广西科学院学报
Journal of Guangxi Academy of Sciences
2010,26(4):520~522
Vol. 26, No. 4 November 2010

一个数字档案馆中的数据挖掘系统工作流程*

A Digital Archives of Workflow for the Data Mining System

罗 艳[1], 黄明初[2], 陆旭安[2], 潘雄伟[2]
LUO Yan[1], HUANG Ming-chu[2], LU Xu-an[2], PAN Xiong-wei[2]

(1.南宁海蓝数据有限公司,广西南宁 530022;2.广西壮族自治区档案局,广西南宁 530022)
(1. Highland Digital Technology INC., Nanning, Guangxi, 530022, China; 2. The Archives Administration of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530022, China)

摘要:根据数字档案馆中的数据挖掘形式与实现方法分析,提出应用在数字档案馆中的数据挖掘系统工作流程。数字档案应用数据挖掘技术处理后,通过对档案情况和用户利用行为信息的分析,了解各档案形成特点、规律和档案利用范围,可以对用户未来利用趋势进行分析预测,为提高档案馆的服务水平提供依据,从而在更深层次上发挥数字档案的作用。

关键词:数字档案馆 数据挖掘 数据分析 工作流程
中图法分类号:TP302.1 **文献标识码**:A **文章编号**:1002-7378(2010)04-0520-03

Abstract: Through a brief analysis and realization methods of the data mining form process for data mining system which applied in digital archives is introduced. After data archives be processed by data mining technique, the future trend of user's archive usage can be predicted and analyzed by the analysis to archives and user's use behaviors information, understand the characteristic of forming, law and available range of each files, which provide a basis for improving service level of archives and exert the function of archive data information.

Key words: digital archives, data mining, data analysis, workflow

在数字档案馆中,如何从浩如烟海的大量数字化资源中提炼、挖掘出有价值的,对数字档案馆的知识积累、知识创新有着数据支撑作用的有效信息,是未来数字档案馆建设所面临的重要课题[1]。数据挖掘正是组织和发现数字档案馆中知识资源的有效途径,为数字档案馆实施知识管理创造了条件。这里的数据挖掘不能看成是纯粹的信息处理技术,它是对信息处理技术集群进行协调和管理的方法和策略。基于知识管理的数字档案馆中的数据挖掘是以网络和数字化资源为基础,立足于多种信息技术的协调和配合,以实施挖掘算法和挖掘模型为手段,以组织和发现数字档案馆中已存在的知识资源,为实施知识管理提供管理对象为目的,让数字档案馆有效利用知识,实现知识创新的过程。本文在简要分析数据挖掘的形式与实现方法的基础上,介绍在数字档案馆中应用数据挖掘系统提炼挖掘有效信息的工作流程。

1 数字档案馆中的数据挖掘形式与实现方法

数据挖掘可以分为描述型数据挖掘和预测型数据挖掘。描述型数据挖掘一般是对数据中存在的规则做出描述,通常根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、宏观的知识,通过对数据的概括、精炼和抽象反映同类事物共同性质。预测型数据挖掘通过对现有数据的分析和处理,得到某类元组中某些属性的内容,或是预测出某

收稿日期:2010-08-13
修回日期:2010-10-11
作者简介:罗 艳(1983-),女,助理工程师,主要从事数字档案管理研究。
*2009年度国家档案局科技项目(项目编号:2009-X-07)资助。

类档案未来形成和使用的规律等[2]。主要有分类和关联两种方法进行描述型和预测型数据挖掘。挖掘得到的知识和结果通过各种直观的图形表示出来,如饼图、柱状图等,以此来获取用户对系统的信任,提高挖掘结果的使用效果。

1.1 分类方法

分类是通过对数据库中属性的分析,将元组划分为不同各类的过程[3]。具体的是,用一组特征不同的类别为一个数据集合中的数据进行分类,然后找出描述这些数据的模型,并根据这个模型将数据划分到不同的类别中,利用这个模型可以预测未知的数据。

应用分类方法可以通过已有用户档案数据库中的数据,揭示用户特征和用户利用行为之间的关系,并按照影响用户行为的程度对这些数据进行分类,用来预测未来的用户的利用需求,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。例如,应用数据挖掘分类方法对近 3 年下半年的档案利用数据与档案利用人数数据进行挖掘,结果(图 1)发现,3 年来每到这个时期由于纪念抗美援朝战争胜利,党和国家会相对出台一些政策,大量史料编修人员和争取个人待遇的老同志查档情况较多。因此,可以预测出今年下半年抗美援朝类档案的使用率将相对偏高。

