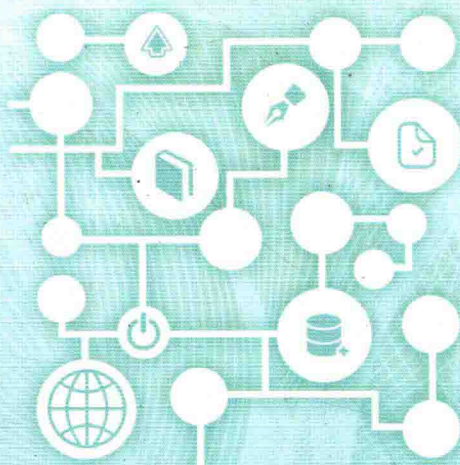


中国社会科学院创新工程学术出版资助项目

大数据时代 专题文献数据库系统设计



孔敬 著

中国社会科学出版社

大数据时代 专题文献数据库系统设计



图书在版编目(CIP)数据

大数据时代专题文献数据库系统设计 / 孔敬著. —北京: 中国社会科学出版社, 2017. 4

ISBN 978 - 7 - 5203 - 0365 - 1

I. ①大… II. ①孔… III. ①专题文献 - 文献数据库 - 数据库系统 - 系统设计
IV. ①G256②TP311. 135. 9

中国版本图书馆 CIP 数据核字(2017)第 080386 号

出 版 人 赵剑英
责任编辑 宫京蕾
责任校对 石春梅
责任印制 李寡寡

出 版 **中国社会科学出版社**
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 [http: //www. csspw. cn](http://www.csspw.cn)
发 行 部 010 - 84083685
门 市 部 010 - 84029450
经 销 新华书店及其他书店

印刷装订 北京市兴怀印刷厂
版 次 2017 年 4 月第 1 版
印 次 2017 年 4 月第 1 次印刷

开 本 710 × 1000 1/16
印 张 25. 25
插 页 2
字 数 414 千字
定 价 98. 00 元

凡购买中国社会科学出版社图书, 如有质量问题请与本社营销中心联系调换
电话: 010 - 84083683
版权所有 侵权必究

本书为国家社会科学基金研究成果

目 录

导论	(1)
第一节 大数据时代专题文献数据库系统的发展与新兴特征	(1)
一 文献的发展变迁	(1)
二 专题文献数据库的定义	(2)
三 大数据时代专题数据库系统的挑战与机遇	(2)
四 大数据时代专题文献数据库系统的内容特点	(5)
五 大数据时代文献数据库系统的功能特点	(5)
第二节 大数据时代专题文献数据库系统关键技术	(7)
一 XML 技术	(7)
二 多媒体信息集成技术	(7)
三 多媒体人机交互技术	(8)
四 软件智能体 (Agent) 技术	(8)
五 本体 (Ontology)	(8)
六 集成检索技术	(8)
七 云计算与 Web Service 等互联网技术	(9)
八 大数据技术	(9)
第三节 本书主要内容与组织结构	(10)
一 主要内容	(10)
二 组织结构	(12)

基础篇 关键技术

第一章 元数据标准与 XML 技术	(15)
第一节 元数据概念	(16)

