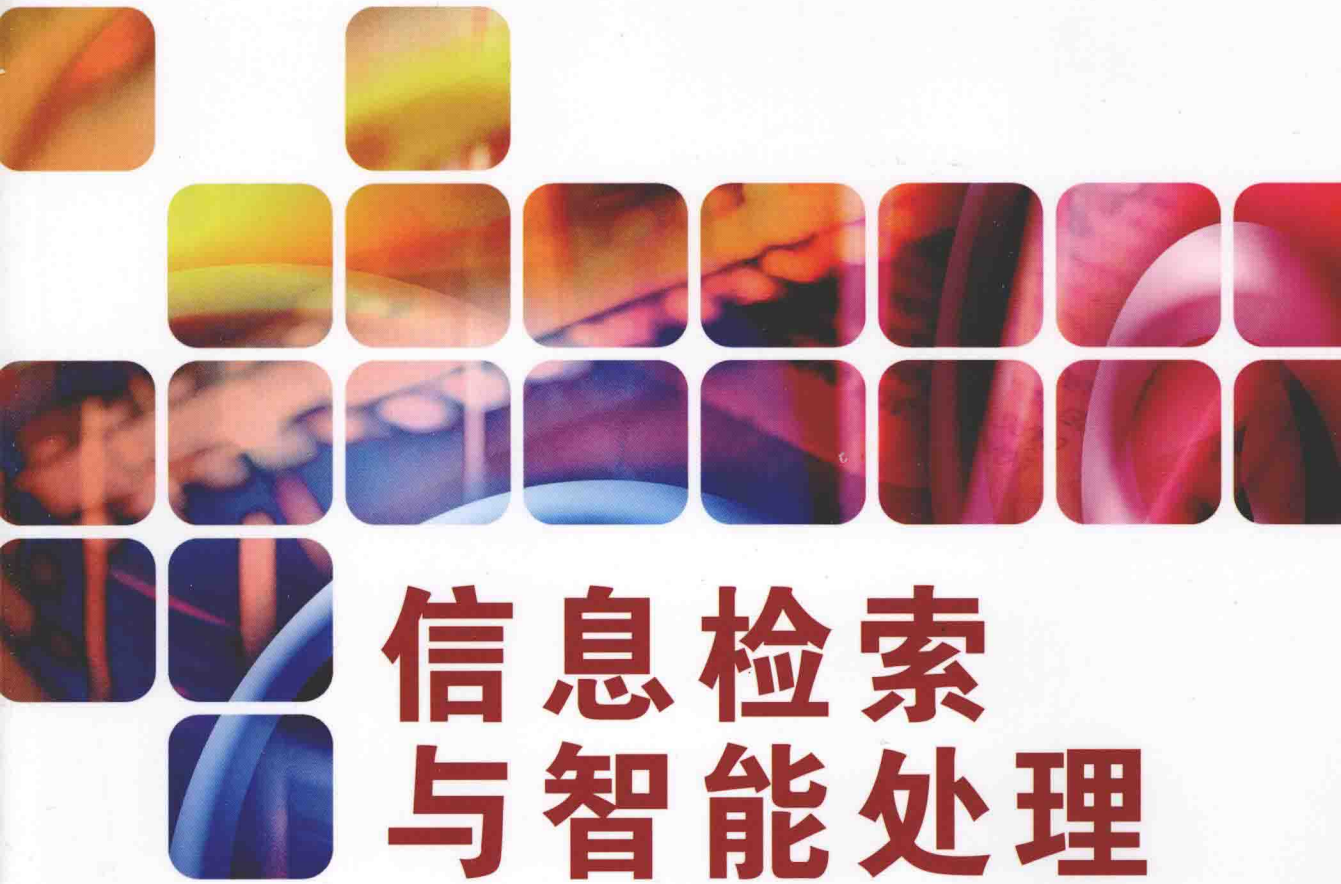

信息检索与智能处理

高　凯　仇　晶　张晓明
王　伟　张华平　著

国防工业出版社
National Defense Industry Press

信息检索与智能处理

高凯　仇晶　张晓明　王伟　张华平　著

国防工业出版社

·北京·

内容简介

本书从多个视角对信息检索和智能处理技术进行了阐述，内容涵盖信息检索系统的架构、检索结果处理、中文自然语言处理、评价方法、Web 检索、网络异构信息采集、网页正文提取与去噪、信息抽取、话题跟踪、主题词标引、分类、聚类、自动摘要、搜索引擎与数字图书馆的开发与应用实践、信息可视化等。全书以模块化的方式进行组织，理论性强，体系完整，内容新颖，条理清晰，组织合理，强调实践。作者团队以认真严谨的科学态度实现了书中绝大部分的主要方法，尽力详尽描述了各种方法的适用环境及取得的效果。

