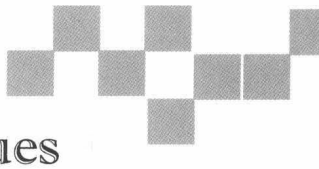
Web Data Management:
Concepts and Techniques

Web数据管理：概念与技术

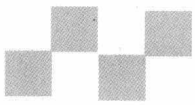
孟小峰 刘伟 姜芳艽 李玉坤 王仲远 编著

清华大学出版社

Web Data Management: Concepts and Techniques



Web 数据管理：概念与技术



孟小峰 刘伟 姜芳艽 李玉坤 王仲远 编著

清华大学出版社
北 京

内 容 简 介

本书介绍 Web 数据管理技术，包括：Web 数据抽取（数据型页面和文档型页面的抽取方法、基于视觉信息的抽取方法、包装器生成与维护及实体识别），Web 数据集成（查询接口集成、模式匹配、查询转换、数据库采样、数据库大小估计及集成系统实现），数据空间（数据空间的模型、索引、查询及系统实现），以及 Web 数据管理新技术（Web 信息可信性、移动 Web 搜索、移动应用集成、大规模知识库构建及社交媒体）。

本书适合作为 Web 数据管理的教科书，也可以作为相关领域研究人员和开发人员的参考书。

图书在版编目(CIP)数据

Web 数据管理：概念与技术/孟小峰等编著. --北京：清华大学出版社，2014
ISBN 978-7-302-37072-7

Ⅰ. ①W…　Ⅱ. ①孟…　Ⅲ. ①互联网络—数据库管理系统　Ⅳ. ①TP393.4

中国版本图书馆 CIP 数据核字(2014)第 146544 号

责任编辑：薛　慧
封面设计：常雪影
责任校对：刘玉霞
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：175mm×245mm　　**印　　张**：23　　**字　　数**：447 千字
版　　次：2014 年 9 月第 1 版　　**印　　次**：2014 年 9 月第1 次印刷
印　　数：1～2000
定　　价：69.00 元

产品编号：058233-01

序 一

Preface 1

自 Tim Berners-Lee 于 1989 年发明 World Wide Web 至今的短短 25 年内，Web 早已成为人类历史上数据量最大、用户最多、使用最频繁的数据源。它对现代社会的各个方面，包括经济、文化、教育、政治等，都产生了广泛及深刻的影响。同时，Web 数据具有大数据(big data)的所有基本特征，包括数据量大(volume)、类型多样化(variety)、实时性强(velocity)、真实性(veracity)及价值大(value)。可以说 Web 数据就是大数据。Web 数据分散在数以千计的相互独立的 Web 服务器上，需要先进的爬虫技术来收集整理。Web 可分为浅层 Web 和深层 Web，前者主要包括有固定 URL 的网页，而后者主要包括需要通过特殊 Web 搜索界面才能访问的数据。深层 Web 的数据量是浅层 Web 数据量的很多倍。如何有效地管理 Web 数据，使之能更好地为各类用户(包括个人、企业及各个机构)服务，这是一个被学术界和产业界共同关注的重大课题。

Web 数据管理是一个既重大又广泛的课题，包含很多具体的子课题，比如数据产生及更新，数据收集及搜索，用户界面设计，数据质量评估，用户数据隐私保护，数据抽取，实体识别，数据集成，语义 Web，数据及搜索个人化，移动搜索，等

等。大批科研及系统开发人员参与了这些课题的研究并取得了很多令人瞩目的巨大成就，使 Web 成为当今社会人们生活中不可缺少的重要部分，并且一定会使未来的 Web 更加精彩。然而，这些研究结果绝大部分都分散于大量的科研论文中，这不利于对这些结果的系统总结和有效利用。把这些结果系统合理地总结成书能大大提高这些结果的传播和利用。