第二节 信息组织描述的元数据技术体系	(17)
第三节 元数据标准的框架体系与设计原则	(18)
一 元数据标准的框架体系	(19)
二 元数据分类	(19)
三 专题文献数据库元数据标准的设计原则	(21)
第四节 元数据标准概述及主要标准简介	(21)
一 元数据标准概述	(21)
二 ISBD、MARC、DC、METS、RDF 等主要标准简介	(24)
第五节 XML 技术体系	(47)
一 XML 核心标准简介	(51)
二 XML 处理关联标准	(56)
三 XML 应用标准举例	(57)
四 XML 信息资源关键要素	(59)
第六节 小结	(60)
第二章 多媒体元数据标准: MPEG-7、MPEG-21	(61)
第一节 MPEG-7 标准基本概念	(61)
第二节 MPEG-7 标准内容简介	(62)
一 MPEG-7 第一至第十四部分简介	(64)
二 MPEG-7 标准描述内容总结	(67)
第三节 MPEG-21 标准概念与开发进展	(68)
第四节 MPEG-21 各部分内容简介	(72)
一 MPEG-21 第一部分	(72)
二 MPEG-21 第二部分	(72)
三 MPEG-21 第三部分	(74)
四 MPEG-21 第四部分	(74)
五 MPEG-21 第五部分	(75)
六 MPEG-21 第六部分	(76)
七 MPEG-21 第七部分	(77)
八 MPEG-21 第八部分	(77)
九 MPEG-21 第九部分	(78)
十 MPEG-21 第十部分	(78)

十一 MPEG-21 第十一部分	(78)
十二 MPEG-21 第十二部分	(79)
十三 MPEG-21 第十四至第二十二部分	(79)
十四 MPEG-21 标准框架小结	(80)
第五节 小结	(81)
第三章 多媒体检索技术	(83)
第一节 多媒体信息检索基本概念与特点	(83)
第二节 基于内容的多媒体检索	(84)
一 基于内容的图像检索技术	(85)
二 基于内容的音频检索技术	(86)
三 基于内容的视频检索技术	(87)
第三节 基于语义的图像检索	(88)
一 语义层次模型	(88)
二 语义表示方法	(88)
三 语义映射方法	(89)
第四节 基于语义的音频检索	(94)
一 音频内容分析的层次结构	(94)
二 语音识别 (低层语义分析) 技术	(96)
三 音频分类 (低层语义分析) 技术	(99)
四 高层语义 (音频场景) 分析技术	(100)
第五节 基于语义的视频检索	(100)
一 时域分割 (分镜头、场景检测) 与关键帧提取	(102)
二 空域分割 (区域分割、视频对象分割) 与语义特征提取 ...	(103)
三 概率统计方法	(104)
四 统计学习方法	(104)
五 基于规则推理的方法	(104)
六 结合特定领域特点的方法	(104)
第六节 小结	(105)
第四章 智能检索技术与模型	(106)
第一节 智能检索系统概述	(107)

一 智能检索系统发展历程	(107)
二 智能检索系统分类类型	(110)
第二节 智能信息检索基础模型	(112)
一 信息检索模型	(112)
二 智能信息检索的基本形式框架与概念模型	(113)
第三节 智能信息检索扩展模型	(116)
一 基于自然语言理解的智能检索模型	(116)
二 基于多 Agent 协作的智能检索模型	(116)
三 基于内容的多媒体智能检索模型	(118)
四 基于认知理论的智能检索模型	(119)
五 基于多媒体人机交互的智能检索模型	(120)
第四节 智能检索模型实例化: 检索算法与知识表示	(121)
第五节 智能检索系统热点技术	(124)
一 Agent 技术	(124)
二 本体技术	(127)
三 个性化技术	(127)
四 多媒体智能检索技术	(128)
第六节 结语	(128)
 第五章 知识本体与自动构建技术	 (130)
第一节 本体引论	(131)
一 何为本体	(131)
二 本体构成	(132)
三 本体构建原则	(134)
四 本体的分类	(136)
五 本体的表示语言	(136)
第二节 本体学习概念与分类	(137)
一 本体学习起源与概念	(137)
二 本体学习分类	(139)
第三节 本体学习系统功能结构	(141)
第四节 本体学习系统实例	(142)
第五节 本体学习关键技术与方法	(144)

