计算机前沿技术在科技管理领域的应用研究

朱东华　荆　雷　徐建国

(北京理工大学管理与经济学院,北京　100081)

摘　要　数据挖掘,在科研领域也称为数据库中的知识发现,是一个利用各种分析工具在海量数据中发现模型和数据间关系的技术。Web挖掘是数据挖掘技术在Web上的应用。Web挖掘处理的对象主要是半结构化的数据,这是它最主要的特点。目前我们的研究工作正是利用该项技术来实现科技信息的网络动态监测和信息的自动获取。自然语言处理技术,是指利用特定的算法使计算机系统能够理解和生成自然语言。它是人工智能活跃的研究领域之一,是人文科学和自然科学的交叉性学科。

关键词　数据挖掘;知识发现;Web挖掘;自然语言处理;文本挖掘

中图分类号　F204　　**文献标识码**　A　　**文章编号**　1001-7348(2003)08-023-03

1　大型科技文献和专利数据库环境下的数据挖掘技术

数据挖掘,在科研领域也称为数据库中的知识发现,是一个利用各种算法和工具在海量数据中发现模型和数据间关系的过程,是从原始数据库中发现有效知识的过程。数据挖掘的结果可以用来预测未来的趋势。数据挖掘技术属于当今国际前沿领域"人工智能"的范畴,是近几年新的研究热点。

1.1　数据挖掘技术研究的兴起

随着大型科技文献数据库和专利数据库系统在科技创新中的广泛应用,全球范围内的科技数据信息急剧增大。传统的数据库技术提供了对科技信息的高效存储、快捷检索,但无法满足当前"高新技术监测分析技术"进行更深层次数据分析的需求。面对庞大的科技信息数据库,人们需要快捷地从中提取出有用和有效的知识。

对数据挖掘技术的研究始于20世纪70年代的知识发现,这一概念一经提出,就受到了普遍关注。研究人员对其理论和技术进行了深入细致的研究,在各应用领域进行了广泛的应用实践,研究工作取得了很大进展。数据挖掘技术的出现,为自动和智能化地把海量数据转化成有用的信息和知识提供了手段。

1.2　数据挖掘的本质

数据挖掘是从大型数据库中抽取隐含的、以前所未知的、具有潜在应用价值的模式的非平凡过程,它在数据库中自己寻找隐含的模型,在本质上是一种"归纳"。这里的模式也就是指所要发现的知识和数据库中有用的信息,它是对数据载体所包含的信息更高层次更抽象的表述。数据挖掘得到的模式必须是正确的和具有创新性的,否则就不能称之为成功的数据挖掘。非平凡是指在数据挖掘中,发现知识的过程和算法应是事先未知的,该过程应具有不确定性。确定的计算过程或计算公式提取的模式,一般称之为平凡知识,平凡知识不是数据挖掘的目标。

数据挖掘基于的环境是大型科技文献数据库和专利数据库,它应用的对象是大规模数据集,待处理的数据规模可能达到GB、TB,甚至更大。在此部分的论述中,我们的研究工作主要针对结构化科技数据库。

1.3　数据挖掘方法

数据挖掘按照应用的技术方法可分为:基于关联规则的发掘方法、基于粗糙集的发掘方法、基于神经元网络的方法、基于统计的发掘方法、数据的可视化发掘方法和文本发掘等等。

(1)关联规则挖掘。所谓关联规则,是指同一个事件中出现的不同项之间的相关性,是指数据对象之间的相互依赖关系,用来描述一组数据的密切度。挖掘关联规则和关联分析就是寻找上述相关性,发现存在于大量数据集中的隐含的关联性。

从数据库中发现关联规则近几年研究得最多,是数据挖掘领域的一个研究热点。其内容包括单一概念层次关联规则的发现和多个概念层次的关联规则的发现。处理对象和过程的概念层次越多,数据挖掘所发掘的知识和信息越具体,实际上这是个逐步深化发现知识的过程。目前典型的关联规则采掘算法有Apriori和DHP两种,它们都是基于数据库遍历算法的。

(2)数据的可视化挖掘方法。数据可视化技术采用直观易懂的方法帮助人们理解数据库中的数据,它以可视化的形式将数据和数据挖掘得到的模式呈现在人们面前,使用户能够完全理解挖掘后产生的数据和所发现的知识。

