知识元挖掘

温有奎　徐国华
赖伯年　温　浩　著

西安电子科技大学出版社
http://www.xduph.com

知识元挖掘

温有奎　徐国华
赖伯年　温　浩　著

西安电子科技大学出版社
2005

图书在版编目(CIP)数据

知识元挖掘/温有奎等著.

—西安：西安电子科技大学出版社，2005.4

ISBN 7-5606-1494-9

Ⅰ.知… Ⅱ.温… Ⅲ.知识学—研究 Ⅳ.G302

中国版本图书馆 CIP 数据核字(2005)第 014122 号

策　　划　戚文艳

责任编辑　杨　璠　戚文艳

出版发行　西安电子科技大学出版社(西安市太白南路 2 号)

电　　话　(029)88242885　88201467　　邮　编　710071

http://www.xduph.com　　E-mail：xdupfxb@pub.xaonline.com

经　　销　新华书店

印刷单位　西安文化彩印厂

版　　次　2005 年 4 月第 1 版　2005 年 4 月第 1 次印刷

开　　本　787 毫米×960 毫米　1/16　印张 16.25

字　　数　209 千字

印　　数　1～2 000 册

定　　价　25.00 元

ISBN 7-5606-1494-9/Z·0017

XDUP　1765001—1

＊＊＊如有印装问题可调换＊＊＊

本社图书封面为激光防伪覆膜，谨防盗版。

内容简介

本书对知识产生、知识组织、知识转换、知识利用、知识更新、知识增值等全过程做了系统研究；对知识管理中“新知识元增加”这一核心环节进行了专题性的深入探索；对知识元、知识元抽取、知识元链接的理论与实现方法有所创新；给出了数值型知识元抽取的试验过程和结果；提出并讨论了认知的三维结构理论，给出了知识与信息之间的变换关系。本书的研究内容是当前文本知识挖掘研究的最新成果。

全书共分为11章。第1章为制约人类利用知识的瓶颈。第2章为知识发现的机理研究。第3章为文献标引与检索的进展。第4章为知识创新与增值的知识链。第5章为知识管理的革命。第6章为知识元链接理论。第7章为信息与知识变换。第8章为知识元的认知理论。第9章为知识元标引。第10章为基于创新点的知识元挖掘。第11章为基于XML的知识元本体推理。

本书取材于国内外最新资料，并总结了作者的研究成果，内容新颖，反映当前该领域的研究水平。本书对从事知识管理、知识发现、知识挖掘的科技人员具有重要的参考价值，可用作计算机、信息技术、图书情报、档案管理、企业管理等专业的硕士、本科高年级学生的教材或参考书。

本专著得到下列基金资助：

国家社会科学基金项目（编号：02BTQ011）

国家自然科学基金项目（编号：70373046）

前　言

当人类进入21世纪时，知识已成为主要的经济资源和占支配地位的——甚至可能是惟一的——竞争优势之源泉。计算机、网络和通信技术的高速发展，带来了信息处理在整个社会规模上的迅速产业化。预计到2020年，人类生产的信息总量将以每73天翻一番的惊人速度发展。《大趋势》的作者John Naisbett疾呼：我们正在被信息所淹没，却又饥饿于知识。

知识的控制单位长期停留在文献这一级上，而人们对知识的需求一般不是以文献为单位的。将知识的控制单位从文献深化到文献中的知识元，实现知识元的链接，这是知识管理上的一场革命。它涉及人类对知识的理解，知识的机器发现、自动抽取、语义表示，知识的检索以及知识的利用(即知识的增值性标引问题)，是信息服务向知识服务过渡的基础。

如何进行知识标引？有文献报导说：目前世界上进行知识标引的报导基本为空白，而基于知识的标引则已经有报导可见。没有人进行知识标引的原因，可能是在文本中进行知识挖掘，比在结构化的数据记录中进行知识发现和数据挖掘更困难(尽管更有意义)。但是，作为今后智能文本处理发展的一个方向，我们应该重视该方面的研究，也许我们可以倡导这样一个方面。

正在兴起的网格技术，为人们对知识信息的需求由文本单元向知识单元深度发展提供了实现的可能性，同时也要求人类采用新的知识组织方式来建立知识管理的平台。传统的图书馆学、情报学已难以完成新的历史任务，网络技术将会带来知识管理的革命，并将引发知识管理学的诞生。知识网格将会成为知识创新服务的大平台。知识网格平台上知识存放的形式是什么？知识网格平台将如何实现知识创新服务？这是时代对计算机科学、信息科学、认知科学提出的挑战性课题。