一 概念抽取	(144)
二 概念关系抽取	(145)
第六节 本体与本体学习在信息系统中的应用	(149)
第七节 面向专题文献数据库系统的本体学习	(151)
一 面向专题文献数据库系统的本体学习环境模型	(151)
二 面向专题文献数据库系统的本体学习系统结构模型	(154)
第八节 小结	(158)
 第六章 个性化技术与用户建模	(160)
第一节 用户建模概念与分类	(161)
第二节 用户建模常用方法与用户知识表示	(162)
第三节 基于 Web 信息检索的用户建模	(168)
一 Web 内容抽取	(169)
二 Web 结构挖掘	(169)
三 Web 用法挖掘 (Web 日志挖掘)	(170)
四 Web 用户性质挖掘	(171)
第四节 用户模型在信息检索中的应用与评价指标	(172)
一 用户模型在信息检索中的应用类型	(172)
二 用户模型在信息检索中的应用案例比较	(174)
三 信息检索系统中用户模型的评价指标	(175)
第五节 小结	(176)
 第七章 信息检索模型算法与策略	(177)
第一节 信息检索概述	(178)
第二节 信息检索模型与算法	(179)
一 布尔模型	(180)
二 向量模型	(184)
三 概率模型	(188)
第三节 信息检索策略	(195)
一 相关反馈	(195)
二 聚类	(198)
三 同义词表与语义网络	(203)

四 其他检索策略	(205)
第四节 信息检索专项策略和算法	(206)
一 跨语言检索策略	(206)
二 并行与分布式检索	(207)
三 Web 检索	(208)
四 可视化检索	(215)
第五节 小结	(217)
 第八章 集成技术	(219)
第一节 集成技术概述	(220)
第二节 集成技术层次框架	(223)
第三节 集成检索技术	(226)
第四节 门户技术	(235)
一 门户及门户技术的概念	(235)
二 门户技术原理	(237)
三 门户技术体系结构	(239)
四 门户技术小结	(240)
第五节 Web 服务技术	(241)
一 Web 服务概念	(241)
二 Web 服务技术原理	(242)
三 Web 服务与门户技术之结合	(244)
第六节 综合多种集成方法的文献信息系统集成流程示例	(246)
第七节 小结	(248)
 第九章 云计算与大数据技术	(250)
第一节 云计算的发展与概念	(251)
一 云计算的发展	(251)
二 云计算的概念	(254)
三 云计算的特点	(255)
第二节 云计算的服务模式和部署模型	(257)
第三节 云计算的技术基础与系统结构	(258)
一 云计算的技术基础	(258)

二 云计算的系统结构	(259)
第四节 大数据发展与概念	(262)
第五节 大数据分析技术	(264)
第六节 小结	(265)

设计篇 分析设计与实例

第十章 体系结构与总体功能设计	(269)
第一节 需求分析	(269)
一 系统环境与定位	(269)
二 系统用户	(270)
三 系统目标	(270)
第二节 体系结构	(273)
第三节 数据流分析	(274)
一 数据流图基础	(275)
二 专题文献数据库系统顶层数据流	(276)
三 专题文献数据库系统 0 层数据流	(277)
四 子系统数据流图示例	(279)
第四节 总体功能设计	(279)
一 系统管理子系统	(282)
二 内容管理子系统	(284)
三 知识库管理子系统	(291)
四 信息服务子系统	(297)
五 信息分析子系统	(304)
第五节 小结	(305)
第十一章 基于实体—联系 (ER) 的关系型数据库设计	(306)
第一节 系统数据库的总体设计	(307)
一 实体—联系图	(307)
二 专题文献数据库系统的实体—联系图 (E-R 图)	(308)
三 专题文献数据库系统 E-R 图到数据库设计的转换	(309)
第二节 数据库配置库元数据设计	(312)