本书可为高校相关专业（如计算机科学与技术、软件工程、情报学、图书馆学、信息管理与信息系统）学生的学习和科研工作提供帮助，同时对于从事信息检索与智能处理技术、社会网络计算的工程技术人员和希望了解网络信息检索技术的爱好者，本书也具有较高的参考价值。

图书在版编目（CIP）数据

信息检索与智能处理/高凯等著. —北京：国防工业出版社，2014. 1

ISBN 978-7-118-09004-8

Ⅰ. ①信… Ⅱ. ①高… Ⅲ. ①情报检索-信息处理 Ⅳ. ①G252. 7

中国版本图书馆 CIP 数据核字（2013）第 223795 号

※

国防工业出版社出版发行

（北京市海淀区紫竹院南路 23 号 邮政编码 100048）

北京奥鑫印刷厂印刷

新华书店经售

*

开本 787×1092 1/16 **印张** 14½ **字数** 320 千字

2014 年 1 月第 1 版第 1 次印刷 **印数** 1—3500 册 **定价** 39. 80 元

前　言

随着手机、个人电脑、互联网等信息工具的快速发展，以及微博、搜索引擎、数字图书馆等的普及应用，个人获取和管理的信息量呈爆炸式增长。由于网上信息的海量性、冗余性及用户需求的多样性等，迄今在网络信息采集、信息抽取、话题跟踪、中文自然语言理解、信息分类与聚类、智能化信息检索、Web 挖掘、信息可视化等方面，尚不能很好地满足用户的需求。如何借助信息检索与智能处理技术，来帮助人们方便、高效地利用网络信息，已成为当前 IT 业的研究热点之一。虽然目前的信息检索工具在一定程度上满足了人们的需求，但随着信息检索的对象从相对封闭、稳定一致、由独立数据库集中管理的内容扩展到开放、海量、冗余、更新快、分布广泛的 Web 内容，同时由于 Web 内容的特殊性（如异构性、网页质量参差不齐、镜像网页的存在、大量非标准的网络用语的普遍使用等），以及中文自然语言处理的特殊性和复杂性，加之使用者由原来的专业检索人员扩展到包括务工人员、管理人士、教师、学生等在内的普通大众等因素的影响，人们对信息检索与智能处理技术也就提出了新的、更高的要求。

信息检索与智能处理技术是解决上述问题的主要方法，它也奠定了现代搜索引擎和数字图书馆的基础。近年来，随着互联网的普及应用以及对海量网络信息挖掘及处理的需要，以信息检索和智能处理技术为基础的搜索引擎、Web 数据挖掘工具等，受到了研究人员的关注和重视。无论是高等院校相关专业方向的研究生，还是对 Web 搜索技术感兴趣的研究者和开发人员，都迫切需要一本全面、专业的书籍。因此，出版有关信息检索与智能处理方面的专著，提高中文自然语言处理和搜索引擎的智能化水平，是一项很有意义的工作。

本书理论性强，体系完整，内容新颖，条理清晰，组织合理，强调实践。它涵盖了信息检索与智能处理的多个重要组成部分，并从多个视角对信息检索和智能处理技术进行了分析，内容涵盖信息检索系统的架构、检索结果处理、中英文分词、评价方法、Web 检索、网络异构信息采集与获取、信息去噪、信息抽取、话题跟踪、中文自然语言处理、主题词标引、分类、聚类、自动摘要、搜索引擎与数字图书馆开发与实践、信息可视化等。通过引用大量的研究文献，作者团队对当今信息检索与智能处理的研究状况给出了综述，并将论述的重点放在了实现和实验上，以便让读者更好地了解到信息检索与智能处理系统的实现细节，同时让读者理解在实践中哪些方法的效果可能会更好。为了这本书，作者团队几乎实现和测试了每一个重要的方法，并进行了多次实验。全书内容分为三篇。第 1 篇“背景知识”涵盖了信息检索与智能处理的背景，其中第 1 章绪论讲述了信息及其分类、信息

