山东财政学院学术文丛 2009

面向Web的个性化语义信息检索技术

林培光 康海燕 著

SHANDONG CAIZHENG XUEYUAN XUESHU WENCONG

中国财政经济出版社

山东财政学院学术文丛

面向 Web 的个性化语义信息检索技术

林培光　康海燕　著

中国财政经济出版社

图书在版编目（CIP）数据

面向 Web 的个性化语义信息检索技术/林培光，康海燕著. —北京：中国财政经济出版社，2009.9

（山东财政学院学术文丛）

ISBN 978-7-5095-1675-1

Ⅰ.面… Ⅱ.①林…②康… Ⅲ.情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字（2009）第 109856 号

责任编辑：刘五书　　责任校对：李　丽

封面设计：邹海东　　版式设计：兰　波

中国财政经济出版社 出版

URL：http：//www.cfeph.cn

E-mail：cfeph@cfeph.cn

社址：北京市海淀区阜成路甲 28 号　邮政编码：100142

发行处电话：88190406　财经书店电话：64033436

北京财经印刷厂印刷　　各地新华书店经销

880×1230 毫米　32 开　9.25 印张　212 000 字

2009 年 9 月第 1 版　2009 年 9 月北京第 1 次印刷

定价：22.00 元

ISBN 978-7-5095-1675-1/TP·0018

山东财政学院学术文丛
编　委　会

总序

Zong Xu

《山东财政学院学术文丛》（以下简称《学术文丛》）是山东财政学院为集中展示本校学人学术研究成果而编辑出版的系列丛书。《学术文丛》的出版，对于落实山东财政学院“学科立校”与“人才强校”的发展战略、繁荣学术研究事业、加强同学术界的交流等具有十分重要的意义。

始建于1986年、由邓小平同志亲笔题写校名的山东财政学院，是在改革开放的春风中，由财政部和山东省人民政府共同创办、实行以地方管理为主的普通高等财经院校。学校目前拥有14个二级学院，43个本科专业，22个硕士学位授权点和MBA、MPA两个专业学位授权点。学校学科门类齐全，已形成以经济学、管理学为主，文、法、理、工等六大学科门类相结合的学科结构。其中财政学、会计学、金融学、企业管理、国际贸易学、管理科学与工程为山东省重点学科，财政学与企业管理为省级重点强化建设学科。在山东省政府确定的“泰山学者”特聘教授设岗学科中，财政学、金融学位列其中。依托于以上优势学科与特色学科，一批批学科带头人与学术骨干脱颖而出，学校也由此成为省内著名、在全国有一定影响的经济学与管理学研究人才高地。

在人才培养方面，山东财政学院广纳全国英才，以“培养基础扎实、知识面宽、业务工作能力强、综合素质高、具有国际视野

的应用型人才”为目标，不断提高教学质量，为社会培养和输送了两万多名优秀毕业生。其中每年都有相当数量的毕业生或被中央机关、国家部委和著名公司录用，或考取名牌高校研究生继续深造。目前山东财政学院的毕业生已得到社会各界的普遍认可，为学校赢得了较高的社会声誉。“学在山（东）财（政学院）”已成为莘莘学子努力追求的人生目标。

在短短的二十多年间，山东财政学院之所以取得如此辉煌的成就，这既是“求是崇真、博学笃行”的校训与“高标准、严要求、好校风、有特色”办学指导思想得以落实的必然体现，也是学校实施“学科立校、人才强校、开放兴校、依法治校”战略、积极推进以提高教学水平与科研水平为核心的综合改革的必然结果。近年来，为把学校建设成为在国内外有一定影响的多科性高水平特色大学，学校在启动本科教学质量工程的同时，启动了研究生学位点建设工程，加快推动学科建设与学术研究工作上层次、上水平。围绕这一目标，学校不断优化学术环境，提倡学术民主，创新学术激励机制，科研工作取得重要突破，并涌现出了一批高水平的学术研究成果。为向外界推介这些学术研究成果，进一步繁荣学术研究事业，校领导审时度势，决定出版《山东财政学院学术文丛》。《学术文丛》就是在这一背景下编辑出版的。

为使《学术文丛》反映、代表我校学科建设与学术研究的最高水平，《学术文丛》在书稿的遴选过程中，严格学术标准，规范评审程序，采用了校外专家审稿与校学术委员会评审确定的机制，最终确定入选《学术文丛》的书稿。经此严格的筛选，这一部部书稿以其较高的研究水准与学术价值得以入选。应当指出的是，这些书稿不仅集中反映了我校学术研究的最新成果，而且展现了我校学人的时代风采。在入选者中，既有名气较大的知名学者，也有砥柱中流的学术中坚，还有崭露头角的学界新秀。在他们中间，或术业有专攻，或名气有大小，或起点有高低，但有一点是共通的：那

就是他们在各自的领域内，瞄准学术前沿，不畏路途艰辛，治学严谨，用力勤苦，最终取得了丰硕的研究成果。可以说，这一部部书稿凝聚了作者多年来潜心学术研究的心血汗水，展现了财院学人勇攀学术高峰的时代风貌。

我们相信，《学术文丛》的出版不仅在加快学术队伍建设、推动学科建设方面起到重要作用，而且在加强同学术界的交流、扩大学校的学术影响力等方面也将产生深远的影响。为此，今后我们还将每年遴选10部左右的书稿出版，推动我校学术研究事业繁荣兴盛，薪传不息。

