.4.2 信息熵约简算法约简结果

Style: train

Stage: 3

Condition attributes number 60

The Number of Condition attributes deleted: 151

The position of Condition attributes deleted: 0 3 4 7 8 10 12 14 19 20 21 23 24 25 29 30 31 32 34 35 36 37 39 41 42 43 44 49 51 53 54 56 57 58 61 64 65 68 69 70 71 72 73 74 75 76 77 78 79 81 82 83 84 85 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 132 133 136 137 139 140 141 144 145 146 147 149 150 151 154 155 156 157 162 163 165 166 167 168 170 172 174 175 176 179 180 181 182 183 184 186 187 188 189 190 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207

208

Records number: 1567

page1 page2 page5 page6 page9 page11 page13 page15 page16 page17 page18 page22 page26 page27 page28 page33 page38 page40 page45 page46 page47 page48 page50 page52 page55 page59 page60 page62 page63 page66 page67 page80 page86 page112 page128 page129 page130 page131 page134 page135 page138 page142 page143 page148 page152 page153 page158 page159 page160 page161 page164 page169 page171 page173 page177 page178 page185 page191 page209 page210

用时 11 分 17 秒。分析同 5.4.2。

5.4.3 Johnson's Reduction 算法约简结果

将 *us_pg.xls* 导入 Rosetta 中,利用 Johnson's Reduction 算法,得到约简后的结果为:

Condition attributes number : 59

The Number of Condition attributes deleted: 152

The position of Condition attributes deleted: 1 4 5 8 9 11 13 15 20 21 22 24 25 26 30 31 32 33 35 36 37 38 40 42 43 44 45 50 52 54 55 57 58 59 62 65 68 69 70 71 72 73 74 75 76 77 78 79 80 82 83 84 85 86 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 132 133 134 137 138 140 141 142 145 146 147 148 150 151 152 155 156 157 158 164 165 166 167 168 169 170 171 173 175 177 178 180 181 182 183 184 185 187 188 189 190 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209

Records number: 1567

page2 page3 page6 page7 page10 page12 page14 page16 page17 page18 page19 page23 page27 page28 page29 page34 page39 page41 page46 page47 page48 page49 page51 page53 page56 page60 page61 page63 page64 page66 page67 page81 page87 page113 page129 page130 page131 page135 page136 page139 page143 page144 page149 page153 page154 page159 page160 page161 page162 page163 page172 page174 page176 page179 page186 page191 page192 page210 page211

用时 42 秒。分析同 5.4.2。

设 5.4.1 的约简结果为 R_1，5.4.2 的约简结果为 R_2，5.4.3 约简结果为 R_3。决策表的核 *Core* 包含 $R_1 \cap R_2 \cap R_3$。由上述 R_1, R_2 得到 $R_{12} = R_1 \cap R_2$ ＝{page1 page5 page6 page7 page9 page11 page15 page16 page17 page18 page26 page27 page28 page33 page38 page40 page45 page46 page48 page50 page52 page55 page59 page60 page62 page63 page66 page80 page86 page112 page128 page129 page130 page131 page134 page135 page142 page143 page148 page152 page153 page158 page159 page160 page161 page164 page169 page171 page178 page185 page191 page209 page210}，其中 page(i+1)才是对应数据库里面的实际页面。即 $R_{12} = R_1 \cap R_2$ 在数据库中实际对应页面为：R_{12}'＝{page2 page6 page7 page8 page10 page12 page16 page17 page18 page19 page27 page28 page29 page34 page39 page41 page46 page47 page49 page51 page53 page56 page60 page61 page63 page64 page67 page81 page87 page113 page129 page130 page131 page132 page135 page136 page143 page144 page149 page153 page154 page159 page160 page161 page162 page165 page170 page172 page179 page186 page192 page210 page211}。*Core* 也包含 $R_{12} \cap R_3$。在这里我们可以粗略将 *Core* 看成是 $Core = R_{12} \cap R_3$ ＝{page2 page6 page7 page10 page12 page16 page17 page18 page19 page27 page28 page29 page34 page39 page41 page46 page47 page49 page51 page53 page56 page60 page61 page63 page64 page67 page81 page87 page113 page129 page130 page131 page135 page136 page143 page144 page149 page153 page154 page159 page160 page161 page162 page172 page179 page186 page192 page210 page211}

记 $card(R)$ 为集合 R 中的基数。由上 R_1, R_2, R_3 得：$card(R_1) = 60$，$card(R_2) = 60$，$card(R_3) = 59$。$card(Core) = 49$。$Costtime(R_1)$ = 3 分 56 秒，$Costtime(R_2)$ = 11 分 17 秒，$Costtime(R_3)$ = 42 秒。由此一般属性约简算法、基于信息熵的属性约简算法和 Johnsons 算法在约简花费时间上看，Johnsons 约简算法明显优于前两个算法。

在第六章中，我们再对基于一般属性约简算法和 Johnsons 约简算法的约简结果进行进一步的聚类。

6. 教务网用户页面聚类

在这章中，我们结合 4.3 用户和页面聚类的方法-矩阵模糊聚类法及 5.4 基于粗集的数据处理结果，对教务网站的页面进行聚类分析。在此采用一般属性(General)约简算法和 Johnsons 算法约简所得处理结果作为聚类算法的数据源。