我们对此问题进行了探讨并认为，从知识标引的阶段开始，把知识分解为最小独立的“知识元”，建立以知识元为单位的“知识元自由集成系统”，是解决知识管理问题的本质和关键。本书探讨了以下问题：

(1) 构造知识元结构。对人类全部文明成果在现有的学科分类体系的范围内，进行系统的整理、甄别、认定，以确定各学科的基本知识元。可组织专家就若干个成熟的基础学科的知识元认定进行实验。

(2) 开展知识链理论与方法的研究，实现知识组织结构由等级式向网络式的转变。研究知识元的链接，实现由知识元组成新的知识系统。知识元之间的不同层次、不同学科的链接，是实现新知识生产、知识传播及知识有效利用的核心。知识链的建立应在传统的图书情报分类的基础上进行，并且深化到科学知识分类上。

(3) 建立知识元平台。知识元的独立性与知识元的链接性是知识创新的途径之一。知识链的实现将依赖于知识元平台的建立。知识元平台将是实现知识创新革命的前提，将构建起整个科学分类体系，包括对诸多的综合科学和交叉学科框架的构建，使人类的知识元成果系统化、有序化。知识元的独立性、拓扑性与链接性是构建知识元平台的有效方法。

(4) 提高隐性知识向显性知识编码转变的技术。隐性知识向显性知识编码转变是知识生产、知识学习和知识利用的基础。知识创新标引与创新检索同用户需求具有耦合共振性。解决知识创新标引与创新检索是知识增进的推动力。应用网格技术建立知识链与知识网络结构，达到对知识元的任意存取。

(5) 知识网格将实现无链接的知识元的自由集成。输入需求信息，输出特定知识是推动知识管理革命的目标。知识网格的建立远远比计算网格和信息网格困难，决不能只靠计算机专家来实现，而必须由情报管理专家、科学管理专家以及各学科的专家通力合作，才有可能实现。

(6) 建立“知识管理学”。开展“知识管理学”研究，并在条件相对成熟的时候，考虑建立“知识管理学”。在原有的文献分类、主题标引

等课程的基础上，建立知识管理学学科。在有关高校开设有关课程，推动对它的深入研究。

本书对知识创造、知识组织、知识利用的知识链过程进行了较为全面的研究，抓住了知识链中知识创新的最本质的东西——知识增量，即新知识元的增加是知识系统中产生知识增量的这一普遍规律，并将信息与知识的有机联系区分开来，对知识元进行了基础性的研究。

本书对汉字语言学进行了研究，认识到表达知识的层次，由字、词、短语、小句、句子、段落、语篇逐渐提升；字、词、短语可以表达概念，但不能表达知识。句子是表达知识的最基本的单位。于是，以知识元为单位的知识标引的研究转入发现句子或小句组成的句群中的知识元对象结构。

本书还从认知论的观点对卡尔·波普尔的哲学的知识观点和理论进行了研究，认为不仅知识与信息是连接三个世界的纽带，计算也是连接三个世界的纽带，由此推论出信息与知识是两个既相互联系又有所不同的东西，信息通过计算变换成知识。这一理论为知识元标引提供了理论根据。

第 1 章探讨制约人类利用知识的瓶颈问题。目前的信息采集、传输、检索的传统理论与方法已成为制约人类利用知识的瓶颈。科学技术的创新发展将会把人类带入智能化的社会。知识创新将会成为社会进步永恒的推动力。知识发现、知识创造、知识组织、知识应用是知识创新链中不可缺少的关键环节，寻找新的知识创新链的有效方法与工具将成为认知科学、信息科学、计算机科学、知识管理科学领域共同努力的目标。

第 2 章探讨知识发现的新理论。知识标引与知识检索的知识链研究的目的在于寻找新的理论和技术，以解决计算机自动发现知识的方法，尤其是从人工智能的角度对文本知识发现的理论和技术作进一步的探索。认为文献知识标引与检索同文本知识发现的理论和技术相结合将是一种趋势。

第 3 章探讨文献标引与检索的进展。文献中的数据、信息和知识都是人类对符号的赋值所表现出来的内容含义。其差别在于数据是对事实性的表达，信息是对概率性的表达，而知识是对规律性的表达。显然三者所体现出的人类的智能程度依次增加，反映出人类认知程度的层次。情报检索语言是一种人工语言，是表达一系列概括文献情报内容的概念及其相互关系的概念标识系统。因此，它们都是建立在概念逻辑的基础上的。

