



中国科学院研究生院教材

Textbooks of Graduate University of Chinese Academy of Sciences

Web信息系统导论

李广建 编著

Introduction to Web Information
System



高等教育出版社
Higher Education Press



中国科学院研究生院教材

Textbooks of Graduate University of Chinese Academy of Sciences

Web信息系统导论

李广建 编著

Introduction to Web Information System



高等教育出版社
Higher Education Press

内容简介

本书主要包括 Web 信息系统基本理论与方法、XML 基础、Web 服务、语义 Web、Web 搜索、Web 挖掘、信息抽取、Web 协同、Web 整合和门户系统。其重点是对不同应用领域中 Web 信息系统的体系结构、关键技术进行分析和介绍。

本书可作为信息系统等课程的研究生教学用书,也可供高年级本科生在学习相关课程时参考使用。

图书在版编目(CIP)数据

Web 信息系统导论/李广建编著. —北京:高等教育出版社, 2008.9

ISBN 978-7-04-024715-2

I. W… II. 李… III. 计算机网络 IV. TP393

中国版本图书馆 CIP 数据核字(2008)第 120858 号

策划编辑	孙惠丽	责任编辑	张海波	封面设计	杨立新
责任绘图	尹莉	版式设计	张岚	责任校对	俞声佳
责任印制	韩刚				

出版发行	高等教育出版社	购书热线	010-58581118
社 址	北京市西城区德外大街 4 号	免费咨询	800-810-0598
邮政编码	100120	网 址	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
经 销	蓝色畅想图书发行有限公司	网上订购	http://www.landaco.com
印 刷	北京中科印刷有限公司		http://www.landaco.com.cn
		畅想教育	http://www.widedu.com
开 本	787×1092 1/16	版 次	2008 年 9 月第 1 版
印 张	22	印 次	2008 年 9 月第 1 次印刷
字 数	430 000	定 价	39.10 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

版权所有 侵权必究

物料号 24715-00

中国科学院研究生院教材编审委员会

主 任：白春礼

顾 问：余翔林

副主任：马石庄（常务） 刘志鹏 韩兴国 苏 刚

委 员（按姓氏笔画排列）：

石耀霖 李家春 李伯聪 李 佩 刘嘉麒

朱健强 张文芝 张增顺 吴 向 汪尔康

汪寿阳 杨 乐 徐至展 阎保平 黄荣辉

黄 钧 彭家贵 裴 钢 谭铁牛

技术学科编审组

主 编：徐至展

副主编：涂国防

编 委：王 珏 王家骥 冯玉琳 冯登国 刘立人

阴和俊 张良益 张雨东 邹谋炎 柳欣欣

唐志敏 顾逸东 顾国彪 阎保平 夏善红

黄伟光 谭铁牛 潘辛平

总 序

在中国科学院研究生院和高等教育出版社的共同努力下，凝聚着中国科学院新老科学家、研究生导师们多年心血和汗水的中国科学院研究生院教材面世了。这套教材的出版，将为丰富我院研究生教育资源，提高研究生教育质量，培养更多高素质的科技人才起到积极的推动作用。

作为科技国家队，中国科学院肩负着面向国家战略需求，面向世界科学前沿，为国家作出基础性、战略性和前瞻性的重大科技创新贡献和培养高级科技人才的使命。中国科学院研究生教育是我国高等教育的重要组成部分，在新的历史时期，中国科学院研究生教育不仅要为我院知识创新工程提供人力资源保障，还担负着落实科教兴国战略和人才强国战略，为创新型国家建设培养一大批高素质人才的重要使命。

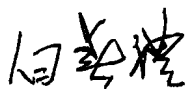
集成中国科学院的教学资源、科技资源和智力资源，中国科学院研究生院坚持教育与科研紧密结合的“两段式”培养模式，在突出科学教育和创新能力培养的同时，重视全面素质教育，倡导文理交融、理工结合，培养的研究生具有宽厚扎实的基础知识、敏锐的科学探索意识、活跃的创新思维和唯实、求真、协力、创新的良好素质。