6.1 原始数据标准化

在构造模糊相似矩阵之前，对数据进行标准化处理，使数据压缩到[0,1]闭区间内。设有 m 个对象 $o_1,o_2,\cdots,o_m$，每个对象具有 n 个对象指标 $y_1,y_2,\cdots,y_n$；x_{ij} 表示第 i 个对象的第 j 个指标。在这里的分析中，对象为页面，对象指标为用户。一般属性约简结果中 $m=60$，$n=1567$；而在 Johnson 算法约简结果中，$m=59$，$n=1567$。

m 个对象第 j 个指标的平均值和标准差分别为

$$\bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij},\quad s_j=\left[\frac{1}{m}\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2\right]^{1/2}$$

原始数据标准化为：$x'_{ij}=(x_{ij}-\bar{x}_j)/s_j$。运用极值标准化公式，将标准化数据压缩到[0,1]内，即 $x_{ij}=\dfrac{x'_{ij}-x'_{\min j}}{x'_{\max j}-x'_{\min j}}$，式中：$x'_{\min j}$ 和 $x'_{\max j}$ 分别是 $x'_{1j},x'_{2j},\cdots,x'_{mj}$ 中的最小值和最大值。

因我们前面的数据存储形式都为行表示用户，列表示页面。逻辑表示上可以记为矩阵

$$M_{m\times n}=\overbrace{\begin{bmatrix}v_{11} & v_{12} & \cdots & v_{1j} & \cdots & v_{1n}\\ v_{21} & v_{22} & \cdots & v_{2j} & \cdots & v_{2n}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ v_{i1} & v_{i2} & \cdots & v_{ij} & \cdots & v_{in}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ v_{m1} & v_{m2} & \cdots & v_{mj} & \cdots & v_{mn}\end{bmatrix}}^{URL}\Big\}UserId$$

为了在程序中计算的方便，首先将矩阵 $M_{n\times m}$ 转置为 $M'_{m\times n}$，记为 $V_{m\times n}=M'_{m\times n}$。在上述计算平均值和方差的过程中，若 m 个对象的第 n 个对象

指标的方差为 0，即 m 个页面的第 n 个用户的访问频度的方差为 0，也就是说，第 n 个用户访问每个页面的频度相同。我们分析数据源可得这种情况只有在用户访问页面的频度都为 0 的时候发生。所以，若 m 个页面的第 n 个用户的访问频度的方差为 0，则在数据标准化中，直接用 0，不必纳入计算。

保存的原始数据标准化矩阵在文件 StdMatrix.txt 中。其中，$record_i$ 为第 i 条记录，即为第 i 个页面的 n 个用户的访问频度记录。由于一条记录有 1567 项，这里省略不列出。

6.2 构造模糊相似矩阵

在 4.3.2 中我们提及到了，模糊相似矩阵：$R^F=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix}$，其中式中 r_{ij} 为两个对象 o_i 与 o_j 之间相似程度的变量，r_{ij} 越接近于 1，表明这两个对象越相似。r_{ij} 可以用距离法、夹角余弦法、相关系数法、主管评分法等来确定。在这里我们使用 Euclidean 函数的 $d_{ij}=\left[\sum_{k=1}^{n}|x_{ik}-x_{jk}|^2\right]^{1/2}$，$r_{ij}=1-pd_{ij}$。在 4.3.2 中我们得到 $0\le p\le\frac{1}{d_{ij}}$ 实际可以表示为 $0\le p\le\frac{1}{\max\{d_{ij}\}}(i=1,\cdots m;j=1,\cdots,m)$。

在上述的距离 d_{ij} 中 $i=1,\cdots m;j=1,\cdots,m$，实际上也可以看成是一个距离矩阵 $DstMatrix_{m\times m}$。d_{ij} 表示 i 对象和 j 对象的距离，因此 $d_{ij}=d_{ji}$，距离矩阵 $DstMatrix_{m\times m}$ 是一个对称矩阵。因对象与自己的距离为 0。故我们只考虑此距离矩阵除去对角线元素以后的 $\frac{n\times(n-1)}{2}$ 个三角元素。并将其保存在 DstMatrix.txt 文件中。

DstMatrix.txt 文件部分显示如下。其中 $[i,j]$ 表示第 i 个对象和第 j 个对象的距离。

[1, 0]: 2.453
[2, 0]: 4.056
[3, 0]: 8.438

[4, 0]: 5.297

[5, 0]: 6.812

[6, 0]: 8.107

……

[56, 55]: 1.436

[57, 55]: 8.625

[58, 55]: 2.746

[57, 56]: 8.621

[58, 56]: 2.758

[58, 57]: 8.904

在 General 约简结果中，我们用上述方法得到其距离矩阵的最大值为 $\max\{d_{ij}\}=d_{17\,18}=24.007$，最小值为 $\min\{d_{ij}\}=d_{56\,57}=1.373$。由此知 $0\le p\le\frac{1}{\max\{d_{ij}\}}$ 即为 $0\le p\le 0.0417$。

在 Johnsons 约简结果中，我们也用上述方法得到其距离矩阵的最大值为 $\max\{d_{ij}\}=d_{9\,19}=24.050$，最小值为 $\min\{d_{ij}\}=d_{20\,39}=1.118$。由此知 $0\le p\le\frac{1}{\max\{d_{ij}\}}$ 即为 $0\le p\le 0.0416$。