第 4 章探讨知识创新与增值的知识链。人类的知识创造、知识组织、知识应用活动形成了知识循环进化的知识链。知识标引属知识链环节中的知识组织环节。研究知识标引的目标是要探索知识如何组织才能被人类有效地检索、应用和再创造，起到知识增值的作用，其理论与方法本身就是一种知识探索和知识创新的过程。

第 5 章探讨知识管理的革命。近来有专家指出，脱胎于传统图书馆职能的数字化图书馆，本质上还只是一本一本“堆积”起来的数字化图书馆。正在兴起的网格技术，为人们对知识信息的需求由文本单元向知识单元深度发展提供了实现的可能性。传统的图书馆学、情报学已难以完成新的历史任务，网格技术将会带来知识管理革命，并将引发知识管理学的诞生。

第 6 章探讨知识元链接理论。网格环境下的知识如何组织才能便于想要的时候临时集成？网格为探讨知识标引与检索的新理论和方法提出了挑战和机遇。为此，我们提出了知识以知识元标引与检索的知识元链接理论，探索文本知识元的分布规律、知识元的组成结构、知识元的抽取机理、知识元的本体表示以及知识元的面向对象的软件实现等新的研究。

第 7 章探讨信息与知识变换。本章提出了信息与知识谱变换的理论，把知识或知识结构看成是由不同谱线的信息组成的，用知识元组织知识的理论和方法。

第 8 章探讨知识元的认知理论。信息与知识是两个不同领域的对象，存在变换关系，它是人类认知过程的两个基本元素。如何实现信

息与知识的变换，是实现认知过程飞跃的关键环节。为此，本章从“三个世界”的哲学概念讨论这一问题，提出了文本中知识元的发现、抽取理论。知识元的发现是通过计算向导信息与知识元的关系来实现的。本章还给出了判断特征标识与知识元的内容的相关计算方法。

第9章研究知识元标引。知识元标引是实现知识标引与检索的核心与具体化。本章进行了具体的知识元类型归类分析、标引规则、人工知识元抽取、软件知识元抽取、人工知识元修改、知识元面向对象表示及知识元对象链接等试验。

第10章研究基于创新点的知识元挖掘。本章对科学论文的创新性和科学性及其表现形式进行了分析，提出了基于创新点构建知识元以解决论文创新知识的有效发现和利用的方法，并采用人工和软件结合的试验方法对基于创新点的知识元挖掘进行了研究。

第11章研究基于XML的知识元本体推理。XML和ONTOLOGY工具为实现文本知识元软件标记及知识推理提供了通用的基础平台。本章研究在XML平台上知识元的本体描述、软件抽取、链接和推理，探讨知识服务模式。

本书是在国家社会科学基金项目(02BTQ011)和国家自然科学基金项目(70373046)研究的基础上完成的。为此，特向国家社会科学规划办、国家自然科学基金委员会表示衷心的感谢。

由于涉及到知识元挖掘的领域和学科非常广泛，我们希望通过知识元的发现、挖掘、组织、利用来推动信息服务向知识服务过渡。此项研究涉及知识管理的基础问题。知识元的提取属于人类的智能行动，有很大的难度，现在的研究还只是刚刚开始。希望各位专家和读者批评指正。

温有奎撰写了本书的全部内容，并和徐国华、赖伯年、温浩一起对书稿进行了讨论和修改。

温有奎

2004年6月

目　录

第 1 章　制约人类利用知识的瓶颈

人类进入 21 世纪后，知识已成为主要的经济资源和占支配地位的——甚至可能是惟一的——竞争优势之源泉。[1]计算机、网络和通信技术的高速发展，在全世界范围内掀起了规模巨大的信息处理产业化浪潮。预计到 2020 年，人类生产的信息总量将以每 73 天翻一番的惊人速度发展。“我们正在被信息所淹没，却又饥饿于知识的客观现状”(John Naisbett，《大趋势》)[2]。目前的信息采集、传输、检索的传统理论与方法已成为制约人类利用知识的瓶颈，如图 1.1 所示。科学技术的创新发展将会把人类带入智能化的社会。知识创新将会成为社会进步的永恒推动力。知识发现、知识创造、知识组织、知识应用是知识创新链中不可缺少的关键环节，寻找新的知识创新链的有效方法与工具将成为认知科学、信息科学、计算机科学、知识管理科学领域共同努力的目标。