研究生教材建设是研究生教育中重要的基础性工作。由一批活跃在科学前沿，同时又具有丰富教学经验的科学家编写的中

中国科学院研究生院教材，适合在校研究生学习使用，也可作为高校教师和专业研究人员的参考书。这套研究生教材内容力求科学性、系统性、基础性和前沿性的统一，使学习者不仅能获得比较系统的科学基础知识，也能体会蕴于其中的科学精神、科学思想、科学方法，为进入科学研究的学术殿堂奠定良好的基础；不但是体现教学内容和教学方法的知识载体、开展教学的基本条件和手段，也是深化教学改革、提高教育质量、促进科学教育与人文教育结合的重要保证。

“十年树木，百年树人”。我相信，经过若干年的努力，中国科学院研究生院一定能建设起多学科、多类型、多品种、多层次配套的研究生教材体系，为我国研究生教育百花园增添一支新的奇葩，为我国高级科技人才的培养作出新的贡献。

中国科学院常务副院长
中国科学院研究生院院长
中国科学院 院士



二〇〇六年二月二十八日

前 言

Web 信息系统是随着 Web 技术的发展及广泛应用而发展起来的一种信息系统,是 Web 环境下的一种新型信息系统,其设计理念、实现方式、应用范围与传统的信息系统有很多差异,正逐渐成为一个重要的研究领域。

本书较全面地概述了 Web 信息系统领域所涉及的有关知识和问题,全书共分为 10 章。第 1 章为 Web 信息系统基础,旨在介绍 Web 信息系统的概念、内容、特征和技术体系。第 2 章是 XML 的基础知识,XML 在 Web 信息系统的研究和实践中占有特别重要的地位,是 Web 信息系统的技术基础。第 3 章和第 4 章分别介绍 Web 服务和语义 Web,Web 服务和语义 Web 不仅仅是两种具体的技术,更重要的是它们代表了 Web 信息系统和 Web 技术的发展趋势。第 5~10 章介绍 Web 信息系统开发中常用的技术,包括搜索技术、挖掘技术、信息抽取技术、协同技术、整合技术和门户技术。

我先后在中国科学院研究生院和北京师范大学管理学院为硕士生开设了“Web 信息系统”课程,本教材是在课程讲稿的基础上编写而成的,承蒙中国科学院研究生院的资助,使得本书得以正式出版。

我要特别感谢中国科学院研究生院教材编审委员会学科编审组的 3 位匿名评审专家,他(她)们认真地审读了本教材的主要内容,并提出了十分中肯的修改意见。正是根据他(她)们的意见,我对讲稿重新进行了组织和修订,增补了有关算法的内容。我至今不知 3 位评审专家的姓名,无法当面对他(她)们表示感谢,只能在此表达我的敬意。

在开设“Web 信息系统”这门课程之初,张智雄、李春旺、黄永文、郭家义、江洪等参加了课程体系的讨论。在申请出版资助的过程中,李春旺协助我做了大量的工作,并执笔撰写了 Web 搜索一章的样稿。

在撰写本书的过程中,牟秋江、徐方、湛怡萍、郭理文在已有讲稿的基础上,进一步搜集了近年来出现的一些新材料,更新了原有讲稿中的旧内容。特别是牟秋江,协助我重新撰写了本书第 6~8 章的内容。

在开设“Web 信息系统”课程的过程中,参阅和引用了大量的文献,不断对讲稿进行补充,由于持续时间较长,加之篇幅有限,本书无法将参考和引用的文献一一列出,在此,对本书具名和未具名的参考文献的作者表示衷心的感谢。

Web 信息系统的内容十分广泛,由于作者水平所限,疏漏之处在所难免,敬请专家、学者和读者批评指正。

李广建

2008 年 5 月于北大燕北园

目 录

第 1 章 Web 信息系统基础	1
1.1 Web 信息系统的基本概念	1
1.2 Web 信息系统的研究活动与研究热点	3
1.2.1 专业会议	4
1.2.2 专业期刊	5
1.2.3 著作	6
1.2.4 教育	6
1.3 Web 信息系统的类型	7
1.3.1 按网络范围划分	7
1.3.2 按信息传递方向划分	9
1.3.3 按功能复杂程度划分	9
1.4 Web 信息系统的技术模式	11
1.4.1 以服务器为中心的技术模式	11
1.4.2 以客户端为中心的技术模式	14
1.4.3 分布式的技术模式	17
1.5 Web 信息系统的设计方法	21
1.5.1 关系管理方法	21
1.5.2 面向对象的超媒体设计方法	24
1.5.3 基于统一建模语言的 Web 工程方法	27
1.5.4 Web 建模语言方法	30
本章主要参考文献	34
第 2 章 XML	35
2.1 XML 概述	35
2.1.1 XML 的产生背景	35
2.1.2 XML 的结构	36
2.2 XML 的基本语法	38
2.2.1 符号	38
2.2.2 文档序言	39
2.2.3 文档主体	40
2.2.4 实体	41