对于每个数据源，我们在处理中，都选则了以下几个 p 作为输入：0.015，0.025，0.035，0.040，0.041。每个 p 对应生成的模糊相似矩阵我们都作为一个独立文件 FzMatrix.txt 保存下来以便在后续模糊等价矩阵和模糊聚类中比较其对结果的影响，从而更好的选择输入参数。

文件部分显示如下：

[1, 0]: 0.899

[2, 0]: 0.834

[3, 0]: 0.654

[4, 0]: 0.783

[5, 0]: 0.721

[6, 0]: 0.668

……

[56, 55]: 0.941

[57, 55]: 0.646

[58, 55]: 0.887

[57, 56]: 0.647

[58, 56]: 0.887

[58, 57]: 0.635

6.3 模糊相似矩阵的模糊等价矩阵

6.3.1 最大生成树

在上述模糊相似矩阵 $R^F=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix}$ 中 $r_{ij}=r_{ji}$，R^F 为一对称矩阵，其中 r_{ii} 为对象自身的相似性度量 $r_{ii}=1$。因此在我们实际的计算中，我们只考虑该矩阵除去对角线元素以后的 $\frac{n\times(n-1)}{2}$ 个下（上）三角元素。将这些元素中的 r_{ij} 从大到小排列，并记录下相应的下标值 i,j。在逻辑上可以看成把用户对象为顶点，连接这些顶点为一个连通图，其中 i,j 分别对应第 i,j 个用户，i,j 连线上的权值用 r_{ij} 表示。若出现回路时，舍弃该边，直至全部样本对象连通为止，从而得到一棵最大生成树。

我们将最大生成树保存在 CreTree.txt 文件中，将 r_{ij} 从大到小排列，同时记录生成树的边及其权值。若选择此边作为最大生成树的边，在文件中用*标记。文件中部分显示如下：

1. [57, 56] *, the edge number in generating tree 1
2. [45, 38] *, the edge number in generating tree 2
3. [54, 38] *, the edge number in generating tree 3
4. [55, 38] *, the edge number in generating tree 4
5. [54, 45]
6. [55, 45]

……

1710. [44, 8]

1711. [43, 8]

1712. [58, 18] *, the edge number in generating tree 59

0th generating tree edge: 57 to 56, pow=0.952

1th generating tree edge: 45 to 38, pow=0.951

2th generating tree edge: 54 to 38, pow=0.951

3th generating tree edge: 55 to 38, pow=0.951
4th generating tree edge: 56 to 38, pow=0.950
5th generating tree edge: 38 to 30, pow=0.949
6th generating tree edge: 41 to 38, pow=0.949
……
55th generating tree edge: 30 to 13, pow=0.646
56th generating tree edge: 30 to 17, pow=0.644
57th generating tree edge: 11 to 8, pow=0.394
58th generating tree edge: 58 to 18, pow=0.280

6.3.2 模糊等价矩阵

在由最大生成树生成等价矩阵的时候，我们先介绍几个前提条件。

(1)若无向图G，设G有n个顶点，$n-1$条边，且G是连通的⇔无向图G的最大生成树⇔无向图G无环。

(2)对于(1)中的无向图G中，若两个顶点v_1和v_2之间没有路径，那么若找到一条路径，那么这条路径就是最短路径。

证明：若两个顶点v_1和v_2之间本没有路径，现找到了两条路径 path1 和 path2，使得v_1和v_2连通，且两条路径的权值$weight_1 < weight_2$，则 path1 与 path2 一定不同，那么v_1和v_2之间有两条不同路径，对于无向图G来说，G中就有环与前提条件矛盾，原命题得证。

由(1)(2)可得若v_1和v_2有边相连，边上的权值就是两个顶点的之间的路径。若没有边直接相连，由于是连通图，总能找到几条边使得v_1和v_2可达。

由 6.2 中模糊相似矩阵及其生成树以及上述前提条件，我们可得等价矩阵的矩阵元素的求解算法。

求等价矩阵的矩阵元素r_{ij}时，若$i=j$：r_{ij}为 1，否则取从i顶点到j的连通图中最小的权值。基本算法如下：

模糊等价矩阵R^*，其中r_{ij}^*（$i,j=1,2,...,n$）是矩阵R^*的元素。

```
BEGIN
FOR i=1 to n
{
   IF (i=j) THEN {r_ij=1};
   ELSE  //r_ij为对象u_i与u_j之间的相似度
```

```
{
对模糊相似矩阵R中对称相似度的一半依次从大到小排列，
并记下行数和列数；生成只有结点没有边的无向图G，每一
个结点对应论域中的每一个个体对象；
  WHILE(图G不连通)
  {
     IF（一条边加到图G后，图G出现回路）
        THEN {丢弃此边}
     ELSE
     {
        按生成的相似序列依次在图G上连接相应的结点，
        并标上权重，即两个个体对象之间的相似度。
     }
  }// while
} // else
}//FOR
FOR (i=1 to n) {
   r*ii =1;
   r*ij=min{树G上结点i,j通路上的最小权值}；
}
END
```