中国人民大学信息学院的孟小峰教授和他的几个已经毕业的学生经过几年的努力，完成了《Web 数据管理：概念与技术》这部著作。该书内容非常丰富，涵盖了 Web 数据管理的很多领域，并被分成四大部分：Web 数据抽取（包括数据型页面和文档型页面的抽取方法、基于视觉信息的抽取方法、包装器生成与维护及实体识别），Web 数据集成（包括查询接口集成、模式匹配、查询转换、数据库采样、数据库大小估计及集成系统实现），数据空间（包括数据空间的模型、索引、查询及系统实现），Web 数据管理新技术（包括 Web 信息可信性、移动 Web 搜索、移动应用集成、大规模知识库构建及社交媒体）。另外，本书还介绍了 XML 基础知识，全书加了绪论，前三部分加了概论，这增强了此书的系统化和结构合理化。本书内容较新，包括了近年来的科研新成果，与科研前沿有很好的接轨。书中基本概念定义清晰，算法叙述准确具体、实施性强，还含有大量实例和插图，大大增强了实用性和可读性。

孟小峰教授是国际上数据库领域的著名学者，其研究涉及 Web 数据管理、XML 数据管理、移动数据管理、云数据管理、闪存数据库、隐私保护等领域。他在这些领域发表论文 150 余篇及出版专著 3 部。孟小峰教授从 2000 年就开始率领他的团队从事 Web 数据管理方面的研究和应用开发。在过去的十多年里，这个团队取得了非常好的成绩，在高质量国际学术会议及杂志上发表了多篇相关论文，并开发出多个广为 Web 用户使用的应用系统（例如文献集成系统 ScholarSpace 和工作通 JobTong）。本书的很多内容就直接来源于这个团队的第一手研究成果。

本书内容适合作为 Web 数据管理领域的教科书，也可以作为该领域研究人员和开发人员的参考书。据我所知，目前在 Web 数据管理领域还没有一本像本书一样内容丰富全面新颖的教科书或参考书。我非常高兴将这本很有价值的书推荐给广大读者。

Weiyi Meng（孟卫一）
2014 年 3 月 31 日
（美国宾汉姆顿纽约州立大学（SUNY Binghamton）教授，
WAIM 指导委员会主席）

序 二

Preface 2

飞速发展的万维网迅速成为全球信息传递与共享的日益重要和极具潜力的资源，这些资源的爆炸式增长使 Web 成为一个巨大的信息库。但是与它的增长速度相比，Web 信息资源的开发和利用还远远不够。如何有效地管理 Web 的海量信息，以满足用户不断增长的信息需求，这已经成为学术界和产业界共同关注并致力解决的问题。

信息资源的基础是数据，高质量的信息需求需要先进的 Web 数据管理技术的支持。与传统的数据库管理技术相比，Web 数据管理的显著特点是：Web 的数据主要呈现海量、异构、大规模分布、噪声普遍、动态增长的特征，这都给它的管理带来很多难题，其中就包括如何有效地对 Web 数据进行抽取与集成，如何有效管理面临“数据爆炸”的个人数据与信息管理等这样一些重要的研究问题。

本书作者多年来在 Web 数据管理方面开展了系统的研究，尤其在 Web 数据抽取、Web 数据集成和数据空间管理三个方面进行了深入的探讨，其目的在于为高效 Web 数据管理提供理论基础和新的解决思路。全书基于作者的研究工作，全面系统地介绍了 Web 数据管理技术，内容涵盖了 Web 数据抽取（包括抽取方法、包装器生成与维护、实体识别等）；

Web 数据集成(包括接口集成、模式匹配、查询转换、数据库采样、数据库大小估计、系统实现等)；数据空间(包括数据空间模型、索引、查询、系统实现等)；以及新技术进展(包括 Web 信息可信性、移动 Web 搜索、移动应用集成、大规模知识库构建、社交媒体等)。全书内容新颖，结构清晰，深入浅出，既反映了 Web 数据管理的最新研究成果，又具有良好的实用性。在当下大数据热潮中，大数据所体现的 4V 特征与 Web 数据的特征几多相似。因此，本书所介绍的 Web 数据管理技术与方法，可以为解决大数据管理问题提供借鉴和参考。