检索的起源和发展、信息检索与其他学科的关系等，以期读者对该研究领域有一个了解；第2章是信息智能处理关键技术综述，内容涵盖自然语言处理与中文分词、异构信息处理与内容表示、文本挖掘、实体关系抽取、命名实体识别与话题跟踪、文本分类与情感分析、文本聚类、自动摘要、全文检索、语义Web与信息集成、大数据处理等；第3章是搜索引擎与信息检索综述，包括搜索引擎概述、网络信息检索与处理的基本流程、开源研发工具简介、信息检索评测、信息检索模型与基本方法、性能评价指标等。第2篇“信息处理”涉及第4～10章的内容，其中，第4章是海量异构信息采集，主要讨论了对网络异构信息的采集、分布式信息采集、动态数据采集算法等；第5章是网页正文提取与解析，提出基于网页文字密度的正文提取算法、基于DOM的网页正文提取等；第6章是实体关系抽取，提出了一种基于核函数及无指导机器学习算法的实体关系抽取方法，并对实验结果进行了分析和比较；第7章是命名实体识别及话题跟踪，对语言模型、基于时间信息和标题信息的话题跟踪进行了介绍；第8章是主题概念自动标引，重点对基于概念分析的主题概念自动标引、基于遗传算法的主题概念权值学习与调整等进行了介绍；第9章是文本自动摘要，研究了基于向量空间模型的文档结构模型，提出了基于相似度的文本摘要算法；第10章是文本自动分类，介绍了基于主题词标引的网页分类算法，提出了基于哈希表的动态向量降维技术，改进了向量余弦相似度算法，构建了基于主题词向量的向量空间模型。第3篇“应用”涉及第11～13章的内容，其中第11章是网络信息检索系统的设计与实现，分别给出基于SQL Server全文检索功能实现的“艺海拾贝”搜索引擎、基于Nutch和Lucene的“校园采风”垂直搜索引擎，以及一个集信息采集、去噪、抽取、摘要、分类等功能于一体的信息检索原型系统的实现；第12章给出有关科技文献共享平台数字图书馆的设计与实现方法；第13章给出基于Prefuse的信息可视化的实现。

作者团队以认真、严谨的科学态度实现了书中绝大部分的主要方法，尽量详尽地描述了各种方法的适用环境以及取得的效果。很多老师、同事、学生也花费了大量的时间帮助我们审阅和组织了与其研究领域相关的章节内容。在本书的写作与相关科研课题的研究工作中，得到了多方面的支持与帮助。这里要特别感谢高国江、王向、李媚、张坤、阮冬茹、马红霞等提供的帮助，感谢作者团队指导的研究生王亚歌、宗宝琴、刘邵博、沈琳、周二亮、王九硕、陶秋红，以及高莘、李洋杰、张林立、阳春晖、李跃鹏、潘东宇等同学，有些内容来自这些同学们的论文和他们直接参与完成的课题。有几个班的研究生同学组织了部分早期的资料，高莘协助完成了相关的文献整理和书稿的校对工作。感谢他们的辛勤付出。

本书得到2013年国家自然科学基金（编号：61272362）、2013年河北省自然科学基金（编号：F2013208105、F2013208107）、2012年河北省科技支撑计划（编号：12213516D）、2012年河北省青年科学研究基金（编号：F2012208016）的支持，并受河北科技大学学术著作出版基金资助。中国科学院龙星计划、CCF前沿技术讲习班、国内外众多的信息检索与智能处理方面的研究和相关网站亦为本书提供了良好的基础。本书的顺利完成也得益于参阅了大量的相关工作及研究成果，在此谨向这些文献的作者以及为本书提供帮助的老

师、同仁、学生和课题组成员致以诚挚的谢意和崇高的敬意。在本书写作过程中，也得到了国防工业出版社的大力支持和帮助，在此一并表示衷心感谢。

由于我们的学识、水平均有限，书中不妥之处在所难免，恳请广大读者批评指正。

高凯　仇晶　张晓明　王伟　张华平

2013 年 10 月

目　录

第1篇　背景知识

第2篇 信息处理

第3篇　应用

第1篇 背景知识

第1章 绪 论

随着因特网的迅速普及和应用，网络已成为人们获取信息的重要渠道。网络信息检索与智能处理使古老的信息处理技术焕发了勃勃生机。应用数据分析与智能挖掘方法，可以帮助人们从海量网络信息中提取知识，可以为网站访问者、站点经营者以及包括电子商务等在内的基于因特网的商务活动提供决策支持。但由于网络信息的海量、冗余、异构等复杂特点，该领域给传统的信息处理技术提出了很多亟待解决的问题。