上述算法用 Java 语言编程实现。我们逻辑上的图在文件中用矩阵上下标及其权值表示。生成树文件保存在 CreTree.txt 中。部分如下所示

```
[1, 0]: 0.899    path: 1--0
[2, 0]: 2.652    path: 2--39--1--0
[3, 0]: 2.425    path: 3--39--1--0
[4, 0]: 2.586    path: 4--39--1--0
[5, 0]: 2.517    path: 5--39--1--0
[6, 0]: 2.458    path: 6--39--1--0
[7, 0]: 2.449    path: 7--39--1--0
[8, 0]: 2.676    path: 8--39--1--0
[9, 0]: 2.788    path: 9--11--39--1—0
```

……

[56, 55]: 1.907 path: 56--39--55

[57, 55]: 3.112 path: 57--12--11--39--55

[58, 55]: 1.847 path: 58--39--55

[57, 56]: 3.111 path: 57--12--11--39--56

[58, 56]: 1.846 path: 58--39--56

[58, 57]: 3.051 path: 58--39--11--12--57

6.4 模糊聚类

用上述图论法求模糊相似矩阵的传递闭包后的模糊等价矩阵的过程中，得到一个最小生成树即一个无环连通图记为G。

模糊等价矩阵放在文件 EqualMatrix.txt 中。由于模糊等价矩阵也是一个对称矩阵，所以我们仍考虑除去对角线元素以外的$\frac{n\times(n-1)}{2}$个下三角元素作为记录。部分记录如下所示：

[1, 0]: 0.899

[2, 0]: 2.652

[3, 0]: 2.425

[4, 0]: 2.586

[5, 0]: 2.517

[6, 0]: 2.458

……

[56, 55]: 1.907

[57, 55]: 3.112

[58, 55]: 1.847

[57, 56]: 3.111

[58, 56]: 1.846

[58, 57]: 3.051

在无环连通图G中，取阈值$\lambda\in(0,1)$，去掉元素值小于λ的元素并记录下相对应的i,j值。剩下元素记录下其i,j，这些i,j对应对象在水平λ下归为一类。

在 6.3 构造模糊相似矩阵的时候，我们选择了几个不同的p值作为比较。对于不同p值生成的模糊相似矩阵和最小生成树以及模糊等价矩

阵都记录在不同文件中。从这些不同的文件中分析最后的最小生成树的情况，对于 p 为 0.015，0.025，0.035，0.040，0.041 的记录文件中，我们可以看出 p 为 0.041 时，模糊相似矩阵中元素在[0,1]区间分布得最好，也就是说，对于 p 为 0.041 时，模糊相似矩阵均匀分布。所以我们选定 p 为 0.041 作为两个数据源的输入参数。选取 λ 为 0.760 作为模糊聚类的阈值进行聚类。

6.5 结果分析

1、对于一般属性约简算法处理结果的模糊聚类

去除模糊等价矩阵中元素值小于 0.760 的元素并记录下相应的 i,j，即相当于在无环连通图 G 中去除边的权值小于 0.760 的边，得到孤立点。最后得到的连通分支数即为聚类结果。

在一般属性约简的结果中，聚类过程中，去掉的边为：

39th generating tree edge: 28 to 9, pow=0.753
40th generating tree edge: 52 to 38, pow=0.751
41th generating tree edge: 38 to 5, pow=0.746
42th generating tree edge: 44 to 38, pow=0.730
43th generating tree edge: 40 to 39, pow=0.717
44th generating tree edge: 40 to 38, pow=0.717
45th generating tree edge: 14 to 11, pow=0.711
46th generating tree edge: 58 to 11, pow=0.692
47th generating tree edge: 38 to 12, pow=0.687
48th generating tree edge: 14 to 10, pow=0.681
49th generating tree edge: 38 to 6, pow=0.676
50th generating tree edge: 58 to 29, pow=0.664
51th generating tree edge: 57 to 2, pow=0.653
52th generating tree edge: 16 to 10, pow=0.638
53th generating tree edge: 57 to 43, pow=0.617
54th generating tree edge: 15 to 14, pow=0.597
55th generating tree edge: 30 to 13, pow=0.585
56th generating tree edge: 30 to 17, pow=0.583
57th generating tree edge: 11 to 8, pow=0.291
58th generating tree edge: 58 to 18, pow=0.156

由于一般属性约简的结果对应网页页面下标从 0 开始，数据库中实际从 1 开始，所以我们将其页面 page+1 即最后得到的 21 类为：

{3}, {6}, {7}, {9}, {10}, {12}, {13}, {14}, {15}, {16}, {17}, {18}, {19}, {30}, {40}, {41}, {44}, {45}, {53}, {59}, {1, 2, 4, 5, 8, 11, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 46, 47, 48, 49, 50, 51, 52, 54, 55, 56, 57, 58, 60}。

表 6.1 表内页面独立成一类的页面表

page_id	page_name	id	page
7	/BeadRoll/Query_Result_t.asp	3	page7
12	/Course/ChooseCourse.asp	6	page12
16	/Course/CourseAppModify.asp	7	page16
18	/Course/CourseList.asp	9	page18
19	/Course/default.asp	10	page19
27	/Course/manual.asp	12	page27
28	/Course/MyCourseApp.asp	13	page28
29	/Course/myCourseList.asp	14	page29
34	/Course/select.asp	15	page34
39	/Course/StudentHome.asp	16	page39
41	/Course/TeaClassList.asp	17	page41
46	/Course/weeksch.asp	18	page46
47	/Default.asp	19	page47
87	/MyJiaowu/main.asp	30	page87
143	/News/NewsHitSum.asp	40	page143
144	/news/newsshow.asp	41	page144
154	/others/freshman.asp	44	page154
159	/others/reexam/query.asp	45	page159
179	/Score/MyCet.asp	53	page179
210	/UserLogin.asp	59	page210

