# Research on Data Mining of Archives Based on Knowledge Management

**Jieping Dong[1], Haitao Han[2]**

[1]*Department of Library Tianjin Polytechnic University, Tianjin, China*

[2] *Department of Archives. Tianjin Polytechnic University, Tianjin, China*

*Email: 397638175@qq.com, tjdagh@163.com*

**Abstract:** As the era of knowledge economy has arrived, to implement the knowledge management on digital archives is the requirement of this era, and is the necessity of the archives' own development as well. Data mining provided pre-preparation and technical support for the effective management of knowledge resources for the digital archives. In this paper, a brief introduction has been made on related theories of knowledge management and data mining. And the processes of data mining of archives base on knowledge management has been put forward in this paper.

**Keywords:** Knowledge management; library; digital library; theory

# 基于知识管理的数字档案馆数据挖掘研究

**董洁萍[1]，韩海涛[2]**

[1]天津工业大学图书馆，天津，中国，300160

[2]天津工业大学档案馆，天津，中国，300160

Email: 397638175@qq.com, tjdagh@163.com

**摘　要：**知识经济时代已经来临,对数字档案馆实施知识管理已是时代的要求,是档案馆自身发展的必然,而数据挖掘为数字档案馆知识资源的有效管理作好前置准备并提供技术保障。本文简单介绍了知识管理和数据挖掘技术的相关理论，提出基于知识管理的档案馆数据挖掘步骤。

**关键词：** 知识管理；图书馆；数字图书馆；理论研究

随着知识经济的到来,数字档案馆的知识管理已成为档案学界研究的热点。在数字档案馆的数据挖掘中，如何运用知识管理的管理理念，吸收知识管理在企业界运用的成功经验来指导和优化数字档案馆数据挖掘的实施，以提高数字档案馆的核心竞争力和对环境的应变能力，是数字档案馆应对知识经济时代挑战的明智之举。

## 1　知识管理概述

柯平教授认为：知识管理是确定、收集和传播共享组织中的知识，包括知识的管理和运用知识的管理，来创造、获取和使用知识以增强组织的应变与创新能力的活动。知识管理理论和方法从 20 世纪 80 年代中期产生以后，迅速成为管理学、工商企业界和信息管理领域的研究热点。知识管理作为一种新的理论或方法，是社会经济和技术发展的产物，也是企业和其他组织的内在需求，成为社会、技术等的许多因素共同驱动而产生的理念和方法，档案馆应用知识管理是档案馆馆事业发展的需要，更是档案馆管理改革的需要[1]。

知识管理作为一种新型的管理理念和管理模式,有其独特的优势和特征:

### 1.1 知识管理的创新性

知识管理要求突破旧的理念、制度、策略、方法,达到新的层次。创新的目的是为了发展,它可以带来更

本文为天津工业大学第九届学生科技作品竞赛立项资助项目《基于知识管理的数字档案馆建设研究》(项目批准号：2009187)成果之一。



978-1-935068-40-2 

广阔的前景,是更为科学的发展途径和更为人性化的发展策略。知识创新包括技术创新、管理创新和制度创新。

### 1.2 知识管理的人本性

以人为核心,重视人在管理中的重要作用,人是生产力中最活跃的因素,知识经济时代,知识成为经济发展的原动力,智慧成为创造财富的源泉,谁拥有了知识和智慧,谁就拥有了无限的财富。如何提高人的素质,发挥人的创造力,是知识管理的重要内容。

### 1.3 知识管理的技术性

信息技术为工具,信息技术的每一次飞跃和发展都会给人类带来生产和生活方式的改变,推动学科向前发展。从某种意义上说,正是信息技术的发展才促使知识管理的产生。知识管理带有浓厚的技术色彩,需要信息技术作支撑。

### 1.4 知识管理的深层次性

知识管理要求数字图书馆提供更深层次的信息服务。做到由传统印刷型信息提供为主转向多元化的信息提供;社会性书目与离散性个体化书目两级并存;对文献的提供转向浓缩的有序的情报信息的提供。

### 1.5 知识管理的增值性

知识管理运用先进的信息技术解决了信息富集与知识缺乏的矛盾。知识管理提供更高层次的服务,能够实现知识的增值[2]。

## 2 数据挖掘概述

### 2.1 数据挖掘的内涵

数据挖掘(Data Mining)又称为数据库中的知识发现(Knowledge Discovery in Database, KDD),是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的、但又潜在有用的信息和知识的过程。数据挖掘主要致力于知识的自动发现,是知识发现研究在数据库系统中的延伸。数据挖掘在没有明确假设的前提下去挖掘信息、发现知识,不仅能对过去的数据进行查询和遍历,并且能够对将来的趋势和行为进行预测并自动探测以前未发现的模式,从而很好地支持人们的决策[2]。