第三节 专题文献数据库元数据设计	(314)
第四节 用户数据库元数据设计	(318)
第五节 知识库元数据设计	(320)
第六节 小结	(322)
第十二章 基于 XML 的数据库设计	(323)
第一节 基于 XML 的数据库设计方法	(324)
一 XML 数据库及其相关概念	(324)
二 XML 数据库设计流程与方法	(325)
三 XML Schema 编辑工具	(329)
第二节 基于 XML 的家谱数据库设计	(330)
一 基于 UML 的概念设计	(330)
二 逻辑设计: XML 模式设计	(332)
三 数据实例: XML 实例文档	(339)
第三节 基于 XML 和 MPEG-7 的多媒体数据库设计	(345)
一 基于 MPEG-7 的中国民族风俗多媒体数据库元数据 设计	(345)
二 基于 MPEG-7 多媒体描述模式的一个 XML 实例文档	(348)
第四节 小结	(356)
第十三章 基于云计算和大数据的专题文献数据库系统设计	(358)
第一节 基于云计算的专题文献数据库系统体系结构	(358)
一 云计算的系统结构	(358)
二 云计算的服务体系结构	(359)
三 云计算的三层服务模式	(360)
四 基于云计算的专题文献数据库系统体系	(364)
第二节 大数据环境下专题文献数据库系统功能设计	(365)
一 大数据环境下专题文献数据库系统功能结构	(365)
二 大数据环境下专题文献数据库系统功能的变化与特点	(368)
第三节 小结	(369)
参考文献	(370)

导 论

随着当今信息的爆发性增长和信息技术的突破,特别是互联网的飞速发展,人类社会正从信息技术(IT)时代向数据科技(DT)时代飞速变革。从IT时代到DT时代,大数据正在对知识和信息的存储与传播产生巨大的变革和影响^①。大数据时代,作为人类传递知识、交流信息和保存文化的重要手段的文献,正经历着有史以来最为重大的变革,由最初较为单一的印刷型文献转变为以网络数据库、电子图书期刊等电子型文献为代表的多元化、海量级数据的文献形式。传统文献概念与形式的扩展,以及传播方式的多样化、网络化,对文献信息系统及其功能产生了巨大影响。这一切变化,促成了文献数据库系统的内涵更加丰富,外延更加扩展。专题文献数据库系统作为文献数据库系统中一个子类,对专业学科文献的收录有较高的广度和深度,它的建立可全面反映各专业学科的研究动态及现状,是深入开发和利用文献信息资源的最有效方式。本书首先对大数据环境下专题文献数据库系统的发展特点进行了剖析,对其系统开发的关键技术、新兴技术进行了归纳分析和阐述,提出了现代专题文献数据库系统的新框架和新理念,并根据新理念、新技术完成了现代专题文献数据库系统的体系结构、功能结构和数据结构设计。

第一节 大数据时代专题文献数据库系统的发展与新兴特征

一 文献的发展变迁

大数据时代,即以云计算、大数据为新兴特征的“互联网+”时代,

^① 阿里研究院:《互联网+:从IT到DT》,机械工业出版社2015年版,第4页。

文献在概念、载体形式和传播方式等方面都有了很大的扩展。文献的定义被扩展为“记录有知识的一切载体”。文献的载体形式由过去以纸型为主发展为如今的纸型和电子型并驾齐驱的局面。电子型文献有磁带、磁盘和光盘等多种载体形式,分为文本、图片、声频和视频等多种媒体格式,它以信息存储量大、检索方便的优势在网络信息时代有着后来居上的趋势。而在文献的传播手段方面,网络的出现更是打破了文献出版发行的严格界限,除传统的出版发行和新闻媒介等传播方式外,网络发布作为一种新型的信息传播方式以其特有的实时、动态和自由等特点迅速地发展起来。文献这一信息媒体正向多媒体、数字化、网络化方向发展。