表 6.2 表内页面聚成一类的页面表

page_id	page_name	id	page
2	/BeadRoll/Default.asp	1	page2
6	/BeadRoll/Query_Result_s.asp	2	page6
8	/BeadRoll/student.asp	4	page8
10	/BookOnline/default.asp	5	page10
17	/Course/CourseInfo.asp	8	oage17
25	/Course/helpme.asp	11	page25
49	/DownLoad/Default.asp	20	page49
51	/Download/down.asp	21	page51
53	/Download/files/chaoxuefen.doc	22	page53
56	/Download/files/default.asp	23	page56
60	/Download/files/zhuanzhuanyc.doc	24	page60
61	/emzsb/default.asp	25	page61
63	/error.asp	26	page63
64	/Evaluate/default.asp	27	page64
67	/guestbook/default.asp	28	page67
81	/MyJiaowu/default.asp	29	page81
95	/MyJiaowu/MyNotes/default.asp	31	page95
113	/News/Default.asp	32	page113
129	/news/html/2004b/files/Default.asp	33	page129
130	/news/html/2004b/files/kb_jsll.doc	34	page130
131	/news/html/2004b/files/kb_xszc.doc	35	page131
132	/News/html/2004b/files/studentreg.doc	36	page132
135	/news/html/default.asp	37	page135
136	/News/html/erzy.doc	38	page136
140	/news/html/pyfa/jsj.rar	39	page140
149	/notice/duizhaobiao/a-bei.htm	42	page149

page_id	page_name	id	page
153	/others/contact.htm	43	page153
160	/others/roomquery.asp	46	pgae160
161	/others/spareroom.asp	47	page161
162	/others/timetable.htm	48	page162
165	/others/YjsBm/query.asp	49	page165
170	/robots.txt	50	page170
172	/Score/default.asp	51	page172
177	/Score/GetPrintScore.asp	52	page177
180	/Score/myNewScore.asp	54	page180
186	/teacher	55	page186
191	/teacher/Files/3040400/2	56	page191
192	/teacher/Files/3040400/2/2004913173952.doc	57	page192
209	/teacher/upload_sec.asp	58	page209
211	/wnl.htm	60	page211

2、对于 Johnsons 属性约简算法处理结果的模糊聚类分析过程同 1，去掉边如下：

37th generating tree edge: 31 to 10, pow=0.755
38th generating tree edge: 39 to 5, pow=0.755
39th generating tree edge: 45 to 39, pow=0.732
40th generating tree edge: 12 to 11, pow=0.730
41th generating tree edge: 51 to 39, pow=0.724
42th generating tree edge: 39 to 11, pow=0.722
43th generating tree edge: 41 to 39, pow=0.718
44th generating tree edge: 41 to 40, pow=0.718
45th generating tree edge: 57 to 12, pow=0.707
46th generating tree edge: 39 to 13, pow=0.698
47th generating tree edge: 39 to 6, pow=0.696
48th generating tree edge: 39 to 7, pow=0.686

49th generating tree edge: 57 to 32, pow=0.672
50th generating tree edge: 39 to 3, pow=0.662
51th generating tree edge: 39 to 17, pow=0.643
52th generating tree edge: 44 to 39, pow=0.622
53th generating tree edge: 16 to 15, pow=0.607
54th generating tree edge: 14 to 11, pow=0.604
55th generating tree edge: 39 to 18, pow=0.593
56th generating tree edge: 11 to 9, pow=0.303
57th generating tree edge: 57 to 19, pow=0.159

聚类结果如下 23 类：

{3}, {5}, {6}, {7}, {9}, {10}, {12}, {13}, {14}, {16}, {17}, {18}, {19}, {32}, {40}, {41}, {44}, {45}, {51}, {53}, {57}, {11, 15}, {1, 2, 4, 8, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 42, 43, 46, 47, 48, 49, 50, 52, 54, 55, 56, 58, 59}。

表 6.3 表内页面独立成一类的页面表

page_id	page_name	id	page
6	/BeadRoll/Query_Result_s.asp	3	page6
10	/BookOnline/default.asp	5	page10
12	/Course/ChooseCourse.asp	6	page12
14	/Course/CourseAppChange.asp	7	page14
17	/Course/CourseInfo.asp	9	page17
18	/Course/CourseList.asp	10	page18
19	/Course/default.asp	11	page19
23	/Course/help.asp	12	page23
27	/Course/manual.asp	13	page27
28	/Course/MyCourseApp.asp	14	page28
29	/Course/myCourseList.asp	15	page29
34	/Course/select.asp	16	page34
39	/Course/StudentHome.asp	17	page39
41	/Course/TeaClassList.asp	18	page41

page_id	page_name	id	page
46	/Course/weeksch.asp	19	page46
81	/MyJiaowu/default.asp	32	page81
139	/news/html/pyfa/index.htm	40	page139
143	/News/NewsHitSum.asp	41	page143
153	/others/contact.htm	44	page153
154	/others/freshman.asp	45	page154
172	/Score/default.asp	51	page172
176	/Score/GetMyPrintScore.asp	53	page176
192	/teacher/Files/3040400/2/2004913173952.doc	57	page192