数据可视化技术一般采用表格、直方图、散点图或自然语言文本报告等可视化和

基金项目:国家自然科学基金重点资助项目(编号:70031010)。

作者简介:朱东华,教授,博士导师,研究方向为科技管理、技术预测、数据挖掘、人工智能、信息系统;荆雷,北京理工大学管理与经济学院;徐建国,清华大学经济管理学院博士生。

收稿日期:2002-12-20

形象化的方式来展现多维多元的数据，用可视化的方法可以方便地将尽可能多的内容同时表示出来。现存的适于进行大型数据库可视化采集的重要技术主要有以下几种：像素定位法、几何法、基于图标的方法等等。

(3)基于粗糙集的发掘方法。粗糙集理论是波兰数学家Z.Pawlak在1982年提出的一种分析数据的数学理论，该理论主要用来处理含糊和不确定性问题。其特点是处理问题之前，只需用户提供必要的数据集合和信息，然后直接从待解决问题的描述集合中找出问题的内在规律性，方法非常简单便捷。这是粗糙集理论最重要的优点。

近年来，粗糙集理论研究和应用都取得了飞速的发展。利用粗糙集理论可以处理的问题包括数据简化、数据相关性发现、数据意义的评估、数据的近似分析等。该理论是以"分类"为基础的，分类即等价替代关系，对知识进行理解也就是对数据的划分和等效替代。

(4)文本挖掘。科技信息的网络动态监测和信息自动获取技术领域的研究是基于自由文本分析(Free TextAnalysis)和自由文本向结构化科技数据库转化等基础技术的。自由文本分析是一个动态的分析过程，属于当今前沿计算机领域—文本挖掘的范畴。

文本挖掘(Text Mining)主要处理半结构化、无结构化和字符型数据。它将数据挖掘技术与信息检索技术相结合，开扩了数据挖掘新的应用领域。其特点是能够更加有效地对文本数据(例如Web页面)进行分析，从而弥补了信息检索技术的缺陷与不足。

目前，对文本挖掘的理论方法和技术实现国内外都在进行深入研究和探讨，我们的研究目标是：利用自由文本分析(Free Text Analysis) 和自由文本信息的结构化等一系列动态分析过程和技术，研制不间断、长期运行的Internet www网页动态监测器，对技术方向系列动态识别、记录和分析，并在结构化的科技文献和专利数据库中寻求相关信息数据支持。

国外的研究成果已经有了一定数量的文本挖掘工具，并且出现了很多融合文本挖掘思想和技术的应用。Semio公司研发的SemioMap工具，可以提供自动的文本处理；IBM公司出品的智能化文本挖掘器(Intelligent Miner for Text)，适合大型软件公司的开发人员使用；Brightware公司的产品Brightware，是一个自动的电子邮件阅读和解释系统。

1.4 数据挖掘的过程

(1)定义问题和目标。明确所要解决问题的性质和数据挖掘的目标。通过学习，熟悉应用领域和问题所涉及的背景知识。

(2)建立目标数据集。根据需求从数据库中提取相关的数据，建立一个独立的目标数据集。

(3)数据预处理与清洗。从目标数据集(即数据挖掘库) 中除去明显错误的数据和冗余的数据，去除噪声或无关数据，去除空白数据域，并进行数据清洗。

(4)数据转换。通过各种转换方法将数据转换成有效形式，为今后的数据开采做好准备工作。

(5)选定数据挖掘算法。根据具体实际情况，选择特定的数据挖掘算法 (如汇总、分类、回归、聚类等)，包括选取模型和参数两项内容。

(6)实施挖掘。用所选择的算法实施数据挖掘工作，并将结果用一定的方法(例如可视化技术)表达成易于人们理解的形式。

(7)模式解释。对发现的模式(知识)进行解释、评估和价值评定。

2 WWW科研、高技术产品网页动态监测

除结构化的数据库资源外，我们的项目重点开展对WWW网页科技信息资源的开发、(自动)动态监测分析。此项研究主要针对非结构化科技信息进行动态扫描监测。

这是一项具有挑战性和较高应用价值的研究，其成果还有望应用在金融、贸易等其它领域。此部分的研究内容具有较大的创新力度，"科技信息的网络动态监测和信息自动获取技术" 属于当今国际前沿领域"WWW Mining"的范畴，我们提出将"利用结构化科技数据信息与网页科技信息交互支持来实现对科技信息动态监测"方案是一个全新的思路。