本书作者孟小峰教授及他所带领的团队长期从事数据库系统和理论的研究工作，学术思想活跃，对研究的前沿动向敏感。自 2000 年起，他们就将研究目标定位在创新数据管理的方向上，为此他们创建了网络与移动数据管理实验室(Web and Mobile Data Management Lab)，致力于 Web 数据管理、XML 数据管理和移动数据管理等，取得了多方面的研究成果。他们先后出版《移动数据管理：概念与技术》(清华大学出版社，2009)、《XML 数据管理：概念与技术》(清华大学出版社，2009)。这本《Web 数据管理：概念与技术》是当初设想的创新数据管理三部曲的最后一部，是作者多年来对 Web 数据管理技术研究的结晶，也是他们坚持长线研究，将理论研究和系统实现有机结合的良好体现。

相信本书对广大的科研工作者和研究生具有重要的学术参考价值，希望本书对推动中国计算机学科的发展及数据管理技术的进步也能发挥应有的作用。

2014 年 3 月

(清华大学教授，中国计算机学会数据库专委会主任)

前言

Foreword

Web数据指能够通过Web访问到的所有数据。基于Web数据访问形式不同，Web又分为浅层网络(Surface Web)和深层网络(Deep Web)。如何有效地管理Web上的大量信息，以满足用户不断增长的高质量的信息需求，成为学术界和产业界共同关注并致力于解决的问题。Web数据所呈现的特征为：海量异构；分布广泛；动态增长；先有数据，后有模式。这使得Web数据无论从数量上还是复杂程度上，都与传统数据库技术所处理的数据显著不同，需要有更先进的技术来管理Web数据。

Web数据管理的主要目的是解决Web上丰富信息资源有效利用的问题，从而大大提高Web应用的开发效率。Web数据管理是指针对特定的主题领域，利用数据抽取和数据集成技术，自动识别Web中与所给主题相关的实体及实体之间的关联，构造面向主题的结构化关联数据，并对这些数据进行有效处理(包括数据质量、动态演化、隐私保护等)，从而为用户提供高质量的信息服务。

传统的数据库技术为传统应用系统的开发提供了有利的支撑，缩短了应用开发周期，降低了系统维护成本。Web数据管理技术与传统的数据库技术一脉相承，其大大降低了Web应用系统开发的难度，同样缩短了应用开发周

期，降低了系统维护的代价。诸如学术集成系统、网络舆情系统、价格比对系统、工作查找系统等应用，利用 Web 数据管理系统可以方便快捷地加以开发，并实现日常的自动增量维护。

当下大数据浪潮一浪高过一浪，大数据所体现的数据量大(volume)、数据多样性(variety)、实时性强(velocity) 、价值大(value)以及真实性(veracity)的特征与 Web 数据的特征几多相似。因此本书所提出的 Web 数据管理技术与方法，本质上提供了将多源异构非结构化数据加以结构化管理的途径，进而为解决大数据管理问题提供了有益的尝试。

本书与同类书目的比较

本书作者自 2000 年即对 Web 数据管理开始连续多年的系统性研究。本书基于作者多年在 Web 数据管理方面的研究积累，全面系统地介绍了 Web 数据管理相关技术。内容涵盖了 Web 数据抽取(包括抽取方法、包装器生成与维护、实体识别等)；Web 数据集成(包括接口集成、模式匹配、查询转换、数据库采样、数据库大小估计、系统实现等)；数据空间(包括数据空间模型、索引、查询、系统实现等)；以及新技术进展(包括 Web 信息可信性、移动 Web 搜索、移动应用集成、大规模知识库构建、社交媒体等)。