《学术文丛》的顺利出版，得到中国财政经济出版社的大力支持，张立宪副总编、林治滨先生为丛书的出版付出了诸多辛劳，在此我们表示衷心的感谢！在《学术文丛》的出版过程中，山东财政学院校领导高度重视，校科研处精心组织，各位作者积极配合，谨此我们一并表示诚挚的谢意！

《山东财政学院学术文丛》编委会

2009年3月12日

前言

Qian Yan

1990年以前，没有任何人能够检索互联网上的信息。应该说，所有的网络信息检索工具都是从1990年的Alan Emtage等人发明的Archie开始的，虽然它当时只可以实现简单意义上的FTP文件检索。随着World Wide Web的出现和发展，基于网页的信息检索工具出现并迅速发展起来。1995年基于网络信息检索工具本身的检索工具元搜索引擎由美国华盛顿大学的Eric Selberg等发明。伴随着网络技术的发展，网络信息检索工具也取得了十足的发展，已成为人们获取信息的重要手段。

本书对信息检索的研究内容和研究目的、信息检索的研究现状、传统检索模型、检索性能评价方法、语义网等基础内容进行了简单介绍；在此基础上，重点介绍了个性化信息检索和面向语义网信息检索的相关理论、算法和技术框架。

全书共分三篇：第一篇对信息检索领域的研究基础进行了概括介绍，包括信息检索的研究目的和意义、信息检索的研究现状、传统信息检索模型、常见的性能评价方法以及语义网、云理论等知识；第二篇主要介绍了个性化信息检索方面的研究，包括个性化信息检索的框架及其理论基础、基于云的泛概念检索模型、检索评价策略和用户相关性判定方法、用户建模技术及兴趣挖掘方法、查询相似度计算和查询后处理和个性化信息检索系统架构等问题；第三

篇主要介绍了面向语义网的信息检索方面的研究，包括基于本体的语义信息检索模型、基于描述逻辑的知识库检索方法、描述逻辑 ALC 的不确定性扩展和基于语义网的分布式信息检索系统等方面的研究内容；最后对本书进行了总结。

本书主要收录了作者攻读博士学位期间所完成的学术论文和近几年所接触的信息检索领域的相关研究成果。由于本书研究属于信息检索领域研究的热点和前沿问题，研究难度较大，其中许多问题仍在研究和探索阶段，加之作者水平有限，虽经几次修改，但难免有许多不足和缺陷，敬请读者、专家、同行朋友惠予指正。

著 者

2009 年 1 月

MuLu 目 录

第 1 篇 信息检索的研究背景和相关基础

第 2 篇　面向 Web 的个性化信息检索

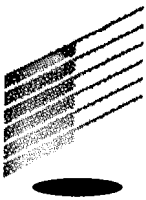
第 3 篇 面向语义网的信息检索

第 1 篇

信息检索的研究背景和相关基础

本篇主要对信息检索的研究目的和意义、信息检索的研究现状、传统信息检索模型、常见的性能评价方法以及语义网、云理论等知识进行简要阐述。

本篇共六章。第 1 章阐述了当前信息检索中存在的关键问题，并由此引出解决现有问题而需要研究解决的问题：不确定性推理、个性化信息检索和基于语义的检索；第 2 章主要对信息检索的发展历程、国内外信息检索的研究现状和语义网信息检索的现状进行了综述；第 3 章综述了当前主流的信息检索模型，并就它们的现状和基本实现原理指出了各自的优缺点；第 4 章给出了常见的检索性能评价方法，并重点介绍了基于召回率和准确率的各种评价方法的基本原理；第 5 章介绍了语义网相关技术框架和语义网的基础——描述逻辑；第 6 章介绍了一种不确定性推理理论——云理论，包括其产生背景、基本概念、定性和定量的表示的双重特性，并对正向和逆向云发生器、云变换和云模型应用现状等进行了介绍。



第1章

信息检索的研究目的和意义

随着计算机的普及和互联网（WWW）的迅猛发展，大量信息（80%左右）以电子文档的形式出现在人们面前。因此，要想从如此海量的信息中找到满足需要的信息无疑是一项极富挑战性的工作。显然，仅靠人工搜索和提取，其操作过程将非常繁琐，并且速度和效率也极低，信息质量也得不到保证。常规的信息检索是基于关键字进行的，但这种检索方式中，信息的获得往往是被动的，必须有用户的参与，获取的信息比较繁杂，不能体现个人兴趣。所以，不论个人还是企业，都面临一个严峻的问题：如何快速、准确地从 Web 页上获取所需信息？解决人们获取知识的困难，迫切需要一些自动化的工具帮助人们迅速找到真正需要的信息，这就是信息检索的任务。信息检索技术是互联网最基础、最核心的技术[①]，一个搜索引擎就是一个信息检索系统，搜索引擎越来越成为互联网的一个操作系统，它掌控着人们从信息海洋中获取有用信息的路径。第一代搜索引擎是以雅虎为代表的人工目录网站导航方式；第二代搜索引擎是以 Google 为代表的关键词搜索方式成为搜索的主

① 徐宝文，张卫丰．搜索引擎与信息获取技术［M］．北京：清华大学出版社，2003.