2.1 WWW Mining的基本概念

WWW Mining，即Web挖掘，是数据挖掘技术在Web上的应用，它是从Web文档的相关资源和Web行为活动中抽取感兴趣的潜在的有用模式和隐藏信息。目前比较可行的方法是将数据挖掘的思想和方法引入WWW信息处理领域，实现科技信息的网络动态监测和信息自动获取。

Web资源可以看成是一个非结构化的数据库，因此Web挖掘较基于结构化数据库的数据挖掘更加复杂，其应用前景更加广阔。Web挖掘的潜力在于应用存在的和最新的数据挖掘算法，分析Internet服务器上的科研信息和高技术产品的外部数据，实现对科技信息的动态监测。

基于Web的信息挖掘可以使用户快捷方便地从WWW网页上获取具有价值的高质量的科技和科研产品信息。我们目前的研究工作在国内和世界上都属于开创性的研究，工作得到不断的开拓进展，研究成果具有很强的实用价值。

2.2 Web挖掘的分类

一般地，Web挖掘可分为3类：Web内容挖掘 (Web content mining)、Web结构挖掘(Web structure mining) 和Web用法挖掘(Web usage mining)。

Web结构挖掘的目的在于发掘出Web页面结构体系的模式，在此基础上对页面进行归类，从而找到科技领域中权威的和重要的页面。Web页面的超链接反映了Web文档间引用和被引用的关系，一个页面被别的页面所引用或指向的次数多少表明了该页面在领域内的重要程度。

Web内容挖掘又可分为Web页面内容挖掘和搜索结果挖掘两种。Web页面内容挖掘是指直接从Web文档的文本内容中提取出有用知识的过程；搜索结果挖掘是指在搜索引擎的基础上对数据作进一步的处理，发现有用的知识。

Web用法挖掘包括一般访问模式跟踪和个性化的使用方法跟踪. 一般访问模式跟踪侧重了解用户的群体访问模式和访问倾向性，从而改进网站的后续建设。而个性化的使用方法跟踪则倾向于分析单个用户的偏好。Web用法挖掘最广阔的应用前景在于电子商务领域的应用，电子商务网站可以利用此项技术为客户提供个性化的服务和产品。在我们所研究的科技信息监测系统中，可以利用此项高新技术通过网页动态监测发现目标技术领域的关键技术和重要产品。

3 基于高新技术监测分析的自然语言处理技术

我们研发的面向我国重要技术管理部门应用环境的高新技术动态监测分析系统，可以在结构化的科技文献和专利数据库中跟踪某项新技术，利用网页动态实时监测器对Internet WWW网页进行长期的、不间断的监测分析。获取的信息，即数据的集合，需要送分析器利用成熟的分析方法和模型进行自然语言分析与处理。

自然语言处理技术，是指利用一定的算法使计算机系统能够理解和生成自然语言。目的在于建立起一种机器与自然语言之间密切而友好的关系，使之进行高度的信息传递与认知活动。

我们的研究工作旨在发展自然语言处理中的文本生成(Text Generation)技术，研制出"技术预警、评估定量分析报告"的自动生成软件系统。该研究成果对在实用阶段根据问题迅速改进模型具有重要意义，并易于在网络上供科技管理人员使用。

3.1 自然语言处理技术的发展

自然语言处理的研究始于机器翻译，1946年，随着第一台ENIAC计算机的问世，英国的A.Donald Booth (布斯) 和美国的W. Weaver (韦弗) 就开始了机器翻译的研究。1954年，在麻省理工学院组织的第一次机器翻译会议上，世界上首次自动翻译运行并取得了初步成功，引起了国际上机器翻译研究的热潮。但是机器翻译的问题很复杂，由于低估了它的困难程度，初步的成功形成了一种假象，以致于又走向了它的反面，出现了低谷。