2.3 DTD	44
2.3.1 定义 XML 元素	44
2.3.2 定义元素属性	47
2.3.3 引用 DTD	51
2.4 XML Schema	52
2.4.1 XML Schema 的结构	52
2.4.2 元素定义	53
2.4.3 属性定义	54
2.4.4 类型定义	54
2.4.5 对元素或属性进行限制	58
2.5 XPath 与 XSLT	62
2.5.1 XPath	62
2.5.2 XSLT	67
本章主要参考文献	76
第3章 Web 服务	77
3.1 Web 服务概述	77
3.1.1 Web 服务的概念和特点	77
3.1.2 Web 服务的应用	79
3.1.3 Web 服务的体系结构与实现原理	80
3.1.4 Web 服务的协议体系	82
3.2 SOAP 协议	83
3.2.1 SOAP 的消息结构	83
3.2.2 SOAP 的主要元素	84
3.2.3 SOAP 编码	88
3.2.4 在 HTTP 中使用 SOAP	89
3.3 WSDL 协议	90
3.3.1 WSDL 文档的结构	91
3.3.2 WSDL 的主要元素	91
3.4 UDDI 协议	99
3.4.1 UDDI 注册中心的结构	100
3.4.2 UDDI 的数据结构	100
3.4.3 UDDI API	106
本章主要参考文献	108
第4章 语义 Web	109
4.1 语义 Web 概述	109
4.1.1 语义 Web 的基本概念和内容	109

4.1.2 语义 Web 的体系结构	110
4.2 资源描述框架 RDF	112
4.2.1 资源描述模型	112
4.2.2 RDF/XML 语法概要	115
4.3 资源描述框架模式 RDFS	118
4.3.1 类的描述	118
4.3.2 属性的描述	121
4.4 本体	124
4.4.1 本体概述	125
4.4.2 本体的构建	129
4.4.3 本体构建工具	133
4.5 语义 Web 描述语言	135
4.5.1 RDF	135
4.5.2 OIL	136
4.5.3 DAML	137
4.5.4 OWL	139
本章主要参考文献	140
第 5 章 Web 搜索	142
5.1 Web 搜索概述	142
5.1.1 Web 搜索的概念	142
5.1.2 Web 搜索的发展简史	142
5.1.3 Web 搜索的评价标准	143
5.1.4 Web 搜索的类型	144
5.2 通用 Web 搜索	147
5.2.1 通用 Web 搜索系统的体系结构	147
5.2.2 通用 Web 搜索的相关技术	150
5.3 专题 Web 搜索	159
5.3.1 专题 Web 搜索系统的体系结构	160
5.3.2 专题 Web 搜索的相关技术	161
本章主要参考文献	167
第 6 章 Web 挖掘	169
6.1 Web 挖掘概述	169
6.1.1 Web 挖掘及其特点	169
6.1.2 Web 挖掘的基本流程	170
6.1.3 Web 挖掘的分类	171
6.2 Web 文本挖掘及其关键技术	174

