

知识元挖掘

温有奎 徐国华 著
赖伯年 温 浩

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 著
赖伯年 温 浩

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 and 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	1
1.1 知识标引的国内外研究现状	2
1.2 知识标引成为知识管理的瓶颈	3
1.2.1 文本标引技术	3
1.2.2 中文文本信息的特点	7
1.2.3 中文理解的困难	7
1.2.4 汉语文献的自动标引研究回顾	8
1.2.5 主题标引的进展	10
1.2.6 自动文摘的进展	11
1.2.7 知识标引成为知识管理的瓶颈	13
1.3 信息科学向内容处理深入	17
1.3.1 概念检索的特点	18
1.3.2 内容检索的特点	18
1.3.3 基于内容的检索方法	18
1.4 知识组织方法的创新	19
第 2 章 知识发现的新理论	21
2.1 知识的基本概念	22
2.2 知识的信息单元	25
2.3 元知识概念	27
2.4 知识的表示模式	29
2.4.1 知识原子	30
2.4.2 知识因子	31
2.4.3 知识因子的一元运算	32
2.4.4 知识项	34
2.4.5 知识表达式	36

2.4.6	知识表达式的 BNF 表示	37
2.5	知识映射的模式识别	38
2.5.1	映射的黑箱式与明晰式	39
2.5.2	狭义的模式识别	40
2.5.3	特征决定模式的表达形式	41
2.6	非数值特征模式分类	42
2.7	数据挖掘与知识发现	47
2.7.1	数据挖掘模式	48
2.7.2	数据挖掘与机器学习的区别	49
2.7.3	数据挖掘与数据库查询的不同	49
2.7.4	数据库中的知识发现的概念	49
2.7.5	文本中的知识挖掘	52
2.7.6	国内 KDD 的研究热点	53
2.8	粗糙集知识发现	54
2.8.1	粗糙集的概念	54
2.8.2	知识的分类观点	55
2.8.3	新型的隶属关系	56
2.8.4	ID3 与 RS 理论	58
2.8.5	知识粒子与知识粒度	59
2.9	非相关文献中的知识发现	61
2.10	归纳是知识发现的基本途径	62

第 3 章 文献标引与检索的进展 64

3.1	检索语言的概念逻辑基础	65
3.2	标引词的统计方法	66
3.2.1	文献标引	67
3.2.2	文献自动标引	68
3.2.3	Zipf 定律	69
3.2.4	Luhn 的自动抽词思想	70
3.2.5	叙词的聚类 and 结合统计	70
3.2.6	标引词向量空间模型	72

3.2.7	文献的词和词的连接矩阵	75
3.2.8	词和文献的结合矩阵	79
3.2.9	文献空间的质心 C	81
3.3	主题层次划分	83
3.3.1	有序聚类方式划分文本层次	83
3.3.2	采用语义网络表示主题的层次概念	84
3.4	标引词的特征提取与选择方法	86
3.4.1	特征提取与选择概念	86
3.4.2	基于熵的方差特征提取方法	88
3.5	文本知识发现的特点与进展	90
3.5.1	Web 上的数据挖掘	90
3.5.2	Web 知识发现的分类	91
3.5.3	Web 知识发现的方法	93
3.5.4	文本特征的处理功能	95
3.5.5	汉语文本结构的自动分析	97
3.6	新形势对情报检索语言的挑战	98
3.6.1	情报检索语言面临的挑战	98
3.6.2	利用文献数据库进行数据和事实检索	100
3.7	引入人工智能的方法	102
3.7.1	人工语言与自然语言结合	102
3.7.2	应用人工智能知识处理的方法	103
3.8	引入面向对象技术	106
第 4 章	知识创新与增值的知识链	109
4.1	知识创造	110
4.1.1	知识成为劳动者的生产要素	110
4.1.2	个人猜想是知识创造的起源	110
4.1.3	知识分为编码模式和人物化模式	111
4.2	知识转换	111
4.2.1	知识的复杂性: 隐性知识与显性知识	111
4.2.2	拉里·普鲁萨克的知识管理	112

4.2.3 知识链	114
4.2.4 知识链模型	115
4.2.5 隐性知识与显性知识转换的四个阶段	117
4.3 知识创新	118
4.3.1 组织是知识成为生产力的放大器	118
4.3.2 知识运用于具体的环境中才产生价值	119
4.3.3 知识是个体对信息的增值	119
4.4 数字图书馆的信息整流	120
4.5 知识增值	124
4.5.1 图书馆受到 Internet 的巨大冲击	124
4.5.2 数字图书馆的知识增值服务	125
第 5 章 知识管理的革命	128
5.1 传统情报学管理知识理论的困境	129
5.1.1 分类法和主题法组织的是文献而不是知识	129
5.1.2 数字化图书馆的不足	131
5.1.3 情报学应研究知识管理	131
5.1.4 情报学构建“知识体系”框架的任务	132
5.2 网格技术推进知识管理革命	132
5.2.1 网格时代的到来	133
5.2.2 网格提供巨大的计算技术空间	134
5.2.3 知识网格对知识管理的挑战	137
5.2.4 有关知识管理革命的几点设想	138
第 6 章 知识元链接理论	141
6.1 知识元标引是知识组织的新方向	142
6.2 知识元标引是知识管理的起点	143
6.3 知识元是构造知识系统的基元	143
6.4 知识元链接理论	144
6.4.1 知识元模块化	144
6.4.2 知识元结构定义	145

6.4.3 知识元链接框架	147
6.5 知识的网格结构	147
6.5.1 信息与知识元的导航	148
6.5.2 建立知识网格平台	149
第7章 信息与知识变换	151
7.1 信息与知识理论的早期贡献	152
7.2 信息科学的理论	154
7.3 信息与知识谱的变换性	156
7.3.1 知识谱	157
7.3.2 知识谱分析	159
第8章 知识元的认知理论	161
8.1 认知的探索	162
8.2 认知对文本知识挖掘的指导	164
8.3 知识元的抽取	166
8.4 实例分析	168
第9章 知识元标引	170
9.1 归类分析	171
9.2 数值型知识元结构	177
9.3 数值型知识元软件抽取试验	177
9.3.1 文本数值数据抽取算法	177
9.3.2 软件功能	177
9.3.3 修改原则	179
9.3.4 检索试验	182
9.4 知识元库构架	183
9.5 期刊论文知识元抽取方案	184
第10章 基于创新点的知识元挖掘	186
10.1 创新点是科学论文的灵魂	187

10.2	学术论文撰写的核心要素	188
10.3	文本“创新点”的特征	190
10.3.1	科技期刊的创新性质	190
10.3.2	科技期刊论文创新点的分布	190
10.3.3	科技期刊论文创新的特点	192
10.3.4	创新点的查新类型	193
10.4	知识元抽取试验	193
10.4.1	人工知识元抽取试验	193
10.4.2	创新点知识元抽取原理	194
10.5	实例分析	196
第 11 章	基于 XML 的知识元本体推理	200
11.1	知识元本体推理模型	201
11.2	XML 平台上的知识元表示	202
11.3	知识元实体	203
11.3.1	知识元的模板描述	204
11.3.2	数值型知识元实体举例	205
11.4	XML 语义网推理	205
11.4.1	语义网上五层次模型	205
11.4.2	语义信息层上的三元组	207
11.4.3	本体信息层上的知识推理	207
11.5	知识元本体推理	208
11.5.1	两个本体定义	209
11.5.2	一个数值知识元本体对话的实例	209
附录	数值型知识元抽取软件部分程序	211
参考文献		236
后记		240

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。