表 6.4 表内页面聚成一类的页面表

page_id	page_name	id	page
2	/BeadRoll/Default.asp	1	page2
3	/BeadRoll/main.asp	2	page3
7	/BeadRoll/Query_Result_t.asp	4	page7
16	/Course/CourseAppModify.asp	8	page16
47	/Default.asp	20	page47
48	/Download/bysj/Default.asp	21	page48
49	/DownLoad/Default.asp	22	page49
51	/Download/down.asp	23	page51
53	/Download/files/chaoxuefen.doc	24	page53
56	/Download/files/default.asp	25	page56
60	/Download/files/zhuanzhuanye.doc	26	page60
61	/emzsb/default.asp	27	page61
63	/error.asp	28	page63
64	/Evaluate/default.asp	29	page64
66	/guestbook/adddata.asp	30	page66

page_id	page_name	id	page
67	/guestbook/default.asp	31	page67
87	/My.Jiaowu/main.asp	33	page87
113	/News/Default.asp	34	page113
129	/news/html/2004b/files/Default.asp	35	page129
130	/news/html/2004b/files/kb_js11.doc	36	page130
131	/news/html/2004b/files/kb_xszc.doc	37	page131
135	/news/html/default.asp	38	page135
136	/News/html/erzy.doc	39	page136
144	/news/newsshow.asp	42	page144
149	/notice/duizhaobiao/a-bei.htm	43	page149
159	/others/reexam/query.asp	46	page159
160	/others/roomquery.asp	47	page160
161	/others/spareroom.asp	48	page161
162	/others/timetable.htm	49	page162
163	/others/YjsBm/printcj.asp	50	page163
174	/Score/getMyAllScore.asp	52	page174
179	/Score/MyCet.asp	54	page179
186	/teacher	55	page186
191	/teacher/Files/3040400/2	56	page191
210	/UserLogin.asp	58	page210
211	/wnl.htm	59	page211

从表 6.2 和 6.4 中我们可以看成，聚成一类的页面都是以学生选课，培养方案为中心。学生用户的兴趣所在为以选修课程，填写培养方案为中心进行页面浏览操作。两个表的页面聚类中心之所以不一样就是因为两个约简算法的差别。由最大生成树和表 6.2，6.4 得出的页面可以看出，用 Johnson 约简算法得到的类中心点为页面 /news/html/pyfa/jsj.rar，即培养方案。各页面之间的关联都是通过 /news/html/pyfa/jsj.rar 得到。用 General 约简算法得到的类中心点

为页面/News/html/erzy.doc，即第二专业辅修专业说明页面。结合我校实际情况，此日志数据的日期正好为我校新学期开学时间。对于教务网学生选课为其核心目标。因此，以 Johnson 约简算法得到的以培养方案为多个页面聚类中心更符合实际。在此过程中，教务网中网络承载量更多的是选课系统。因此可以针对此特定时期，加大选课系统的网络维护，更好的为学生服务。还可以结合网络拓扑图和所得结论，调整一下网络拓扑结构。

论文总结

论文主要工作

论文就基于粗糙集理论和聚类算法的 Web 使用挖掘研究做了以下几个工作：

(1) 介绍了数据挖掘，Web 数据挖掘的定义、过程、分类和使用的技术，分析了 Web 挖掘的难点，总结了一段时间以来国内外在 Web 使用挖掘中的研究现状。

(2) 介绍了粗糙集的产生和发展及其基本理论和其约简算法。

(3) 就聚类算法介绍了其中的模糊聚类算法及其分类算法。

(4) 在上面理论引入的基础上，研究了基于粗糙集的 Web 使用挖掘，主要是针对日志数据的粗糙集数据挖掘研究。针对日志数据建立模糊聚类算法的理论模型，并进行分析讨论。

(5) 结合我校教务网日志数据，在上面理论分析的基础上，利用粗糙集国内软件 Ridas 和国外软件 Rosetta 进行数据预处理，进一步用 Java 编程实现了模糊聚类算法得到教务网页面聚类。

论文的创新之处在于结合了粗糙集理论和模糊聚类算法对日志数据进行分析处理。其中，预处理部分的时间离散化方法，模糊聚类算法中对象之间的相似度求解中参数 p 的求得，模糊等价矩阵及图的模糊聚类算法的具体化实现等都是本文在基于以前方法的改进之处。论文还可以利用此方法类似的对用户聚类。

进一步研究的问题

本文对粗糙集约简算法进行了讨论和分析，并对其数据集教务网日志数据进行了验证。由于对教务网拓扑结构并不是很清楚，所以约简时有些无效页面并没有约简掉。比如页面框架/BeadRoll/main.asp 仍在约简结果中。进一步工作可以在粗糙集的约简算法上进行改进并编程实现。

