信息电子类专业
研究生教学用书

Web搜索

Web Search

郭 军

高等教育出版社

信息电子类专业
研究生教学用书

Web搜索

Web Search

郭 军

高等教育出版社

内 容 简 介

本书内容包括导论、文本检索、图像检索、音频检索、信息过滤、信息推荐以及发展前沿。对 Web 搜索的基本概念进行定义，阐述其科学价值和研究状况，根据 Web 搜索所涵盖的检索、过滤以及推荐技术，论述其中的核心问题、基本概念和基本方法，并介绍 Web 搜索若干新的研究方向。

本书的最大特点是将 Web 上的信息检索、过滤和推荐等技术定义为 Web 搜索，使其具有比较宽泛的内涵。将 Web 检索、过滤和推荐统一在一个体系中，既符合这三项技术发展的现状和趋势，又便于读者进行系统的学习和研究。另外，本书紧跟近年来的最新研究进展，具有显著的先进性和独特性。

本书可以作为信息、通信、计算机类研究生或高年级本科生的教材和教学参考书，也可作为专业技术人员的阅读和培训资料。

图书在版编目(CIP)数据

Web 搜索/郭军. —北京：高等教育出版社，2009.8
ISBN 978-7-04-027817-0

Ⅰ.W… Ⅱ.郭… Ⅲ.主页制作-程序设计
Ⅳ.TP393.092

中国版本图书馆 CIP 数据核字(2009)第 121818 号

策划编辑 许怀锷 **责任编辑** 唐笑慧 **封面设计** 李卫青
责任绘图 尹 莉 **版式设计** 范晓红 **责任校对** 殷 然
责任印制 宋克学

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	**咨询电话**	400-810-0598
邮政编码	100120	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京新华印刷厂	**畅想教育**	http://www.widedu.com
开 本	787×960 1/16	**版 次**	2009 年 8 月第 1 版
印 张	19.25	**印 次**	2009 年 8 月第 1 次印刷
字 数	350 000	**定 价**	31.00 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。

物料号 27817-00

前　言

当今时代，如何从源源不断、无边无际的海量 Web 数据中搜索信息已经成为一个对社会的政治、经济、文化、安全等具有全方位影响的重大课题。在这一背景下，以信息检索、过滤和推荐为主要内容的 Web 搜索引起了全球学术界、产业界以及各国政府的极大关注。商用搜索引擎巨头迅速崛起，强有力地带动了社会经济的发展。相关的学术研究异常活跃，为自然科学和社会科学的多个领域的研究注入了活力。

Web 搜索是一种高度智能化的信息处理技术。在目前已经形成的理论和技术体系中，融合了模式识别、自然语言处理、机器学习、数据挖掘等多个学科的成果，综合性和交叉性十分突出。此外，海量信息处理、Web 网页自动获取及分析、网页索引、网页链接分析、社会网络挖掘等内容更是具有独特性和新颖性。这门技术也因此走入了大学的课堂，并迅速受到了广大学生的青睐。目前，国内外 IT 背景较强的大学都至少在研究生层次上开设了相关的课程。

相对于这种旺盛的教学需求，Web 搜索的教材建设明显滞后，特别是中文教材非常稀缺。即使是外文教材也在系统性和前沿性等方面落后于技术的发展现状。因此，编写出版紧跟最新技术进展的 Web 搜索的大学教材有十分紧迫的需求。

作者长期从事模式识别和网络技术领域的研究和教学工作，近年来对 Web 搜索产生了浓厚的兴趣，带领一支十多人的教师团队指导上百名研究生对该领域进行了多方面的深入研究。通过研究工作的不断积累，对 Web 搜索的技术体系和主要内涵有了比较深刻的认识和理解，感到值得将其梳理和总结为一部主要面向研究生教学的教材，为解当前的燃眉之急贡献一份力量。

本书将 Web 上的信息检索、过滤和推荐等技术定义为 Web 搜索，使其具有比较宽泛的内涵。这样做的好处是将 Web 检索、过滤、推荐等既联系紧密又相互区分的技术统一在一个体系中，便于进行系统地学习和研究。这是本书的一个显著特色。

本书紧跟技术的最新进展，讨论和介绍重要的研究成果，以及不断涌现的挑战。在写法上以 Web 搜索所包含的主要任务和核心问题为纲、以典型理论模型为例介绍研究的进展，分析对比不同方法在不同方面的优劣，并着力指出它们的局限、当前的研究重点和发展趋势。这一点与通常的教材一般只对成熟的理论进行系统总结相比有很大的不同。