本章对信息检索与智能处理领域的相关技术进行综述，内容包括信息及其分类、信息检索的起源和发展、信息检索与其他学科的关系等，以期读者能对信息检索与智能处理的研究范畴、背景和研究意义以及与相关学科的关系等有一个大致了解，以便能更好地理解后续章节中提到的相关技术，为更进一步的学习和研究打下良好基础。

1.1 信息及其分类

在社会信息化的今天，信息已经成为全社会宝贵的资源。就信息本身而言，大致可将其分为结构化信息、半结构化信息和非结构化信息三类。

结构化信息一般指那些经过整理并按照一定格式编码存放且不含无关内容的信息。一般来说，结构化信息的字段含义确定、清晰。作为管理结构化信息的有效手段，数据库系统对当今科研部门、政府机关、企事业单位等都是至关重要的；作为数据库系统中的核心，数据库管理系统 DBMS，特别是关系型数据库管理系统(如 Microsoft SQL Server、Oracle、Sybase 等)，可用于高效创建和维护结构化信息。同时，数据库技术也是计算机科学与技术领域中发展飞快的一个分支，在其将近半个世纪的发展历程中，已造就了包括 C. W. Bachman、E. F. Code、James Gray 等在内的多位图灵奖得主，并发展成为具有很大工程实践价值的学科。

虽然传统的关系数据库管理系统在管理结构化信息方面具有得天独厚的优势，但在网络迅速普及的信息化社会中，网络信息的规模和种类比传统信息有了很大提高和扩大。其中，半结构化信息和非结构化信息所占比重逐步增大，它们已经逐渐成为主要的信息组织方式。据统计，半结构化和非结构化信息占整个信息量的 80%以上，而由于关系型数据库自身结构的缘故，它管理大量半结构化和非结构化信息显得有些不方便，且查询这些信息的速度比较慢，因此针对半结构化数据和非结构化信息的全文检索技术日益受

到人们重视。全文检索技术也是开发信息检索系统与搜索引擎、构建大规模语料库、开发数字图书馆等诸多重要应用的基础[1]。

一般来说，半结构化信息的特点是它们往往具有一定的结构，但语义不十分确定。所谓的“非结构化”信息，并非指该信息没有任何结构，而是指其结构不是显式(而是隐式)存在的。要找出其中的结构，需要运用某种文本处理技术，如中文分词技术可把中文词从句子中分割出来，而隐性语义分析技术则从词汇/文档关系的挖掘中发现文本的深层结构[2]。非结构化信息一般是无法用统一的结构来明确表示的，也因此较难按照统一的模式进行信息结构化和信息抽取。对于非结构化信息(如大规模文本信息、语料库)而言，必须通过阅读并理解文本的方式才能挖掘到相关信息。

本书的主角，即信息检索，可以帮助我们从半结构化或非结构化的信息集合中找出与用户需求相关的信息。在这个信息处理过程中，需要用到智能处理相关技术，如自然语言处理、网络信息检索与数据挖掘、机器学习、标引、分类、聚类、自动文摘、话题跟踪、语义 Web、信息可视化技术等。上述智能化技术也是本书所关注的重点。

1.2 信息检索的起源和发展

信息检索是伴随着人类社会的进步而发展起来的。从广义上来说，信息检索一般指用户为处理解决各种问题而查找、识别、获取相关的事实、数据、文献的活动及过程(如用户在图书馆查询相关信息资料的行为就属于一种信息检索行为)。而狭义的信息检索一般指用户在计算机信息检索系统上进行信息查询的行为[3]。一般认为，信息检索是经过了手工检索、脱机批处理检索、联机检索、光盘检索等多个阶段后，逐步发展到今天的网络信息检索阶段。