### 2.2 常用数据挖掘技术

#### 2.2.1 人工神经网络

仿照生理神经网络结构的非线性预测模型,主要由“神经元”的互联,或按层组织的节点构成,通常由输入层、中间层和输出层三个层次组成,在每个神经元求得输入值后,再汇总计算输入值;由过滤机制比较输入值,确定网络的输出值。

#### 2.2.2 决策树

决策树是一个类似流程图的树型结构,其中每个内部节点表示在一个属性上的测试,每个分枝代表1个测试输出,而每个树叶点代表类或类分布。树的最顶层节点是根节点。目前,在数据挖掘中使用的决策树方法有多种,典型的在国际上影响较大的决策树方法是Qninlan研制的ID3算法。

#### 2.2.3 遗传算法

遗传算法是模拟生物进化过程的计算模型,是自然遗传学与计算机科学相互结合渗透的计算方法。遗传分析应用搜索技术,先找出两个合适的父样本,通过“交叉”“变异”等带有生物遗传特点的操作产生下一代样本,对子样本反复“交叉”“变异”操作直到子样本收敛为此,再找另外两个合适的父样本重复上述过程,就能得到下一代的样本集。由此得到当前样本集较可能的发展方向。

#### 2.2.4 近邻算法

用该方法进行预测的基本概念就是相互之间“接近”的对象具有相似的预测值。如果知道其中一个对象的预测值后,可以预测其最近的邻居对象。

#### 2.2.5 规则推导

根据统计意义上对数据中的规则“如果条件怎么样、怎么样,那么结果或情况就怎么样”,对给定的一组项目和一个记录集合,通过分析记录集合,推导出项目间的相关性。

#### 2.2.6 聚类方法

聚类分析方法按一定的距离或相似性测度将数据

分成系列相互区分的组,它是不需要预定义知识而直接发现一些有意义的结构与模式。可采用拓扑结构分析、空间缓冲区及距离分析、覆盖分析等方法,旨在发现目标在空间上的相连、相邻和共生等关联关系。

#### 2.2.7 可视化技术

可视化技术在数据挖掘过程中的数据准备阶段是非常重要的,它能够帮助人们进行快速直观地分析数据。利用可视化方法,很容易找到数据之间可能存在的模式、关系和异常情况等[3]。

### 2.3 web 数据挖掘的主要流程

通常可以将数据挖掘分为四个步骤:数据预处理、数据挖掘、后续处理和反馈，其挖掘步骤见图 1 所示。

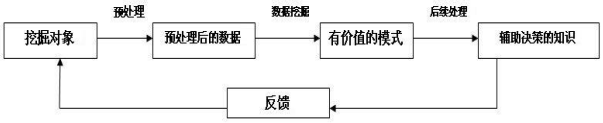


**Figure1. Data mining steps**

第一步:数据预处理。这是用户导航信息挖掘最关键的阶段,数据预处理包括数据清洗、用户识别、会话识别和事务识别 4 个步骤。第二步：数据挖掘是将经过预处理的数据送到数据挖掘算法中，生成有价值的模式。第三步：后续处理是要识别出模式中有用的部分。有很多评估和可视化技术可以用来做此项决策。第四步：要把辅助决策的结果反馈给挖掘对象，以便于优化新一轮的决策。

## 3 基于知识管理的档案馆数据挖掘步骤

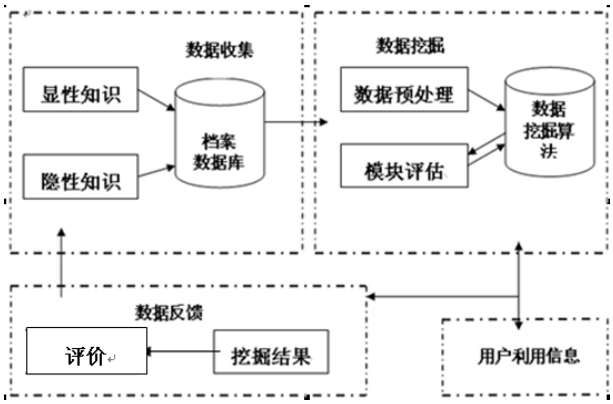


**Figure2. Knowledge management-based data mining steps Archives**

### 3.1 原始数据的收集

#### 3.1.1 显性知识

这是存在于数字档案馆中的固化资源,即记录于一定物质载体上的知识,包括:已数字化的馆藏资源、现行电子文件、检索工具、编研成果,与数字档案馆工作相关的各种法律法规、规章制度、行业标准等,围绕数字档案馆建设所产生的研究成果、技术资料及有助于数字档案馆发展的其它相关知识[4]。