本书共有7章，第1章为导论，第2章～第4章为信息检索，第5章为信息过滤，第6章为信息推荐，第7章为发展前沿。

第1章给出Web搜索的定义，对其发展背景、挑战性、科学价值以及研究状况进行介绍。

第2章通过文本检索介绍Web检索的基本理论和方法。主要内容包括Web信息采集、文本的保存与索引、检索模型、概率模型、查询重构、文本聚类、文本分类等。

第3章讨论图像检索，主要内容包括图像的文本自动标注、物体识别、文字识别、人脸检测与识别等典型算法及最新研究进展，以及视频检索中的镜头切分和视频摘要的典型算法。

第4章讨论音频检索，对声学特征抽取、基于HMM的语音识别统一框架、语音关键词发现技术、语音词汇检测技术、音频的声学模型、音频的语义模型、声学空间与语义空间的联系等技术进行重点介绍，并对音乐检索等非语音音频检索的模型和方法进行研究和讨论。

第5章阐述Web搜索第二个方面的技术——信息过滤，重点讨论信息过滤系统中的分类器选择、学习和优化问题，对现有的典型算法进行介绍，并以垃圾邮件及垃圾短信过滤系统、话题检测与跟踪系统为例，讨论信息过滤系统的具体实现技术。

第6章阐述Web搜索第三个方面的技术——信息推荐，首先对基本的关联规则挖掘和协同过滤算法进行介绍，然后重点讨论提高算法的效率和质量的途径。其中若干算法研究直接针对信息推荐系统中的实际问题。

第7章对内网检索、对象检索、博客观点检索等前沿研究进行介绍。以企业检索为例介绍内网检索的基本技术内涵，以Web学术检索和产品检索为例介绍对象检索要解决的主要问题，以典型系统为例分析博客观点检索的主要技术特点。

本书的内容来自3个方面：经典信息检索理论，同行学者近期发表的论文，作者及所指导的研究生的研究成果。在这里，对同行及研究生的工作所给予的帮助表示衷心的感谢。

由于作者水平有限，加之涵盖的内容尚在迅速发展之中，本书难免存在不足、不当甚至错误之处，希望同行及广大读者批评指正。

作者

2009年3月

目　录

第1章 导 论

1.1 Web搜索的定义

当今信息时代,Web上的信息搜索已经成为影响人类物质文明和精神文明进程的重大问题。其原因在于:信息作为与物质和能量同等重要的资源为人们所认识和利用,社会的发展、技术的进步、物质文化生活水平的提高空前地刺激了人们对信息的需求。

本书所定义的Web搜索,是指在以万维网World Wide Web为典型代表的网络上检索、过滤和推荐信息的理论、方法、技术、系统和服务,也称网络搜索。需要注意,本书所定义的Web搜索包含Web信息的检索、过滤和推荐3个方面。

检索、过滤和推荐这三者既有密切的联系,又有显著的区别。具体如下:

检索是由用户提出查询需求,系统根据这个需求对Web信息进行查询并给出结果。例如,用户通过搜索引擎查询某研究方向的论文。

过滤是系统根据预先设定的条件,对Web中与该条件相符的信息进行获取、隔离或封堵。例如,情报侦听、垃圾邮件过滤、黄色图像过滤。

推荐是系统将用户需要的重要信息从大量的一般信息中抽取出来,并主动推荐给用户。例如,在电子商务系统中向顾客有选择地提示商品信息。

之所以将Web信息的检索、过滤和推荐统称为Web搜索,一方面是因为这三者都需要系统在Web中"寻找"与需求相符的信息,并且这种"寻找"通常是很花"力气"的,是在Web上的一种"搜索"。另一方面,与Web信息搜索有关的研究、开发和应用虽然名目繁多,但主要内容均可归纳到检索、过滤和推荐这3个方面。

另外,传统的信息检索等研究一般是针对封闭数据集的,这与在Web这个数据海量无边、数据特征高度动态的环境中的信息搜索是有巨大差别的。因此要十分明确地将Web搜索与传统的信息检索区分开。