6.2.1 Web 文本挖掘的基本流程	174
6.2.2 Web 文本表示模型	176
6.2.3 Web 文本特征提取技术	179
6.2.4 Web 文本分类技术	182
6.2.5 Web 文本聚类技术	185
6.3 Web 结构挖掘	187
6.3.1 Web 结构挖掘的基本原理	187
6.3.2 基于 PageRank 改进算法的挖掘技术	189
6.3.3 基于 HITS 改进算法的挖掘技术	191
6.3.4 PageRank 和 HITS 相结合的算法	193
6.4 Web 使用挖掘及其关键技术	194
6.4.1 Web 使用挖掘的基本流程	194
6.4.2 会话识别与事务识别技术	196
6.4.3 关联规则发现	198
6.4.4 序列模式挖掘技术	202
本章主要参考文献	204
第7章 信息抽取	206
7.1 信息抽取概述	206
7.1.1 信息抽取及其主要学术活动	206
7.1.2 信息抽取的基本流程	208
7.1.3 信息抽取的分类	212
7.1.4 信息抽取的评价指标	215
7.2 自由文本中的信息抽取技术	215
7.2.1 有指导的机器学习法	216
7.2.2 无指导的机器学习法	221
7.3 基于封装器的网页信息抽取	223
7.3.1 网页抽取的基本原理	223
7.3.2 半自动方式生成封装器	226
7.3.3 全自动方式生成封装器	233
本章主要参考文献	237
第8章 Web 协同	238
8.1 Web 协同概述	238
8.1.1 Web 协同概念及其特点	238
8.1.2 Web 协同系统的类型	240
8.1.3 Web 协同系统体系结构	242
8.1.4 Web 协同系统的实现方法	243

8.2 协同工作的基本模型	244
8.2.1 协作模型	244
8.2.2 群体感知模型	248
8.3 Web 协同中的关键技术	253
8.3.1 群组通信技术	253
8.3.2 应用共享技术	256
8.3.3 同步机制	258
8.3.4 并发控制	259
8.3.5 多用户协作界面	261
8.4 群件	262
8.4.1 消息系统	262
8.4.2 计算机会议系统	263
8.4.3 协同编著及设计系统	265
8.4.4 群组决策与会议室系统	266
8.4.5 workflow 管理系统	267
本章主要参考文献	269
第9章 Web 整合	270
9.1 概述	270
9.1.1 Web 整合的概念	270
9.1.2 Web 整合的类型	272
9.2 Web 整合系统的体系结构及其关键技术	277
9.2.1 数据仓库架构及其关键技术	277
9.2.2 Mediator/Wrapper 架构及其关键技术	284
9.3 Web 整合的标准规范	293
9.3.1 元数据收割协议 OAI-PMH	293
9.3.2 元数据链接协议 OpenURL	298
本章主要参考文献	306
第10章 门户系统	307
10.1 概述	307
10.1.1 门户及其特点	307
10.1.2 门户技术的发展	308
10.1.3 门户的类型	311
10.2 门户系统的结构和功能	314
10.2.1 门户系统的体系结构	314
10.2.2 门户系统的主要功能	316
10.3 门户关键技术	319

10.3.1 Portlet 技术	319
10.3.2 Portlet 远程调用技术	327
本章主要参考文献	335

第 1 章 Web 信息系统基础

Web 信息系统 (Web Information System, WIS 或 WebIS) 是随着 Web 技术的发展及广泛应用而发展起来的一种信息系统, 是 Web 环境下的一种新型信息系统, 其设计理念、实现方式、应用范围与传统的信息系统有很大的不同, 正逐渐成为一个重要的研究领域。

1.1 Web 信息系统的基本概念

据不完全统计, 目前对 Web 信息系统的定义有数十种之多, 这里简要介绍其中一些典型的 Web 信息系统定义。

A. R. Dennis 于 1998 年提出: Web 信息系统是系统, 不是网页。Web 信息系统首先是信息系统, 其次才是 Web 系统, Web 页被设计成能像翻阅杂志那样进行浏览。与 Web 页不同的是, Web 信息系统允许用户进行事务处理操作, 这种操作本身要比 Web 更加复杂。

A. R. Dennis 在这个定义中将 Web 信息系统和普通的网页区别开来, 强调 Web 信息系统的信息系统特性。在 Web 信息系统中, 用户操作的对象是具有原子性、持久性、隔离性和一致性的事务, 这远比 Web 上的浏览功能复杂得多。

T. Isakowitz 等人于 1998 年也指出: Web 信息系统与一组 Web 页有明显的区别, 即 Web 信息系统支持并通常与其他非 Web 信息系统 (如数据库、事务处理系统等) 紧密结合; 同时 Web 信息系统也区别于传统的信息系统, Web 信息系统需要采用新的设计和开发方法, 能够为更广泛的受众所使用, 通常是大众推动的结果, 这些不同导致了管理和技术方面的变化。