Principle of Data Integration (Anhai Doan, Alon Halevy, Zachary Ives,2012,MK)是与本书内容最相近的一本书。该书是有关数据集成技术的集大成之作，其中译本亦由我们实验室翻译，近期将由机械工业出版社(华章)出版。数据集成技术的研究历时 20 多年，大致可分为两个阶段：起初主要关注企业内部异构数据库范畴下的数据集成，随着 Web 的出现，开始关注动态海量 Web 数据源范畴下的数据集成。两个阶段的侧重点有所不同，技术和方法也有所差异。Alon Halevy 的研究背景涉及了这两个阶段，因此该书试图将这两个阶段的研究成果中的共性技术同时呈现给读者。该书以教科书逻辑整理有关内容，强调知识的基础性和理论性。其第一部分主要介绍数据集成的基本知识，主体基本来自数据库集成的内容，如查询的表示、数据源的描述(GAV、LAV)、模式匹配、查询处理、集成方法等；穿插补充了 Web 数据集成的内容，如包装器、数据匹配(实体识别)等。第二部分主要介绍数据集成的扩展知识，主要包括 XML、语义 Web、不确定性、数据溯源等。第三部分介绍各种新的集成技术，包括 Web 数据集成、基于关键字的按需集成、对等集成、协同集成等。

这本名为《Web 数据管理》的书直接以 Web 数据为研究对象，系统地介绍了 Web 数据管理的关键技术，早期的数据库集成技术并未涉及，那是 *Principle of Data Integration* 一书的优势。但论及 Web 范畴下的数据集成，

本书的体系则更为系统、具体、丰富、完整，并有系统实现的内容，本书相关的章节有第9章包装器、第8章数据匹配(对应本书的第一部分数据抽取)、第15章Web数据集成(对应本书第二部分数据集成)。后者显然具有侧重基础知识上的优势，而本书具有侧重Web数据管理的系统化优势。可以说两者相互补充，相得益彰。当然后者在知识体系上的厚度和广度是本书所不能比拟的。

Advanced Metasearch Engine Technology (Weiyi Meng, Clement Yu, Morgan & Claypool Publishers, December 2010)和*Deep Web Query Interface Understanding and Integration* (Eduard Dragut, Weiyi Meng, Clement Yu, Morgan & Claypool Publishers, 2012)是有关Web数据管理的两本专著，其作者是这个领域的资深学者，特别在元搜索引擎方面颇有研究。

Web Data Mining: Exploring Hyperlinks, Contents and Usage Data (Bing Liu, Second Edition, July 2011 First Edition, Dec 2006, Springer)一书是近期与Web数据相关的比较优秀的一部专著，书中的内容主要侧重数据挖掘的基本知识和Web数据挖掘。书中部分内容与本书相关，如Web爬取、Web搜索、结构化数据抽取、信息集成等。总之该书还是定位在数据库挖掘领域。

在作者的书架上还有几本早期的与Web数据相关的书籍，但与本书讨论的内容均关系不大。比如*Data on the Web: From Relations to Semistructured Data and XML* (Serge Abiteboul, Peter Buneman, Dan Suciu, 2000, MK)，该书当年因其书名吸引了很多人，但从副标题可以看出其讨论的主要问题是半结构化数据模型和XML。该书作者Serge Abiteboul与*Principle of Data Integration*的作者Alon Halevy其实都是属于斯坦福流派，20世纪90年代斯坦福的学者做了大量异构数据库集成技术的理论和系统研究工作，该书是其研究工作在半结构化数据上的扩展，其实与本书所讨论的内容基本没有交叉。*Web Data Management* (Sourav S. Bhowmick, Sanjay K. Madria, Wee Keong Ng, 2003, Springer)，书名虽与本书相同，但内容差异很大。作者从数据库的角度对Web信息的有效管理提出了基于数据仓库的方法，试图用对象模型对Web数据建模。此类工作在2000年前后提出了很多，如W3QL, WebSQL, WebOQL等，都试图用数据库建模的思想组织Web数据并提供类似SQL的查询语言。其实这里的本质问题是：把Web数据**看成**结构化来处理是一回事，如何把Web数据**变成可处理的**结构化数据甚至知识库(如知识图谱)是另外一回事。后来表明此类工作因没有能够针对Web数据的特殊处理要求，渐渐淡出了人们的视野。相反本书介绍的Web数据抽取、实体识