大约到了20世纪70年代，涌现出了一大批新的理论与方法，ATN文法分析、LUNAR模型、SCHRDLU模型、语义网络理论等。这些理论的不断发展，将自然语言处理的研究引向非常广阔的应用领域。近年来，新一代计算机和智能机器人的研究开发，使得自然语言处理技术成为当今人工智能中最活跃的研究领域之一。

3.2 自然语言处理的概念和过程

自然语言处理(理解)，有时也称为计算语言学，是计算机科学中一个富有挑战性的课题，是人工智能早期活跃的研究领域之一。自然语言处理研究的目标是寻找一种计算机模型(算法)，这种计算机模型能够象人那样理解和分析自然语言。

自然语言处理是语言学、逻辑学、生理学、心理学、计算机科学和数学等相关学科发展和结合而形成的，是人文科学和自然科学的交叉性学科。自然语言处理系统包括自然语言人机接口、机器翻译、文献检索、自动文摘、自动校对、语音识别与合成、字符识别等等。

对自然语言的理解可以分为3个层次：词法分析、句法分析和语义分析。词法分析是通过分析词汇的各个要素，从中获得语言学信息；句法分析是对句子和短语的结构进行分析；语义分析是通过分析找出词义、结构意义及其结合意义，从而确定语言所表达的真正含义和其所表达的知识。在自然语言理解中，语义分析是研究的重点，越来越受到人们的关注。

3.3 机器翻译

机器翻译，是让计算机模仿人类翻译语言的思维过程，把一种自然语言转变成另一种自然语言的过程。自然语言研究初期阶段的工作主要是针对机器翻译。机器翻译系统是典型的、其应用价值也是最明显的自然语言处理系统。

机器翻译的过程一般包括3个阶段：原文输入、原文分析和译文输出。原文分析包括两个阶段，查词典和进行语法分析。机器翻译的逻辑过程又可划分为：分析阶段、转换阶段和生成阶段。

当前机器翻译研究的重点是，在理论研究的基础上，建立一种形式系统，该系统不仅可以用来表达不同的语言知识，而且要表述出不同语言内部表示之间的可计算性，也就是可以通过特定的计算机算法和程序模块进行自然语言之间的准确转换。

3.4 自然语言处理技术在搜索引擎中的应用

将自然语言处理技术嵌入和应用于搜索引擎技术当中是自然语言处理技术今后的一个重要发展方向。近年来自然语言处理技术发展非常迅速，特别是机器翻译与语义理解被广泛应用于搜索引擎，取得了很大成功，有效地扩展了自然语言处理的应用领域。

应用了自然语言处理技术的搜索引擎称之为智能搜索引擎。通过对嵌入了机器翻译功能的智能搜索引擎的研究，将使得网络用户可以使用母语搜索非母语的网页，并以母语浏览搜索结果。语义理解通过将语言学及人文科学的研究成果同计算机技术结合在一起，实现了计算机智能化地对语言在语义层次上的理解和认知。

4 结论与展望

本文全面综述了计算机前沿技术在科技管理领域的应用与研究现状，对数据挖掘、Web挖掘和自然语言处理3个计算机技术领域进行了全面深入的综合论述。并对我们目前的研究工作和工作成果也进行了概述性的简单介绍。

计算机前沿技术在科技管理，甚至是整个管理科学界的应用前景都是被人们所非常看好的，"中英文兼容、关键词互译平台环境下高新技术动态监测和分析软件系统"是我们研究工作的最终目标和成果。这一研究具有很高的应用价值，整个项目方案目前在国际上属首创、开拓性的研究工作。

参考文献

1 Donghua Zhu, Alan L Porter. Automated extraction and visualization of information for technological intelligence and forecasting.Technological Forecasting & Social Change.69(2002)495-506

2 王咏，倪波，丁尉，承斌.20世纪计算机软件技术的发展[J].现代图书情报技术，2000(6)

3 吕安民，林宗坚，李成名.数据挖掘和知识发现的技术方法[J].测绘科学，2000(12)

4 邹涛，黄源，张福炎.基于WWW的文本信息挖掘[J].情报学报，1999(8)

(责任编辑 曙 光)