H. W. Gellersen 等人于 1999 年提出: Web 信息系统是依靠 Web 来保证其正确执行的应用软件。

H. W. Gellersen 等人的定义虽然简短, 但内容要比前两个定义更广泛。这个定义没有刻意强调 Web 信息系统的用户交互性, 因此在理论上还包括网络上的应用程序, 例如搜索引擎中的网络爬虫等, 用户并不能与网络爬虫程序进行直接交互, 但可以使用网络爬虫所获取的信息资源。H. W. Gellersen 等人将这类应用系统也看作是 Web 信息系统。

C. Gnaho 于 2001 年提出: Web 信息系统是可以通过 Web 访问复杂数据和进

行交互服务的信息系统。

在这个定义中,C. Gnaho 强调了 Web 信息系统的运行环境是 Web 网络,处理的对象是复杂数据(注意,不是一般性的数据或简单的数据),其功能是为用户提供交互服务。

A. Scharl 等人于 2001 年对 Web 信息系统提出了以下观点:Web 信息系统是指信息系统的一个子类,这类系统一般能使得分布在多个位置的大量随机用户通过自我服务联机进行信息检索和执行例行任务。

A. Scharl 等人在这个定义中说明了 Web 信息系统所面对的用户群体的特征,强调用户使用信息系统的普遍性以及系统服务的用户自主性,另外还说明了 Web 信息系统的分布性和实时(联机)性。

J. Holck 于 2002 年提出:Web 信息系统或 WIS 是一种由计算机支持的信息系统,它使用了 WWW 技术,并且允许大量用户通过浏览器来访问。

在这个定义中,J. Holck 强调了 Web 信息系统的技术基础是计算机和 Web 技术,并将 Web 信息系统的客户端限定为浏览器。

M. Villanova-Oliver 等人于 2002 年提出:与传统的信息系统一样,Web 信息系统可以看做是用于进行信息搜集、组织、存储、管理与传播的技术手段的集合。与传统信息系统的不同之处在于,Web 信息系统的这些活动都是在 Web 基础架构上完成的,因此,Web 信息系统中涉及的信息数量巨大、来源广泛、格式多样,并且经过不同应用软件的处理。此外,Web 信息系统所管理的服务更为复杂,例如,在电子商务中管理用户购物车服务等。

M. Villanova-Oliver 的这个定义强调了 Web 信息系统的 3 个方面,一是 Web 信息系统以 Web 架构为基础,二是 Web 信息系统的数据具有复杂性,三是 Web 信息系统的服务要比传统信息系统更为复杂。

K. Schewe 等人于 2005 年对 Web 信息系统提出的定义是:Web 信息系统是一种有后台数据库支持的信息系统,这种系统在 Web 上实现并且分布在 Web 上,用户可以通过 Web 浏览器加以访问,可以通过包含导航结构的网页和系统外的站点来获取系统信息。此外,系统还应能实现对底层数据库的信息检索和更新操作。

这是对 Web 信息系统比较综合的定义,它不仅指出了 Web 信息系统的技术基础和分布式结构的特点,强调了导航结构的重要性,同时还指出了 Web 信息系统应包括除导航以外的其他交互操作,但仍然把 Web 信息系统的操作界面限于 Web 浏览器。

还有研究者认为,当前软件系统的特征与数年前的系统有着极大的区别,存储和多媒体信息检索的需求以及 Internet 的存在使得复杂的分布式软件系统应运而生,这些系统关注不同的用户角色。此外,此类系统还具有更为丰富的功能,例如,导航与用户界面就是两个极为重要的因素,具有这些特点的系统称为 Web 信息系统。

在这个定义中,研究者除了强调分布式和页面导航的特点外,还强调了多媒体技术、用户角色和系统界面的重要性。

之后,又有研究者提出:鉴于巨大的存储需要与复杂的功能需求,Web 信息系统是一类基于 Internet 的分布式信息系统,用户界面、信息恢复、导航和多媒体是这种信息系统的基本要素。

在这个定义中,研究者将 Web 信息系统定义进行了总结,将界面、导航以及多媒体技术等归纳为这种新型信息系统的基本要素。

分析以上各种 Web 信息系统的定义不难看出,人们对 Web 信息系统的认识在不断深化。早期,人们更加强调 Web 信息系统与 Web 页以及传统信息系统的区别,认为不能将它们等视之。随着研究的深入,人们进一步从更全面的角度(如技术架构、系统结构、数据特征、用户界面等)去界定 Web 信息系统,从而对 Web 信息系统的认识越来越清晰。