别、接口集成等研究成为这方面的主流技术，更能解决 Web 数据处理的实际需要，日渐为人们重视。

本书的内容和组织结构

本书分为四个部分，共 25 章。前两章为基础知识。

第 1 章是本书总述，主要讨论 Web 数据的特点、Web 数据管理基本概念，以及涉及的关键技术。

第 2 章介绍 XML 基础知识。XML 在 Web 数据管理中发挥了重要作用，首先 Web 应用中大量使用 XML 表示形式，其次 Web 数据抽取和集成目前普遍采用 XML 作为中间形式来完成。本章介绍了 XML 数据模型和查询语言(XQuery)，主要涉及本书用到的 XML 基本概念和基础知识。

第一部分为 Web 数据抽取技术。主要讨论从 Web 页面中抽取数据所涉及的关键技术，具体包括第 3 章至第 8 章。

第 3 章从 Web 页面的分类开始，介绍了 Web 数据抽取的定义，基本方法和评价标准。

第 4 章重点介绍数据型页面抽取技术，包括针对单记录数据型页面的抽取技术和针对多记录数据型页面的抽取技术。

第 5 章重点对文档型页面抽取技术进行讨论，主要介绍文档型页面的结构特征和主要的抽取技术。

第 6 章主要介绍包装器的生成与维护方法，讨论如何使数据抽取程序快速地适应 Web 页面结构和内容变化这一问题。

第 7 章针对 Web 页面的视觉特征，讨论基于视觉的 Web 数据抽取技术。基于视觉的抽取方法可以独立于任何特定的页面编程语言，因此具有更广泛的适用性。

第 8 章重点讨论 Web 页面中的实体识别技术，主要包括 Web 实体属性的分类、实体识别技术框架、属性匹配计算方法。

第二部分为 Web 数据集成方法和技术。主要包括数据集成框架、查询接口集成、不确定模式匹配、查询转换、Web 数据库采样、Web 数据库大小估计、系统实现等内容。具体包括第 9 章至第 15 章。

第 9 章对 Web 数据集成框架进行阐述，基于该框架对 Web 数据集成所需要解决的关键问题进行分析。

第 10 章重点讨论 Web 数据集成中的查询接口集成技术，包括接口集成框架、领域匹配方法以及查询转换算法。

第 11 章关注的是用户查询转换过程中模式匹配的不确定性问题，重点对

数值型数据的相似度计算方法、不确定模式匹配方法进行介绍。

第 12 章重点讨论集成查询接口和各数据源自身查询接口之间的查询转换问题，重点介绍基于动态规则的转换方法和基于谓词的转换方法。

第 13 章对数据库采样技术进行分析，内容涉及记录选择、查询的生成、采样过程的终止、偏差的修复几个方面。

第 14 章讨论 Web 数据库大小估计的方法，重点介绍基于词频的估计方法。

第 15 章介绍 Web 数据集成系统实现技术，主要涉及 Web 数据管理系统体系架构，以及静态集成、动态集成方法，并通过两个案例分别对这两种集成方式进行介绍。

第三部分对数据空间的基本概念和相关技术进行介绍，包括数据模型、索引技术、查询技术、系统实现几个方面。具体包含第 16 章至第 20 章。

第 16 章对数据空间相关概念和关键技术进行总述。

第 17 章对数据空间模型进行阐述，重点介绍数据空间模型、任务空间与核心数据空间。

第 18 章主要介绍数据空间索引技术，包括倒排索引、面向属性的倒排索引、面向动态演化的索引技术。

第 19 章重点讨论数据空间查询技术，重点介绍基于同义词的查询方法、基于任务的查询技术和基于核心数据空间的查询技术。

第 20 章重点介绍数据空间系统实现技术。以数据空间原型系统 OrientSpace 为例对数据空间构建、存储等实现技术进行介绍。

第四部分介绍 Web 数据管理的新技术进展和应用，包括 Web 信息可信性、移动 Web 搜索、移动应用集成、大规模知识库构建、社交媒体等几个方面，具体包括第 21 章至第 25 章。