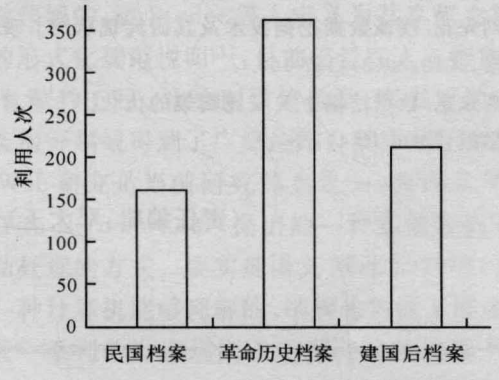

图 1 2007～2009 年下半年档案利用情况

在分类过程中可以选取被广泛研究和应用的决策树(Deeis Tree)算法[4]实现档案数据挖掘。决策树算法的实现过程如图 2 所示。

1.2 关联规则集方法

数据挖掘的关联规则集方法描述数据库中数据项之间存在的相关特性,即挖掘出隐藏在数据项之间的相互关系。具体来说,若其中两项数据或多项数据存在着某种关联,其中一项数据就能依据其它数据进行预测。例如,在档案工作中,能经常发现 A 类人群经常使用 B_1 类档案,或是 C 类人群使用 B_2 类档案的同时又有 75%的人使用了 B_1 类档案,通过数据挖掘得出这些相互关系,当用户检索某类档案时,可以对应提供具有关联的相关其它档案,提高档案利用率。

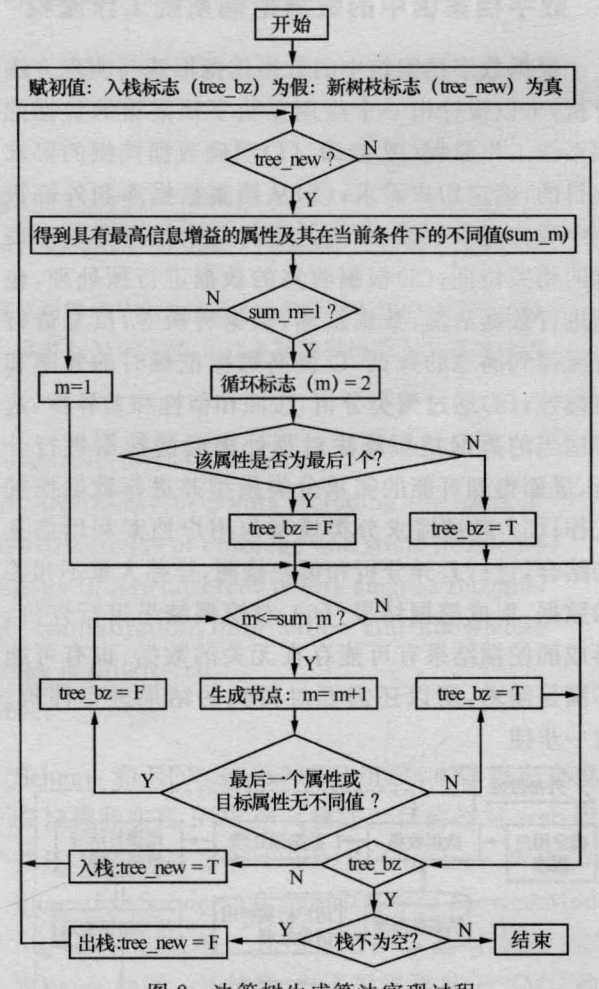

图 2 决策树生成算法实现过程

挖掘关联规则集可以采用的 Apriori 算法[5]。Apriori 算法是一种较有影响的挖掘单维关联规则频繁项集的算法,主要是利用迭代方法逐层搜索。其实现过程如下。

(1)确定频繁项 1^- 集,记作 L_1;

(2)自然连接(natural join)L_1 产生 2^- 项集;

(3)利用"任何非频繁的$(n-1)^-$项集都不可能是频繁 n^- 项集的子集"这一性质从 2^- 项集中去除非频繁的;

(4)扫描事务数据库,从剩余的 2^- 项集中确定频繁 2^- 项集 L_2;

(5)由 L_2 自然连接产生 L_3,由 L_3 自然连接产生 L_4,……直至找不到频繁$(n+1)^-$项集;