1.2 Web搜索的发展背景

Web搜索是在网络和数字内容等信息技术的强力推动下发展起来的。

近20年来,网络技术日新月异。当前,随着光纤骨干传输网络、新一代无线

移动接入网络的相继建成，一个可以综合提供文本、图像、视频、音频服务的宽带信息高速公路正在向世人开放。同时，Web 上的信息也在随着数字内容采集制作技术的日益成熟，以及大容量存储器的廉价化而迅猛增加。时事新闻、数字音像、科学文献、远程教育、动漫游戏、金融信息、政府公告、网络论坛、网络博客、音视频播客等数字内容应有尽有，形成了无边的信息海洋。

Web 信息的海量化，一方面为满足人们的需求提供了无尽的可能，另一方面也使人们的信息查询反而变得更加困难。因为在信息海洋中找到最需要的东西常常是宛如大海捞针。人们永远怀疑得到的东西是不是最好的，因为无法证明是否还有更好的。更糟糕的是，如果没有有效的技术手段，找到比较好的、比较满意的东西都是困难的。

搜索引擎在这种背景下应运而生，并迅速得到大众的青睐。使用先进的搜索引擎，可以为用户的信息检索提供有效的帮助，使用户在较短的时间内找到比较满意的结果。大众对搜索引擎的依靠也迅速使其成为信息产业的发展热点，几年之内就造就了多家世界顶级企业。

此外，Web 中非法信息、有害信息和垃圾信息的大量传播严重污染了 Web 信息环境，干扰和妨碍了人们的信息利用。不法分子甚至利用 Web 造谣惑众，挑起事端，危害社会稳定和公共安全。而利用 Web 泄露他人隐私，谋取不法利益的行为也很严重。例如，肆意公布知名人士的家庭住址、私人照片、电话号码等。同时，人们也会由于各种需要从 Web 上抽取指定类别的信息，如军事情报、反恐信息等。面对上述问题和需求，Web 信息过滤受到了世界各国学术界、产业界和政府的高度重视，各类过滤技术的研究和开发活动十分活跃。

在 Web 信息极其丰富的条件下，如何为用户提供有用信息是一个非常重要的问题。例如，如何为订阅新闻的用户提供他所关心的消息，如何为电子商务系统的用户提供他所感兴趣的商品信息等。解决这类问题需要利用相关数据对用户的特点、偏好等进行挖掘，从而对其感兴趣的信息进行筛选。这种强烈的需求有力地促进了人们对信息推荐的研究，形成了一个生机勃勃的研究领域。

同时，Web 信息的检索、过滤和推荐在研究和开发中相互渗透、相互交叉，形成了你中有我、我中有你的局面，并且这种整体性的趋势正日益明显。

1.3 Web 搜索的挑战性

全世界不同领域众多学者和技术人员多年的研究实践表明，Web 搜索是一个极具挑战性的任务。主要原因在于其中包含了数据海量、数据稀疏、媒体多样、大量并发请求、数据特征演进、主观客观交叉等困难问题。

如前所述，Web 中的信息完全可以用无边的海洋来形容。这使得人们在处

理一个具体搜索任务的时候,经常面对海量的数据。这一点只要看一下搜索引擎为人们的一个查询所返回的成千上万的结果就足以了解,更何况这成千上万的结果是在比其大得多的数据集内筛选出来的。这样巨大尺度的数据处理问题带来了对算法的全新要求,许许多多在中小尺度问题中适用的算法变得不再适用。面向海量数据处理的算法研究成为 Web 搜索的一个关键和难点。

另外,在 Web 搜索中还常常存在数据稀疏的问题。所谓数据稀疏,通俗地讲是指由于数据不完整所造成的算法所需要的数据的严重缺失。Web 中有海量的数据,但它们又是非常不齐全、不全面和不统一的。例如,在建立网页索引时,常常需要网页的标题,而大量的网页是没有标题的。许多理论模型也会带来数据的稀疏问题,例如,著名的向量空间模型 VSM 是用预先确定的 N 个词表示所有文档,这个 N 通常是万数量级的。那么对于较小的文档,它的 VSM 就会非常稀疏,因为它包含的词很少。数据稀疏会严重影响算法的有效性,有时还会导致零分母、奇异矩阵等难题。