第 21 章根据不同的网络应用场景各自的特点，重点对 Web 数据可信性问题及相关技术进行讨论。

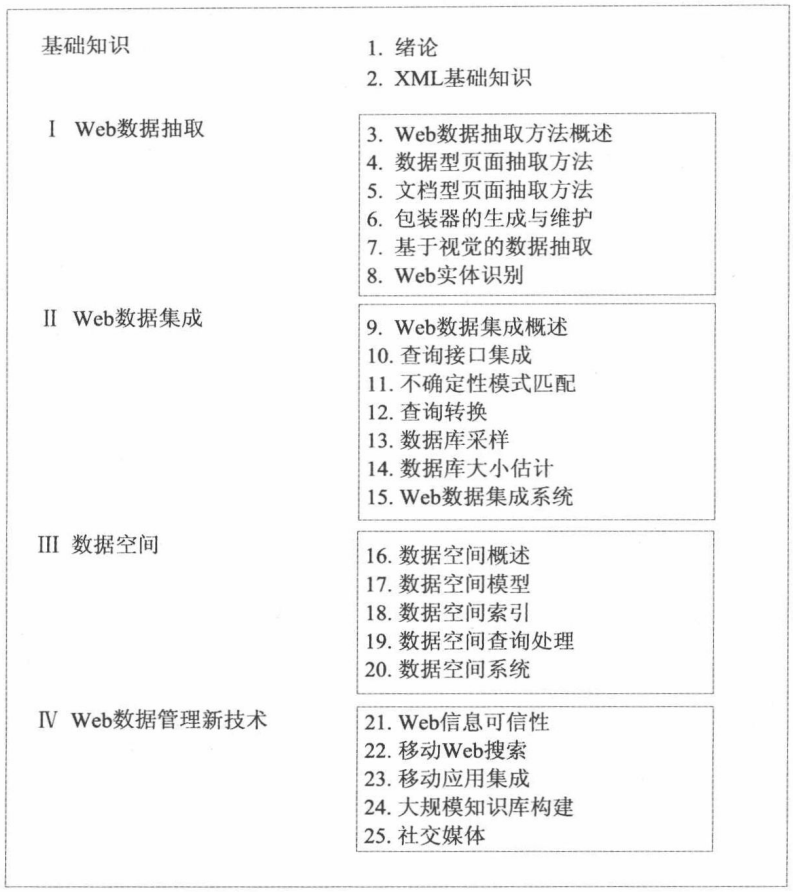
第 22 章重点对移动 Web 及其特性进行分析，讨论移动 Web 搜索与互联网搜索的差异，以及移动 Web 搜索流程及关键技术。

第 23 章主要对移动应用集成相关技术进行分析，具体包括移动应用数据抽取、移动应用匹配、移动应用搜索与推荐。

第 24 章对大规模知识库构建相关技术进行讨论，包括关联数据的基本概念和相关技术，以及知识图谱相关概念。

第 25 章介绍社交媒体相关概念和技术，具体包括社交媒体特点、社交媒体数据带来的挑战，以及社交媒体中的短文本分析、事件发现与处理框架。

全书各章均附有参考文献。



章节结构图

本书的对象和使用

本书深入详细地介绍了 Web 数据管理相关概念和技术，内容涵盖 Web 数据抽取（包括抽取方法、包装器生成与维护、实体识别等）；Web 数据集成（包括接口集成、模式匹配、查询转换、数据库采样、数据库大小估计、系统实现等）；数据空间（包括数据空间模型、索引、查询、系统实现等），内容系统全面；本书同时强调内容的先进性，将 Web 数据管理的最新成果涵盖进来，包括 Web 信息可信性、移动 Web 搜索、移动应用集成、大规模知识库构建、社交媒体等等。本书在阐述上力求简洁明了，对概念方面的内容，辅以一些图表加以说明，对技术方面的内容，则给出精炼的实现算法。本书在编写中考虑到教学需要，给出了大量实例。每章的内容安排大致相当，并附有相关文献及注释。书后附有总文献和词汇索引。