#### 3.1.2 隐性知识

这是存在于数字档案馆中的智力资源,是存在于档案馆行政管理人员、政策法规研究人员、信息技术人员、对外协调人员等头脑中所储备的大量非编码智力资源,包括:各种管理方法、计算机处理技术、处理问题的能力等。由于人是知识管理的核心,是知识管理中最活跃的最主动的因素[5],所以对这部分知识的挖掘也是数字档案馆知识挖掘的重点。

### 3.2 数据挖掘

数据预处理模块的主要功能是对数据源传过来的数据进行初步处理,并对多个数据源的数据进行集成、概化、编码和规约,使之变成数据挖掘方法库中数据挖掘算法可以处理的、一致的、完整的、可理解的数据。这些处理包括:移除和过滤掉冗余及不相关的数据,预测和填充数据中丢失的值,移除噪声数据,对数据进行转换和编码以及处理任何不一致问题。

模式评估模块包括两个过程:模式发现和模式评估。模式发现过程是从数据挖掘方法库中取出各种数据挖掘算法,然后对数据预处理模块得到的数据进行处理。模式分析过程是对用户的使用模式进行更进一步的分析,我们可以运用查询或 OLAP 等方法来进行,然后建立用户的行为和偏好模型,输出用户需要的数据。这个阶段的主要任务是模式过滤、聚集以及特征化[6]。

### 3.3 数据反馈

在参照用户信息的反馈，利用特定的数据挖掘模型形成挖掘结果。用户的利用行为信息包括两方面,利用信息和反馈信息。利用信息是用户为了解决现实问题,满足学术、科研、生产等需求,在实施具体利用行为时所产生的信息,包括:访问内容、访问频率、访问时


978-1-935068-40-2 © 2010 SciRes.

间等,它们反映出用户对数字化资源的个性化、多样化需求及利用规律。反馈信息是在档案利用这一连续活动中,档案利用者发现的问题和情况、提出的要求、意见、评价和效益等[7]。对这些数据的挖掘,可用于对用户未来利用趋势的分析预测,以及提出在此基础上的管理决策,为提高数字档案馆的服务水平提供依据。之后便对挖掘结果进行评价,形成的挖掘结果有可能存在无关的数据,也有可能不满足需求,如果不符合挖掘要求和目的,整个数据挖掘过程就要退回到数据收集阶段,并重复挖掘过程,反之则达到数据挖掘要求,能为数字档案馆知识管理所用,并充实到原有数据库中,实现档案馆的知识创新。

## 4 结语

随着数字档案馆的不断发展进步，知识管理的作用会越来越重要，只有不断提高知识管理的手段、技术、理念等水平才能对数字档案馆数据挖掘工作进行最优化管理。因此,知识管理理论对于数字档案馆数据挖掘工作具有重大意义,研究基于知识管理的数字档案馆数据挖掘是非常必要的。以知识管理的相关理论和数据挖掘方法进行数字档案馆的建设，是数字档案馆在知识经济时代应对机遇和挑战的必然，是数字档案馆应对知识经济时代挑战的明智之举。

## References (参考文献)

[1] Ping Ke，Knowledge Management Application in Library [J].Journal of Library Research ，2003,(9): 8-12
柯平，知识管理在图书馆中的应用研究[J]. 图书馆学研究，2003,(9): 8-12

[2] Chunjia Chi，Zhiyong Mao ，Library based on data mining decision support of book purchasing program [J].Journal of Modern Information，2009,29(7):108-110
迟春佳，毛志勇.基于数据挖掘的高校图书馆图书采购计划辅助决策研究[J]. 现代情报，2009,29(7):108-110

[3] Hui Zhang, Data Mining Application in the procurement of books [J]. Journal of Library and Archives Management, 2006(3):158-159
张晖. 数据挖掘技术在图书采购中的应用初探[J]. 图书与档案管理，2006(3):158-159

[4] Xiaozhong Huang, Jiang Shi ,Digital Archives Knowledge Management based on Data Mining [J]. Journal of Newsletter archives, 2008,(4):58-60
黄小忠，史江. 基于知识管理的数字档案馆中的数据挖掘[J]. 档案学通讯，2008,(4):58-60

[5] Xin Li, Comparison of Information Management and Knowledge Management Analysis [J]. Journal of School of Guiyang Municipal, 2006，5
李昕:信息管理与知识管理比较分析[J],贵阳市委党校学报,2000.19(2)521-523

[6] equal to [2]

[7] Jjiang Shi, Jinfeng Li, File using feedback Problems and Countermeasures [J]. Journal of Journal of Newsletter archives, 2007.3
史江、李金峰:档案利用信息反馈工作的问题与对策探讨[J],档案学通讯,2007.3