对以上定义加以归纳和总结,可以对 Web 信息系统有如下认识:

Web 信息系统是以 Web 技术为基础构建起来的,在 Web 上运行的,对各类 Web 信息进行搜集、组织、存储、管理与传递,通过适当界面形式为用户提供服务的分布式信息系统。

这个认识包括以下几层含义:

(1) 从系统功能上看,Web 信息系统本身是信息系统的一种特定类型,它和一般信息系统一样,承担着对信息进行搜集、组织、存储、管理与传递的任务。

(2) 从技术实现上看,Web 信息系统的基础是 Web 技术,因此,Web 信息系统的实现要广泛地应用网络传输、网络数据库、超文本、网络多媒体等诸多技术。

(3) 从系统结构上看,Web 信息系统是 Web 上的分布式信息系统,可以提供与位置无关、不受时间限制的信息服务。

(4) 从用户界面来看,这里并没有将 Web 信息系统的界面限定于 Web 浏览器,尽管在实践中大多数 Web 信息系统都使用 Web 浏览器,但这并不是 Web 信息系统的必备特征,Web 信息系统的界面可以使用 Web 浏览器,也可以不使用 Web 浏览器。

(5) 从处理对象看,Web 信息系统的处理对象是 Web 信息,Web 信息的特点是数量大、来源广泛、结构复杂,这相应地对 Web 信息系统提出了更高的要求。

1.2 Web 信息系统的研究活动与研究热点

Web 信息系统是近年来一个较为热门的研究领域,目前国内外对该领域的研究都非常重视。例如,发表在计算机科学、信息管理等领域中的相关研究论文;对于 Web 信息系统研究领域的形成和发展较有影响的学术活动,包括定期举办的专

门以 Web 信息系统为研究对象的学术会议及出版发行 Web 信息系统的专业杂志和著作;在大学中开设的 Web 信息系统相关课程。这些活动集中反映了当前 Web 信息系统的发展情况,体现了当前 Web 信息系统的研究热点。

1.2.1 专业会议

Web 信息系统工程国际会议(International Conference on Web Information Systems Engineering)是目前全球最重要的 Web 信息系统专业会议,由 Web 信息系统工程协会(Web Information Systems Engineering Society, WISE Society)主办,每年举行一次。Web 信息系统工程协会是一个非营利性的组织,其目标是促进 Web 信息系统以及相关领域的学术交流,举办 Web 信息系统工程国际会议是该协会的重要工作之一。

Web 信息系统工程国际会议始于 2000 年,称为“第一届 Web 信息系统工程国际会议”,以后每年举办一次,会议顺序编号,会后出版会议论文集,截至 2006 年,已成功地举办了 7 届。表 1-1 列出了历届 WISE 会议的基本情况。

表 1-1 历届 Web 信息系统 WISE 会议的基本情况

WISE 会议	举办时间	举办地点	主要议题
第一届	2000 年 6 月 19 ~ 21 日	中国香港	Web 数据库、XML 应用、工作流、Web 查询与搜索、元数据、数据挖掘、移动计算
第二届	2001 年 12 月 3 ~ 6 日	京都	Ontology、电子商务、Web 搜索与检索、XML 应用、多媒体、Web 地理信息系统
第三届	2002 年 12 月 12 ~ 14 日	新加坡	Web 信息检索、Web 挖掘、XML 应用、元数据与 Ontology、信息抽取
第四届	2003 年 12 月 10 ~ 12 日	罗马	Web 查询与检索、Web 计量、XML 应用、基于 Web 的合作、语义 Web、Web 服务(Web Services)、数据与知识整合
第五届	2004 年 11 月 22 ~ 24 日	布里斯班	Web 信息系统建模与设计、信息抽取、链接分析与文档聚类、XML 应用、Web 搜索与个性化、工作流、深层 Web、Ontology、多媒体
第六届	2005 年 11 月 20 ~ 22 日	纽约	Web 挖掘、Web 信息检索、元数据、Ontology 与语义 Web、XML 应用、Web 服务、Web 协同、信息抽取
第七届	2006 年 10 月 23 ~ 26 日	武汉	Web 搜索与检索、Web 服务、工作流、Web 挖掘、XML 应用、语义 Web 与整合、Web 文档分析、多媒体