本书主要面向各类研究人员和开发人员，既可以作为相关领域的教科书，

也可以作为技术参考书。基于本书可开设的课程有如下建议：

Web 数据管理技术：面向高年级本科选修课程，内容包括第Ⅰ，Ⅱ；

Web 数据管理研究：面向研究生的课程，包括Ⅰ～Ⅳ；

Web 数据管理专题 1：与本科数据库管理系统课程结合，补充内容Ⅱ；

Web 数据管理专题 2：与本科数据库管理系统课程结合，补充内容Ⅲ；

大数据专题：与相关研究生课程结合，补充内容Ⅰ～Ⅳ。

致谢

作者从 2000 年起即开始从事 Web 数据管理的研究工作，2007 年开发计算机中文文献系统 C－DBLP，在此基础上开发多领域文献集成系统 ScholarSpace，包括 SearchScholar，EasyScholar 和 SocialScholar。之后还开发了工作通 JobTong、个人信息管理系统 OrientSpace 等。本书即是作者在多年来形成的研究成果的基础上，经过总结和整理而成。

首先感谢国家自然基金委多年长期的资助，和国家 863 计划一贯的支持。在连续的十多年间，具体有如下的项目资助：

- 2003—2005 国家自然科学基金项目“Web 数据抽取和集成技术研究”，编号：60273018
- 2002—2004 国家 863 项目“基于 Web Service 的 Web 数据集成技术”，编号：2002AA11304
- 2007—2009 国家高技术研究发展计划(863 计划)项目“海量数据空间模型、查询与索引技术研究”，编号：2007AA01Z155
- 2011—2013 国家自然基金面上项目“Web 信息可信性研究”，编号：61070055
- 2014—2017，国家自然基金面上项目“面向移动用户的 Web 集成技术研究”，编号：61379050
- 2012—2014，国家 863 项目“海量非结构化数据管理系统结构、测试与标准”，编号：2012AA011001
- 2011—2016，中国人民大学研究基金重大基础研究项目“社会计算若干关键问题研究”，编号：11XNL010
- 2013—2015 高等学校博士学科点专项科研基金优先发展领域课题“云计算环境下的在线聚集技术研究”，课题号：20130004130001
- 2012—2015，国家自然科学基金面上项目“基于图的个人数据空间模型与查询方法研究”，项目号：61170027

特别感谢两位在此领域颇有建树的学者百忙中拨冗为本书作序：美国宾汉姆顿纽约州立大学(SUNY Binghamton)孟卫一教授和清华大学周立柱教

授。孟卫一教授在 Web 数据管理方面是国际上的知名学者，作为元搜索引擎的开拓者在国际上享有盛誉，自 2000 年以来多次回国开办 Web 数据管理方面的讲习班，与本研究团队交流密切，对我们的工作多有指导和帮助，他目前是 Web 时代信息管理国际会议(WAIM)指导委员会主席。周立柱教授领导的研究团队在大规模 Web 数据管理和知识提取方面有出色的研究工作，对本实验室的研究工作常年给予指导和帮助，他目前担任中国计算机学会数据库专业委员会主任委员。他们对本书作了整体概括和推介，在此深表谢意。

本书的形成凝聚了中国人民大学网络与移动数据管理实验室(http://idke.ruc.edu.cn)集体智慧。特别感谢实验室的博士研究生和硕士研究生，先后有若干届的学生参与到本项目的研究中来，他们是博士生刘伟、姜芳艽、李玉坤、张金增、马如霞、马友忠、李勇，以及硕士生谷明哲、王海燕、胡东东、李宇、李忺、林灿、凌妍妍、王仲远、艾静、赵婧、胡享梅、贾琳琳、张相於、寇玉波、陈威、邓云、童薇、王淼、赵可君等。刘伟(第一部分)、姜芳艽(第二部分)、李玉坤(第三部分)、王仲远(第 15、21 章)等直接参与写作并在资料收集和文献整理方面做了大量工作。

本书涉及面广，内容丰富，参考文献众多。值得指出的是，在全书的撰写和课题的研究中，尽管投入了大量精力、付出了艰苦努力，但受知识水平所限，书中不当之处在所难免，诚恳希望读者批评指正并不吝赐教。如果有任何建议或意见，可发电子邮件至 xfmeng.ruc@gmail.com。

孟小峰

2014 年 3 月于北京

目录

Contents

第一部分 Web 数据抽取