(6)利用频繁 n^- 项集的结果产生关联规则。假

设最终结果为频繁3⁻项集{ABC}，可以得到A⇒B∧C，B⇒A∧C，C⇒A∧B，A⇒B∧C，A∧C⇒B，B∧C⇒A六种情况，通过计算每种情况的支持度和置信度找出强关联规则。

2 数字档案馆中的数据挖掘系统工作流程

根据数字档案馆中的数据挖掘形式与实现方法分析，可以设计出一个应用在数字档案馆的数据挖掘系统工作流程(图3)为：(1)明确数据挖掘的要求和目的，确定用户需求；(2)从档案数据库和外部数据中收集提取数据，并对其进行概念描述归纳出需求的相关特征；(3)根据收集的数据进行预处理，包括进行数据清洗、数据推测、数据转换等，反复进行直至得到满意的数据，以提高数据挖掘时的效率和准确性；(4)通过聚类分析，按照相似性和差异性，选择适当的数据挖掘算法对预处理后的数据进行分析，进而得到可能的需求分类模型并进行数据挖掘工作；(5) 通过需求分类模型与用户档案利用信息的结合，进行差异分析和偏差检测，排除大量不相关的数据，形成挖掘结果；(6) 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，所以还需要对得到的结果进行评价。这一步使用的方法包括用数据供稿进行验证，也可以根据常规的经验进行一些判断，一般由数据挖掘具体操作而定。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，并充实到决策知识库中，实现档案馆的知识创新和重用。

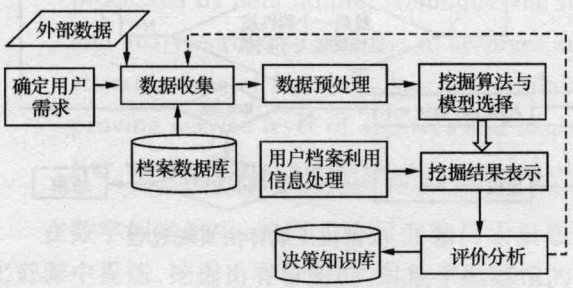

图3 数字档案馆中的数据挖掘系统工作流程

3 结束语

数据挖掘可以快速有效地分析和处理来自数字档案馆内外的海量数据和信息，使隐性知识显性化，显性知识结构化。将数据挖掘应用在数字档案馆的建设中，可以为数字档案馆的科学管理和服务水平的不断提高提供有力的支持，使档案馆向着知识化的方向发展。随着信息技术的不断深入和挖掘算法的不断改进，数据挖掘必将与数字档案馆的知识管理结合得更加紧密，显现出更加强大的生命力。

参考文献：

[1] 黄小忠. 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.

[2] 宇然. 数据挖掘技术与档案管理[J]. 兰台世界，2002(8)：24.

[3] 苏新宁. 数据挖掘理论与技术[M]. 北京. 科学技术文献出版社，2003：25.

[4] 刘先花. 浅谈数据挖掘技术及其研究现状[J]. 现代情报，2010，30(3)：167-169.

[5] 高永惠. 数据挖掘中关联规则集的优化[J]. 吉首大学学报，2010，31(4)：38-42.

(责任编辑：邓大玉)

科学家首次实现信息转化为能量

日本研究人员在实验室让一个直径为287纳米的聚苯乙烯小球沿电场制造的微小旋转阶梯向上爬动，并将小球拍照。小球可以随机朝任何方向运动，由于向上爬会增加势能，因此其往下一层的概率更大，如果不人为干扰，小球最终会掉至最底层。在实验中，当小球沿阶梯向上爬一层后，研究人员就使用电场在小球爬上的那层阶梯加一面“墙”，让小球无法回到低的那一层，这样小球就能一直向上爬。该小球能爬阶梯完全由“自己的位置”这一信息所决定，研究人员无需施加任何外力(比如注入新能量等)，仅需一个感应系统(比如摄像机)。另外，他们也能精确地测量出有多少能量由信息转化而来。

1871年，英国物理学家詹姆斯·克拉克·麦克斯韦提出了“麦克斯韦妖”设想：一个绝热容器被分成相等的两格，中间是由一种机制控制的一扇活板门，容器中的空气分子做无规则热运动时会撞击门，门则可以选择性地将速度较快的分子(温度较高)放入其中一格，将速度较慢的分子(温度较低)放入另一格，这样，两格的温度就会一高一低。麦克斯韦认为，整个过程中使用的能量就是“分子是热的还是冷的”这一信息。日本研究人员首次在实验中实现了“麦克斯韦妖”设想。 (据科学网)