Web 上的信息媒体越来越多样,总体上有文本、图像、视频、音频、二进制程序代码、二进制数值等形式。而每种形式又分为众多的子类,例如文本文档有 TXT、WORD、PDF、RTF、HTML 网页、XML 网页等多种格式,图像文档有 JPEG、TIFF、GIF、BMP 等多种格式。媒体的多样化为 Web 搜索带来了困难,因为不同媒体的解码和处理需要不同的算法。一个系统要综合处理各种媒体信息,其开发成本和系统资源成本是很高的。更为不利的是,这种多样性使得许多核心算法在面对不同媒体时难以做到统一和一致。

Web 搜索面向的是亿万大众,因此用户服务请求的并发性非常高。在研究算法时,要格外注意存储和计算资源的开销,要对高度并发的操作进行有效和可靠的应对,防止系统瘫痪和崩溃。这种要求已经超越了普通的信息检索、过滤和推荐的研究领域,进入到了可信计算、并行计算、分布处理等领域。原本在单机或小型网络中适用的算法会变得不再适用。

Web 数据总是动态变化的,这种变化包括主题、媒体、数据量等多方面的变化。变化的结果导致 Web 数据的特征分布的不断演进。而这一特点在面向数据库的信息检索等系统中是不存在的。由于数据特征模型是 Web 搜索的核心,因此其演进性将从总体上影响搜索的性能。如何自动跟踪 Web 数据特征的演进,保证搜索系统不与数据特征相偏离,是 Web 搜索研究面临的一个十分严峻的挑战。

Web 上的信息是客观存在,但这种客观存在是受大众的主观需求影响的。一方面,大众关心什么,需要什么,Web 上就会有什么,就会增加什么。另一方面,Web 搜索更是直接面向用户的需求和偏好,用户想要什么,就应该去搜索什么。因此,Web 搜索不仅要研究 Web 上客观存在的信息,还要研究用户的主观

意志、行为偏好。这又是不同于普通自然科学研究的一个重要方面。

1.4 Web 搜索的科学价值

Web 搜索广阔的应用领域、巨大的社会经济作用以及高度的技术挑战性使其充满了科学研究价值。

第一,Web 搜索所研究的是一个崭新的科学问题,即如何在无边的动态的 Web 信息中寻找最符合用户需求的信息。这个问题不仅在尺度上空前巨大,而且约束条件非常不确定。因为系统通常难以了解用户真正的信息需求。用户总是希望以最简单的提问或最便捷的操作,如输入少量关键字的方式来表达自己的请求,因而系统得到的指示是十分笼统和模糊的。我们应该认识到,Web 搜索在计算规模和约束的不确定性方面已经将人类的科学研究带到了一个新高度。

第二,Web 搜索既要考虑信息的客观性,又要考虑信息的主观性。所谓信息的客观性,是指信息的数据形式在 Web 中是客观存在的,不论面对哪个主体(用户),承载信息的数据都是相同的。而信息的主观性是指同样的数据给用户提供的信息(量)是不同的。一篇介绍摄影常识的文章对初学者来说可能"很有信息量",而对一个摄影师来说信息量几乎为零。在 Web 搜索中,上述客观性因素和主观性因素都会影响搜索结果的正确性(质量)。这种特点在普通的自然科学研究中是很少见的,因此引起了人们更大的研究兴趣。

第三,Web 搜索强有力地带动了相关学科,特别是智能学科的发展。智能学科中的自然语言理解、模式识别、机器学习、数据挖掘等在 Web 搜索中找到了巨大的发展空间,近年来已经形成了空前高涨的研究热潮。例如文本分类、多媒体识别、海量数据挖掘、在线增量机器学习、在线分类和聚类、信息抽取、信息摘要、命名实体识别等研究都紧密地与 Web 搜索联系了起来。商用搜索引擎的智能化趋势也正是在这些研究的基础上形成的。甚至可以预期 Web 搜索将成为一个大面积涵盖智能学科的新兴独立学科。

1.5 Web 搜索的研究状况

Web 搜索的研究已经在全球范围内掀起了高潮。各国学术界、产业界和政府部门都对其给予了高度的关注,得到了各类国家计划、研究基金和企业项目的大力支持。在我国,国家 863 计划、国家 973 计划以及国家自然科学基金都在积极开展有关的研究。国际上,SIGIR(Special Interest Group on Information Retrieval, ACM 的年会)、SIGKDD(Special Interest Group on Knowledge Discovery and