由于教务网用户都是在校学生，从结果中看出，聚类的效果并不明显，可以进一步研究改进模糊聚类算法对于不同类型数据集的聚类效果。

致 谢

论文的顺利完成，首先要感谢我的导师尹治本教授。感谢尹老师在我三年的硕士研究生涯中的指导、关心和帮助。尹老师不仅给予我研究方向的指导，还指点了我有关研究的重要思想方法。导师科学严谨的态度、深厚扎实的理论知识、忘我敬业的工作精神、脚踏实地的工作作风、平易近人的待人处事方式，这一切都深深的影响了我，成为我毕生学习的榜样。我所取得的每一点进步，都与导师的鼓励和指导密不可分。

感谢粗糙集研究小组的蒋朝哲博士和涂瑞同学，感谢他们在我研究生生活中给予的大量帮助和关心！感谢他们带给我丰富的思想和无限的乐趣！感谢我的朋友杨美成、温海峰，感谢他们在我论文数据处理部分给予的无私帮助！感谢 01402 实验室所有的同学和曾经帮助过我的所有朋友，感谢他们给予我的支持和帮助。感谢教务处老师给我提供了教务网日志数据，使得我的理论研究有了实际数据的验证。

同时对百忙中抽出宝贵时间对本文进行评阅的专家和学者表示诚挚的谢意！

感谢我的父母，一直以来你们都在默默的关心和支持着我，给予我精神上无限的力量！

参 考 文 献

[1][SCDT2000] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. Web usage mining: Discovery and Applications of Usage Pattern from Web Data ACM SIGKDD, VOL.1, Issue 2.2002

[2] Krishna Bharat, Tomonari Kamba, Michael Albers. Personalized, Interactive news on the Web. Multimedia Systems. 1998-6

[3] 姚远。两种网站个性化方案及其实现[J]，计算机应用研究，2001.12:64-66

[4]Eui-Hong Han, Daniel Boley, Maria, Robert Gross, Kyle Hasings, George Karypis, Vipin Kumar, Bamshad Mobasher, Jerome Moore, "WebACE:Aweb Agent for Document Categorization and Exploration", proceedings of the second international conference onAutonomous Agent, 1999

[5]T.-P. Hong, K.-Y. Lin, S.-L. Wang. Mining linguistic browsing patterns in the world wide web. Soft Computing 6(2002)329-336

[6]Oyanagi, S, Kubota, K, Nakase, A, Application of Matrix Clustering to Web Log Analysis and Access Prediction, WEBKDD-2001

[7]R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules, Proc. 20th VLDB Conf. 1999. PP:487-499

[8]Pang-Ning Tan, Vipin Kumar. Mining Indirect Associations in Web Data. Lecture Notes In Computer Science; Vol.2356:145-166

[9] Joshua Zhexue Huang, Michael K. Ng, Wai-Ki Ching, Joe Ng, David Wai-Lok Cheung. A Cube Model for Web Access Sessions and Cluster Analysis. Lecture Notes In Computer Science. Vol. 2356:48-67

[10]Alexandros .Nanopoulos, Dimitris Katsaros, Yannis Manolopoulos, Effective Prediction of Web-user Accesses: A Data Mining Approach. In Proceeding of the Workshop WEBKDD http://citeseer.ist.psu.edu/nanopoulos01effective.html

[11][BM98]AlexG. Buchner, Maurice .Mulvenna. Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. ACM SIGKDD Record. ISSN

0163-5808, vol. 27, No4, pp. 54-61, 1998

[12] 邓英，李明。用户访问模式挖掘中的数据预处理问题的研究[J]，计算机工程与应用，2002.01：188-190

[13] 夏火松。数据仓库与数据挖掘技术[M]，北京：科学出版社，2004

[14] 朱杨勇，数据挖掘软件分析 .PPT, http://www.dmgroup.org.cn/pptdown050322/pptdownload1.htm

[15]Oren Etzioni. The World Wide Web: quagmire or gold mine. Communication of the ACM, 1996, 39(11)

[16] 朱丽红，赵燕平。Web 挖掘研究综述[J]，情报杂志 2004.7:2-5

[17] Buchner AG, Mulvenna MD. Discovering Internet Marketing Intelligence Throgh Online Ananlytical Web Usage Mining. ACM SIGMOD Record. 1998(4)

[18] 童恒庆，梅清。Web 日志挖掘数据预处理研究[J]，MODEN COMPUTER 2004.3：6-10

[19]Catledge L, Pitkow J. Characterizing browsing behaviors on the World Wide Web. Computer Networks and ISDN Systems. 1995, 27(6):1065-1073

[20]Chen M S, Park J S Yu P S. Data mining for path traversal patterns in a web environment. In:Proc of the 16th Int' l Conf on Distributed Computing System. Hong Kong. 1996:385-392

[21]Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences 1982, 11: 341-356

[22]Pawlak Z, Rough Sets: Theoretical Aspects of Reasoning about Data [M], Dordrecht: Kluwer Acasemic Publisher, 1991

[23]Slowinski R, Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory [M], Dordrecht: Kluwer Academic Publisher, 1992

[24] 王国胤 编著，。Rough 集理论与知识获取[M]，西安：西安交通大学，2001。

[25] 张文修，吴伟志，梁吉业，李德玉 编著。粗糙集理论与方法[M]，北京：科学出版社，2001。

[26]Aleksander, Rosetta Technical Reference Manual.