二 专题文献数据库的定义

文献数据库随着计算机技术,特别是数据库技术的产生而产生,跟着计算机技术与通信技术的发展而发展。文献数据库系统的普及与应用使得传统图书馆、资料室的文献搜索、整理手段由过去单一的手工方式,转为自动化或半自动化方式;而随着计算机技术的发展,文献数据库系统所管理的内容,知识的载体——文献,也发生了前述的巨大变化。根据中国大百科全书的定义,数据库是为满足某一部门中多个用户多种应用的需求按照一定的数据模型在计算机系统中组织、存储和使用的互相联系的数据集合。专题文献数据库就是对某一专业学科文献广泛而深入地收录,并在计算机网络系统中组织、存储和使用。它的内容将全面反映该学科的研究动态及现状,是该专业人员或其他用户获取和利用该专业文献信息资源的最有效、最方便的途径。在云计算、大数据互联网时代,专题文献数据库就是建构在当前互联网和信息技术之上的以用户为中心,以数据分析处理为核心的专业数据库信息系统。

三 大数据时代专题数据库系统的挑战与机遇

在云计算、大数据互联网时代,信息环境已发生巨大变化。2008 年和 2011 年 Nature 和 Science 杂志先后出版了“大数据”和“数据处理”专刊^{①②},从互联网技术、网络经济学、超级计算、环境科学、生物医药

① Special issue: Big data [J]. Nature, 2008, 455: 1-136.

② Special issue: Dealing with data [J]. Science, 2011, 334: 692-729.

等多个方面介绍了海量数据带来的挑战,以及数据洪流所带来的机遇。大数据具备4个基础特征,用4个V来概括表示:数据体量大(Volume)、数据多样性(Variety)、价值密度低(Value)、处理速度快(Velocity)^①。大数据逐步从概念发展为一个新兴的研究方向,并对全人类的发展产生深远影响。这些发展变化所产生的问题,对专题文献数据库系统功能提出了以下新的挑战与机遇。

- 大数据时代,数据的网络化分布式存储和异构性,要求文献数据库系统提供集成检索信息服务。在大批量分布式异构数据库群环境下,文献数据检索不可能针对所有可用数据源而进行,对于检索的数据源必须有所选择。这要求现代专题文献数据库系统能管理异构的、动态扩展的和开放的数据源,自动识别异构数据的特征,根据用户交互行为,动态选择适当的数据源集合用于检索。这改变了传统文献数据库检索中数据源的稳定性和一致性。过去信息检索的对象通常是特定的数据库,用于检索的数据源相对稳定,数据结构一致。而今现代专题文献数据库系统需要在大规模分布式环境下实现异构信息的集成检索。

- 大数据时代,数据呈现几何级增长,数据体量大要求文献数据库系统提供个性化、专业化的更为精准的信息检索服务。海量多媒体信息的文献激增,导致“信息过载”、“信息迷向”等问题日趋严重。这要求现代专题文献数据库系统在用户交互中不断明确用户的信息需求,过滤大量冗余信息,筛选出满足用户个别需求的高质量、高精度信息,实现基于专业化、个性化的信息检索。

- 大数据时代,数据呈现多样性、复杂性,要求文献数据库系统能处理复杂的数据对象。在大数据环境下,文献信息数据种类繁多,数据量庞大,含有大量半结构化、非结构化的数据。特别是包含了文本、图像、语音等多种媒体格式的海量数据,又给现代专题文献数据库系统的开发提出了大规模多媒体复合数据对象的描述、存储、索引和检索等多项技术难题。

- 大数据时代,全文数字化、事实数据等电子文献大规模增加,

^① 李军:《大数据:从海量到精准》,清华大学出版社2014年版,第9页。

为文献数据库智能检索、知识抽取奠定了良好基础。印刷型文献全文数字化、电子图书期刊、网络百科工具书、网页信息资源等飞速发展,使基于文本的自动信息抽取更容易实现。这使得现代专题文献数据库系统由过去以收录文献题录为主转向以收录原文为主,为实现基于文本的智能检索奠定了良好基础。特别是百科事实数据类的各个专题文献数据库的建设,将为知识抽取提供良好的数据源基础。