Data mining, ACM 的年会)、TREC(Text REtrieval Conference, NIST 举办的年会和测试)、TDT(Topic Detection and Tracking, NIST 主办的测试)、MUC(Message Understanding Conference, DARPA 主办的测试)、ACE(Automatic Content Extraction, NIST 主办的测试)等国际会议和评测活动十分活跃,吸引了全世界的注意,强有力地推动了研究进展。

Web 搜索在理论研究方面取得了长足的进步。关于文本搜索,基于 Markov 过程的 N-gram 模型和 Salton 的向量空间模型(Vector Space Model, VSM)是目前普遍采用的特征表达模型。而词频-倒文档频度法(TF-IDF)、信息增益法(IG)、CHI 统计量法、互信息法(MI)等提供了有效的特征选择方法。主成分分析、线性鉴别分析和奇异值分解等方法被用于特征降维,并衍生出了潜语义标号(Latent Semantic Index, LSI)的重要概念。Bayes 分类器、支撑向量机、自组织映射、k 近邻以及向量相似度等模型提供了多样性的分类方法。

关于语音搜索,有两种不同的技术路线。第一种是先利用 ASR (Automatic Speech Recognition)技术将语音文档转换成文本文档,然后再用文本过滤的方法进行处理。TDT 测试中的技术就属于这一类。这类技术的主要问题是系统的精度和速度受到语音识别的制约。第二种是基于音频检索、语音关键词定位和语音鉴别(说话人识别、语种鉴别、性别鉴别等)等技术抽取语音文档的声学特征向量,然后进行内容识别和过滤。这种技术直接针对内容识别和过滤的任务要求,有更深的研究潜力。关于 Web 语音内容过滤系统,在 TDT 技术体系之外,基于音频检索的技术比较常见。

关于图像搜索的理论研究也取得了许多重要进展。此项研究与物体图像识别、计算机视觉等关系密切。在物体图像识别和图像检索方面,提出了以星群模型(Constellation Model)、二维多分辨率隐 Markov(马尔可夫)模型(2DMHMM)和高斯混合离散余弦变换模型(GMM-DCT)等为代表的有效方法;在视频检索和计算机视觉方面,镜头切分、故事切分、关键帧抽取、场景分析、动态特征抽取、视频聚类等关键技术已经取得许多突破。

在系统模型研究方面,TREC 会议的测试任务发挥了重要的引导作用。早期的研究主要集中在对经典的 Ad-hoc 检索系统的模型改进上,目前该方向的研究已经进入了高原期,因而转向了其他模型。比较重要的包括 Novelty、HARD、QA 等。Novelty 是一种新颖性检索系统模型,它首先将与查询相关的文档排成一个序列,然后逐个文档抽取与查询相关的句子,内容相同或类似的句子第一次抽取后就不再抽取。这是一种集成了段落查询和信息过滤的检索模型。HARD 代表 High Accuracy Retrieval from Documents,即高精度文档检索。它是一种用户个性化信息检索模型,系统在反馈查询结果时会根据不同的用户以及用户以往的查询经历给出不同的结果。QA 代表 Question Answering,即问答式检

索。它允许用户直接提出问题,系统根据问题去寻找答案,而不是文档。例如,如果用户提问"哪位美国总统打开了中美交往的大门",系统要直接回答"尼克松",而不是提供相关文档。

此外,TREC 的 Enterprise 检索和 Spam 过滤任务也很重要。Enterprise 提出了企业检索也就是内网(Intranet)的检索任务,它不同于互联网上的检索,其研究重点是如何将一个机构内部的信息进行有效的组织和整合,以便对命名实体、主题文件进行检索,如专家检索、邮件检索等。Spam 是 TREC 设立的第一个内容过滤任务,主要目的是推动垃圾信息过滤的研究。

在多个成功商用搜索引擎等技术的推动下,Web 搜索的应用已经普及。除了公众所熟悉的 Web 信息检索应用之外,还包括政府部门的信息内容过滤,国防及安全部门的情报获取,电子商务系统中的商品信息推荐等。

虽然研究、开发和应用已经取得了长足的进展,但 Web 搜索仍然处于发展的初级阶段。在理论上,许多核心问题,如用户需求的把握、文档内容的理解和提炼、相关文档的排序、数据模型演进的跟踪等都是悬而未决的开放问题。当前阶段,人们的主要努力方向是个性化筛选、多媒体融合、专业性划分、语义级匹配等。