[27] 孙惠琴，熊璋。基于粗集的模糊聚类方法和结果评估[J]，复旦学报(自然科学版)。2004.10 Vol 43(5):143-146

[28]Zadeh L A. Fuzzy set[J]. Information and Control, 1965, 8:338-353

[29] 梁伍七，江克勤。数据挖掘中的模糊聚类分析及其应用[J]，安庆师范学院学报（自然科学版)。2004.5 Vol 10 (2)：150-152

[30]Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Exploration, 2000, 1(2):12-23

[31] 陈健,印鉴。Web 使用挖掘技术研究综述[J]，计算机工程。2005.5. Vol.31(9):4-6

[32]Lin W, Alvarez S A, Ruiz C. Efficient adaptive support association rule mining for recommender systems[J]. Data Mining and Knowledge Discovery, 2002, 6(1):83-105

[33]Mobasher B, Dai H, Luo T, et al. Effective personalization based on association rule discovery from web usage data[A]. Mobasher B ed. 3rd Int Workshop on Web Information and Data Management (WIDM 2001)[C]. New York: ACM Press, 2001: 9-15

[34]Fu X, Budzik J, Hammond K. Mining navigation history for recommendation [A]. Riecken D ed. Proc Int Conf Intelligent User Interfaces[C]. New York: ACM Press, 2000: 106-112

[35] 邢东山，宋擒豹，沈钧毅。一种新的 Web 事务模糊聚类算法的研究[J]，西安交通大学学报。2002.8 Vol 36 (8)

[36] 宋擒豹，深钧毅。Web 日志的高效多能挖掘算法[J]，计算机研究与发展。2001.3 Vol 38 (3)：143-148

[37] 骆洪青，吴小俊，曹奇英。模糊聚类分析的一种新方法研究[J]，华东船舶工业学院学报。2000.6 Vol 14 (6)：24-27

[38]http://www.cnpaf.net/2004/8-2/194156.html

[39] 陈恩红，徐涌，王煦法。Web 使用挖掘：从 Web 数据中发现用户使用模式[J]，计算机科学，Vol.28(5):85-88, 2001

[40] 行小帅,焦李成。数据挖掘的聚类方法[J],电路与系统学报。2003.2 Vol 8 (No 1)：120-128

[41] 郭伟刚。电子商务网站用户访问模式挖掘中的预处理技术[J]，计算机应用，2005.3 Vol25(3):691-694

[42] 史忠值。知识发现[M]，北京：清华大学出版社，2002.1

[43] Robert Grossman, Supporting the Data Mining Process with Next Generation Data Mining Systems, http://www.lac.uic.edu/grossman/paper/esj-98.htm.1998
Robert Grossman, The Terabyte Challenge Discoverying Information in Distributed and Massive Data. KDD'01

[44] 章成志。数据挖掘研究现状及最新进展[J]，南京工业职业技术学院学报，2003.6 Vol3(2):1-5

[45]Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. Proceedings of the ACM SIGMOD conference on management of data, 1993.207-216

[46]Aggarwal C, Yu P. Data mining techniques for personalization[J]. IEEE Data Engineering Bulletin, 2000, 23(1):4-9

[47]R Cooley, J Srivastava. Web Mining: Interesting patterns disvoery from web data. Department of Computer Science, University of Minnesota. Tech Rep: TR99-023, 1999

[48] 邓松林。基于粗糙集的 Web 用户模式挖掘研究，重庆大学 2003 硕士学位论文

[49] 田海山。基于 Web 日志的数据挖掘研究。河北工业大学 2003 硕士学位论文

[50] 郝先臣，刘小玲，周建华，赵海。模糊聚类挖掘方法在电子商务中的应用[J]，东北大学学报（自然科学版），2001.8. Vol.22，No.4:389-392

[51]Pawan Lingras, Rough set clustering for Web mining, Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on Volume 2, 12-17 May 2002 Page(s):1039 - 1044

[52]Lingras, P.; Rough set clustering for Web mining. Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference onVolume 2, 12-17 May 2002 Page(s):1039

- 1044
[52] 刘靖，陈福生。结合粗糙集和模糊聚类方法的属性约简算法[J]，计算机应用软件 2004.11. Vol.21, No.11:72-74
[53] 何清。模糊聚类分析理论与应用研究进展[J]。模糊系统与数学，1998，12（2）：89-94
[54] 梁伍七，江克勤。数据挖掘中的模糊聚类分析及其应用[J]。安庆师范学院学报（自然科学版），2004.3 Vol.10, No.2:65-67
[55] 刘景芬。聚类分析在股票研究中的应用[J]。天津纺织工学院学报，1995.Vol 14.No 4:82-86

攻读学位期间发表的论文和参与项目

发表论文

1、高晓琴，蒋朝哲，涂瑞。Web 使用挖掘研究 微计算机信息 2006.11；

2、蒋朝哲，高晓琴，涂瑞。基于 DNA 计算机理的粗糙集约简算法构想 计算机科学 2005.8.A

3、蒋朝哲，胡培，高晓琴。多属性决策问题方案筛选的粗糙集方法 2005 交通运输工程博士论坛·天津

4、蒋朝哲，关秦川，涂瑞，高晓琴。基于粗糙集理论的项目评价方法研究 管理工程学报 2005 增刊

5、高晓琴，蒋朝哲，涂瑞。基于粗糙集的决策排序法 （已投稿）

参与项目

西南交通大学博士创新基金项目[20005]