1.2.1 手工检索

顾名思义，手工检索就是指人们以手工为主的方式进行信息检索。在长期的社会和生产实践中，手工检索曾经扮演了重要的角色，特别是在计算机出现以前的漫长岁月中。这种传统方法主要是指人们利用人工或借助简单的机械工具，对记录在纸质媒体上的资料进行检索。例如检索者通过书本式目录、卡片式目录等检索工具，利用人工查找文献，就属于这种方式。手工检索基本是依靠人脑的思考、比较和选择进行的。这种方法的灵活性高，也便于对自己的检索策略进行修改。一般来说，检索到的文献也基本能符合检索者的要求，查准率比较高。但是其缺点也是显而易见的，那就是检索速度慢，效率不高，而且检索结果易受检索资源的限制[1, 3]。

1.2.2 脱机批处理检索

这种检索方式大致出现在 20 世纪 50 年代中期至 60 年代中后期。所谓脱机批处理检索方式，是指定期由专职检索人员把许多用户检索需求汇总，一次性批量处理检索要求，并把结果提供给用户。与手工检索相比，脱机批处理有其自身的优点(如它可同时进行多项检索，可处理检索关系复杂的检索词汇，且一次输入检索提问后可多次或以多种方式输出检索结果)。但批处理也有其不足之处，最明显之处就是缺乏人机交互，而且如果用

户的信息需求和查询结果之间可能有偏差，也不便于进一步修改检索条件和进行二次检索[1, 3]。

1.2.3 联机检索

联机检索是计算机技术、信息处理技术和现代通信技术三者的有机结合。用户在检索过程中可以修改检索策略，但是联机检索指令比较复杂。从20世纪60年代中期到70年代初，由于计算机分时技术的发展及通信技术的改进，用户可以通过检索终端与检索系统的中心计算机进行交互，从而对远距离的数据进行检索。早期的大规模联机检索系统是美国NASA的RECON系统，此外还有美国的DIALOG系统、ORBIT系统、BRS系统及欧洲的ESA-IRS系统等[3]。

1.2.4 光盘检索

光盘是20世纪80年代出现的利用激光和光电集成技术等实现的存储体。在使用网络信息服务比较困难的地区，光盘检索就会显示出它特有的优势，比如它可以免去联机检索必须支付的联机使用费，且存储方便、容量大。光盘检索除可提供追溯检索、定题服务外，还可用于自建库和做联机检索前的预处理[1, 3]。

1.2.5 网络信息检索

随着互联网的崛起，网络信息检索工具越来越受到重视。网络信息检索能够使人们在很短的时间里，尽可能全面地查找到相关信息(如网页、论坛、软件资料、图像文件、声音文件等)，而网络信息检索反过来也促进了互联网的发展和普及应用，使网上信息的利用率进一步得到提高。例如，企业会根据搜索引擎或网站的知名度及日流量来选择是否要投放广告，普通网民则会根据搜索引擎的性能和技术选择自己喜欢的搜索引擎来查找资料。可以说，网络信息检索正在发挥越来越重要的作用。信息检索工具(特别是网络信息检索工具)已经成为现代商业社会和日常生活中不可或缺的一部分。在国外，“Google”已经不再单纯是一个名词，有时它充当的则是动词的角色；在国内，“百度一下”等流行语的出现，也说明现在人们越来越依赖于搜索引擎。

网络搜索和挖掘是建立在信息检索、数据挖掘基础上的。虽然信息检索有着很长的历史，但它真正受到人们的关注还是近半个世纪的事。近年来，随着互联网的普及应用，网民规模也日渐增大。据2013年7月中国互联网络信息中心发布的“第32次中国互联网络发展状况统计报告”显示，截至2013年6月底，我国网民规模达到5.91亿，互联网普及率为44.1%。互联网已经成为人们获取信息的首选途径。如何在浩如烟海的网络信息中快速检索和挖掘到所需信息，已经成为时代发展的需要。可见，网络搜索正在成为互联网中的基础应用，它也正在发挥着日益重要的作用。表1.1是中国互联网络信息中心在2013年7月发布的“第32次中国互联网络发展状况统计报告”统计的2012.12—2013.06期间各类网络应用使用率情况，从中可以看出，截至2013年6月，我国搜索引擎用户规模超过4.7亿，较2012年6月底增长了1928万人，在网民中的渗透率为79.6%，搜索引擎也稳居互联网第二应用之位。另据2012年10月中国互联网络信息中心发布的“互联网发展信息与动态”报告显示，2012年上半年“搜索引擎”的用户数占总覆盖人