1.6 本书的内容

本书重点从理论上阐述 Web 搜索的研究进展以及不断涌现的挑战。由于 Web 搜索尚处于迅速发展阶段,所以本书并不试图进行理论上的系统总结,而是以包含的主要任务和核心问题为纲,以典型理论模型为例介绍研究的进展,分析对比不同方法在不同方面的优劣,并着力指出它们的局限、当前的研究重点和发展趋势。

同时,为了避免不同任务和问题中论述内容的重叠和割裂,本书对各章节的内容进行精心安排,力图既较完整地涵盖本领域的主要理论进展,又整体有序、不出现累赘和冗余。

本书包括导论、文本检索、图像检索、音频检索、信息过滤、信息推荐以及发展前沿等内容。导论对 Web 搜索的基本概念进行定义,并阐述其科学价值和研究状况。第 2 章至第 6 章根据 Web 搜索的主要任务对其核心问题、基本概念和基本方法进行阐述。第 7 章介绍 Web 搜索若干新的研究方向。具体内容包括:

第 2 章通过讨论文本检索中的主要问题对 Web 信息检索的基本理论和方法进行介绍。Web 搜索的研究是从文本检索开始的。无论是产业界的搜索引擎、门户网站等应用系统,还是学术界的理论研究,文本检索都是首要的核心内容。Web 文本检索涉及许多问题,主要包括 Web 信息采集、文本组织索引、文本

内容表示、用户查询方法、相关文本排序、文本聚类、文本分类等。本章对解决上述问题的主要模型和算法进行系统介绍，对其中的若干难点问题，如文本聚类、文本分类以及用于聚类和分类中的特征选择和变换进行深入讨论和分析。

第3章讨论图像检索。图像检索（包括视频检索）是Web信息检索的一种重要形式，虽然系统总体架构和基本技术与文本检索相类似，但涉及图像分析、识别、标注等特殊技术。本章针对图像检索中的上述特殊问题，讨论图像的文本自动标注、物体识别、文字识别、人脸检测与识别等典型算法及最新研究进展。对视频检索中的镜头切分和视频摘要的算法进行介绍。

第4章讨论音频检索。音频检索也是Web信息检索的一种重要形式，但相对文本检索和图像检索而言发展较慢。主要原因是早期的研究采取了基于自动语音识别将语音变成文本的技术路线，使得语音检索变成了语音识别+文本检索的问题，失去了作为一个独立的研究方向的必要性。近年来，基于声学特征的音频检索研究开始兴起，推动了音频检索技术的发展。本章对声学特征抽取、基于HMM的语音识别、语音关键词发现及检测、音频的声学模型、音频的语义模型、声学模型与语义模型的联系等技术进行重点介绍。并对音乐检索等非语音音频检索的模型和方法进行研究和讨论。

第5章阐述Web搜索第二个方面的技术——信息过滤（Information Filtering）。与信息检索不同，从本质上讲，信息过滤是“流环境”下的二元分类问题。这样的本质决定了信息过滤的技术核心和难度，即以模式分类为技术核心，高效高精度地处理数据流。本章针对上述特点，重点讨论信息过滤系统中的分类器选择、学习和优化问题，对现有的典型算法进行介绍。并以垃圾邮件与短信过滤系统、话题检测与跟踪（TDT）系统为例，讨论信息过滤系统的具体实现技术。

第6章阐述Web搜索第三个方面的技术——信息推荐（Information Recommendation）。与信息检索和信息过滤不同，信息推荐的特点是用户的需求不确切，只能通过历史数据和相关数据进行挖掘（预测）。信息资源也是不断变化的，系统需要根据预测得到的用户需求主动向其推荐信息。关联规则挖掘（Association Rules Mining）和协同过滤（Collaborative Filtering）是信息推荐系统中挖掘商品项之间的关联关系以及用户兴趣和需求的两类主要算法。本章首先对基本的关联规则挖掘和协同过滤算法进行介绍，然后重点讨论提高算法的效率和质量的途径。其中若干算法研究直接针对信息推荐系统中的实际问题。

第7章对内网检索（Intranet Retrieval）、对象检索（Object Retrieval）、博客观点检索（Opinion Retrieval）等前沿研究进行介绍。内网检索的意义在于网络中80%以上的信息存放在企业等各类组织内部计算机网络之中，并且信息源复杂，异构性强，不能简单地利用因特网检索技术加以实现。对象检索试图以对象为单位进行信息的抽取和整合，以对象为单位进行信息的呈现，以提高检索系统的