- 大数据时代,数据的分布式庞大体量使文献数据库的数据规范生产难以控制。“互联网+”环境下,微博、微信等自媒体的发展更加迅速,基于云平台的多机构和多个体参与的虚拟文献数据库系统联合建设成为可能,数据信息制造的参与者无限增多,从处理局部数据变为处理全局数据,使信息制作的标准化和一致性难以实现,导致信息组织的异构性和无序性。这一方面给现代专题文献数据库提出了元数据描述标准化的问题,另一方面,自动信息加工、组织、知识发现和语义捕获等技术将在数据、信息和知识的规范化、规模化生产中发挥重要作用。

- 大数据时代,数据价值密度低,使得文献数据库系统抽取有价值的信息更为困难。这就需要从低价值的原始海量数据中进行深度挖掘和计算,总结出具备高价值的信息。面对原始的、海量的、复杂的多样性大数据,专题信息资源数据库如何从大量无序的低价值密度的数据中抽取专业性、知识性的高价值信息为用户提供高质量的专业信息服务变得更加困难和重要。

- 大数据时代,数据时效性强,处理速度快,使得文献数据库系统需要更快的并行处理能力。大数据区别于传统数据挖掘最显著的特征,就是数据处理的速度越快、越及时,其价值越大,发挥的效能越大。这就要求专题文献数据库系统能从各类型的原始数据中快速获取有价值的信息。

- Web 信息系统已成为文献数据库系统开发的主流模式,并且正由 B/S (Browser/Server) 系统架构走向 B/C (Browser/Cloud 浏览器/云) 系统架构。互联网的普及与发展使得各种数据库信息系统的开发转向了基于 Web 信息系统的开发模式,将数据库在网站上发布,面向更多用户提供方便快捷的信息检索服务,已成为数据库信息系统开发的主流模式。云计算服务模式将为整合来自分布式网络的文献信

息资源服务,建设虚拟化的大数据专题信息资源数据库提供实现基础。

四 大数据时代专题文献数据库系统的内容特点

传统专题文献数据库收录的数据是以书刊等印刷型文献为主。随着科学技术的发展,缩微资料、视听资料、机读资料、光盘资料、网站网页等新型文献不断涌现,它们在专业文献数据库中的比例将不断上升,并成为其主体。因此,在云计算、大数据互联网环境下,专题文献数据库的建设内容不仅包括本地的文献信息和电子信息资源,还应包括整合远程的外部可访问信息与资源。其内容建设有以下发展趋势:

一是海量多媒体信息。数据库收录内容从单一媒体向多媒体信息发展。不但收录文字信息,还将收录图片、音频、视频和超媒体文件等多种媒体格式。

二是网络化资源整合。在“互联网+”时代,文献数据库系统的资源建设由着重于本地文献资源收录转向虚拟文献资源整合发展。当前数字图书馆、虚拟图书馆等概念已被广泛接受,专业化的数字图书馆是没有围墙的图书馆,它以网络信息技术为基础,将世界范围内本专业的图书馆、个人、机构及商业公司连接起来,提供专业信息资源的联合存储与检索服务。由此,大数据时代,专题文献数据库建设是基于互联网的多种实体来源的联合资源整合,以实现学科资源共建共享。

三是数据信息知识化。从单纯文献采集向知识发现与知识重组发展。知识管理改变了传统的数据库建设理念。传统的文献数据库建设侧重于表达知识的记录型信息,即显性知识的管理,而忽视隐性知识的管理。现代专题文献数据库则更加重视其收录内容的知识化,即在显性知识的基础上进行知识发现与知识重组。知识重组有效地组织隐性知识,精练、提取、发现隐含在信息中的有用知识,并对其进行集合组织,设立一个包含信息使用者所需的相关知识信息的内容和地址的知识库,以实现知识的转换、创新、共享和交流。

五 大数据时代文献数据库系统的功能特点

- 基于 Web。文献数据库系统在经历了由单机版向网络版转变之后,又转向以 Web 信息系统方式为主,其系统架构从以 C/S (Cli-